



République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la recherche Scientifique

Université Mouloud Mammeri de Tizi-Ouzou

Faculté de Génie Electrique et d'informatique

Département d'informatique

Mémoire de fin de cycle

**En vue de l'obtention du diplôme de Master Académique
en Informatique**

Thème :

Les Big Data : Etat de l'art & Challenges

Promotion : 2013 / 2014

Option : Systèmes informatiques

Réalisé par : Senani Ahmed

Dirigé par : Mr. Idir Rassoul, MCA

Membres du Jury :

- | | |
|-------------------------|---|
| - M. AOUGHLIS Farida, | Maître de Conférences Classe A, Présidente. |
| - Mr. HAMMACHE Arezki, | Maître de Conférences Classe B, Examineur. |
| - Mr. OUAMRANE Mohamed, | Maître Assistant Classe A, Examineur. |
| - Mr. RASSOUL Idir, | Maître de Conférences Classe A, Rapporteur. |

Date de soutenance : 03 juillet 2014

Remerciements

J'aimerais en premier lieu remercier le bon dieu de m'avoir donné la volonté et le courage pour la réalisation de ce travail.

Je tiens à exprimer ma profonde gratitude et mes sincères remerciements à mon encadreur Monsieur Idir Rassoul, Maître de Conférences au département d'informatique, pour avoir accepté de diriger ce travail de recherche. Son soutien, sa clairvoyance, ses compétences et ses précieux conseils, furent d'un apport considérable et m'ont été d'une aide inestimable.

De même je remercie les membres du jury qui m'ont fait honneur d'accepter d'évaluer et de juger mon travail. Qu'ils puissent trouver à travers ce dernier, le témoignage de ma sincère gratitude et de mon profond respect.

Enfin, mes remerciements les plus chaleureux vont à mes parents pour leur soutien inconditionnel, et à tous les membres de ma famille qui ont participé de près ou de loin au bon accomplissement de mon travail.

Ahmed Senani.

Résumé

L'accroissement continu des données numériques dans notre société et la démocratisation des terminaux mobiles (Smartphones, tablettes, GPS, ... Etc.) ainsi que la baisse des prix d'accès à internet ont conduit à l'émergence d'un nouveau phénomène appelé «Big Data».

L'information, aujourd'hui, existe sous tous types de formes et est produite par toutes sortes d'acteurs, ce qui la rend difficile à gérer et à analyser ... avec les outils traditionnels en tous cas.

En quoi consistent les Big Data ? Quels sont les avantages qu'apportent les Big Data, de même pour ses inconvénients ? Quelles conséquences sur la vie privée d'autrui et la sécurité des données ?

Mots clés : Big Data, Masses de données hétérogènes, Bases de données hétérogènes, Cloud Computing, Data Mining, Business Intelligence.

Table des matières

Chapitre I : Introduction aux bases de données

I.1. Définition	1
I.2. Utilité d'une base de données	1
I.3. La gestion des bases de données	2
I.4. Quelques principaux SGBD	3
I.5. Typologie des bases de données	4
I.5.1. Selon le nombre d'utilisateurs	4
I.5.2. Selon l'usage	4
I.5.3. Selon le contenu	5
I.6. Naissance d'un nouveau « Paradigme »	8

Chapitre II : Big Data et bases de données non structurées

II.1. Introduction aux Big Data	10
II.2. Définition du Big Data	15
II.3. Les caractéristiques des Big Data	17
II.3.1. Volume	18
II.3.2. Variété	24
II.3.3. Vitesse (Vélocité)	28
II.3.4. Véracité	30
II.3.5. Valeur	32
II.3.6. Visibilité	33
II.4. Pourquoi Big Data, quels apports pour l'entreprise ?	37
II.4.1. En quoi le Big Data concerne-t-il votre entreprise ?	39
II.4.2. Comment mettre en place un projet Big Data ?	40
II.5. Les défis et enjeux liés aux Big Data	42
II.5.1. Défi technologique	42
II.5.2. Un enjeu économique	44

II.5.3. Un défi organisationnel	46
II.5.3.1. La conduite du changement	46
II.5.3.2. L'apparition de nouveaux métiers « Data Scientist »	47
II.5.4. Un défi Juridique	48
II.6. L'analyse, le point clé du Big Data	49
II.6.1. Ce qu'est l'analyse Big data et ce qu'elle n'est pas	50
II.7. Quand le Big data rencontre le Cloud	51

Chapitre III : Les technologies émergentes pour la Gestion des Big Data

III.1. L'Architecture « Shared Nothing » pour les systèmes évolutifs	54
III.2. Le framework « Hadoop »	55
III.2.1. Définition	55
III.2.2. Histoire et origines	55
III.2.3. Concepts fondamentaux d'Hadoop	56
III.2.3.A. Hadoop Distributed File System (HDFS)	57
III.2.3.B. Map/Reduce	62
III.2.4. Les Avantages d'Hadoop	67
III.2.5. L'écosystème « Hadoop »	68
III.2.6. Les distributions d'Hadoop	82
III.3. La solution NoSQL	89
III.3.1. Définition	89
III.3.2. Histoire et origines	90
III.3.3. Les Avantages des systèmes NoSQL	91
III.3.4. Fonctionnement	91
III.3.5. Types de bases de données NoSQL	94

Chapitre IV : La sécurité dans les Big Data

IV.1. Analyse des risques	108
IV.1.1. Risques de non-conformité	108
IV.1.2. Risques de non disponibilité	111
IV.1.3. Risques de perte de confidentialité	112
IV.1.4. Risques opérationnels	113
IV.1.5. Risques liés aux modes d'intégration	114
Conclusion générale & Perspectives	116
Glossaire	119
Bibliographie & Webographie	122

Chapitre I : Introduction aux bases de données

I.1. Définition :

Une base de données est un conteneur servant à stocker des données : des renseignements bruts tels que des chiffres, des dates ou des mots, qui peuvent être retraités par des moyens informatiques en vue de produire une information; par exemple, des chiffres et des noms assemblés et triés pour former un annuaire téléphonique. Les retraitements sont typiquement une combinaison d'opérations de recherches, de choix, de tri, de regroupement, et de concaténation [1].

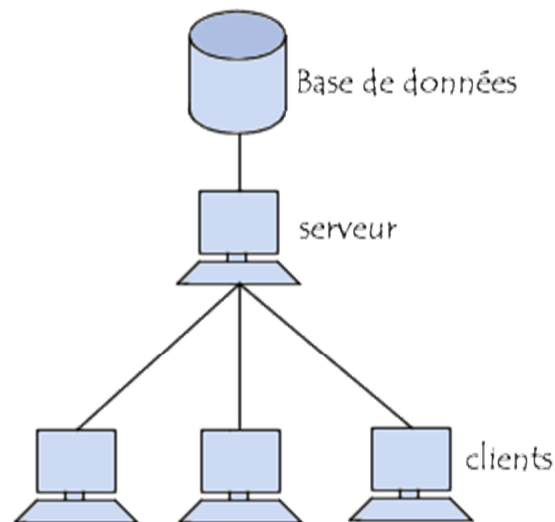
Le recours aux bases de données est une alternative au procédé classique de stockage de données, par lequel une application place des données dans des fichiers manipulés par l'application. L'utilisation d'une base de données en lieu et place de fichiers facilite le partage des informations, permet le contrôle automatique de la cohérence et de la redondance des informations, la limitation de l'accès aux informations et la production plus aisée des informations synthétiques à partir des renseignements bruts. La base de données aura, de plus, un effet fédérateur: dans une collectivité qui utilise une base de données, une personne unique, l'administrateur de bases de données, est chargée d'organiser le contenu de la base de données d'une manière qui soit bénéfique à l'ensemble de la collectivité, ce qui évite les conflits dus à des intérêts divergents entre les membres de la collectivité.

I.2. Utilité d'une base de données :

Une base de données permet de mettre des données à la disposition d'utilisateurs pour une consultation, une saisie ou bien une mise à jour, tout en s'assurant des droits accordés à ces derniers. Cela est d'autant plus utile que les données informatiques sont de plus en plus nombreuses [2].

Une base de données peut être locale, c'est-à-dire utilisable sur une machine par un utilisateur, ou bien répartie, c'est-à-dire que les informations sont stockées sur des machines distantes et accessibles par réseau.

L'avantage majeur de l'utilisation de bases de données est la possibilité de pouvoir être accédées par plusieurs utilisateurs simultanément.

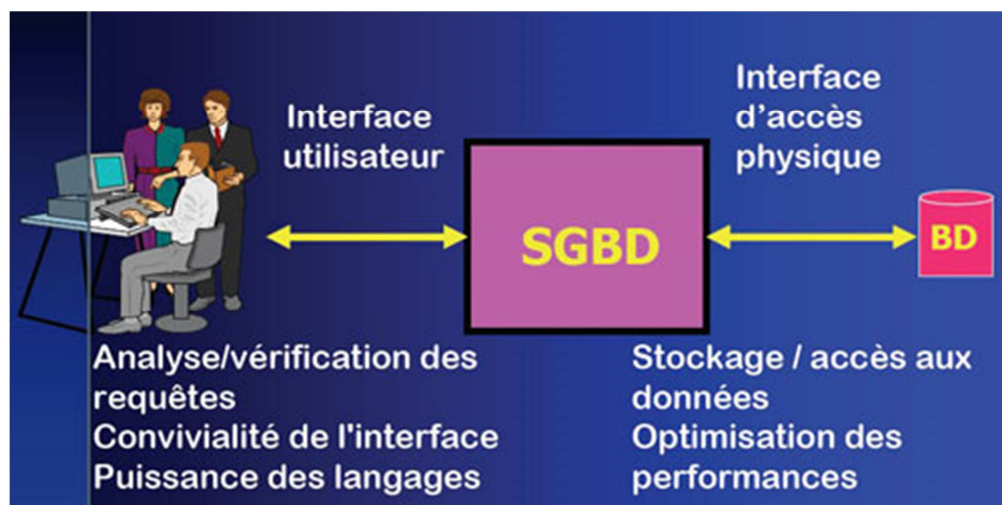


I.3. La gestion des bases de données :

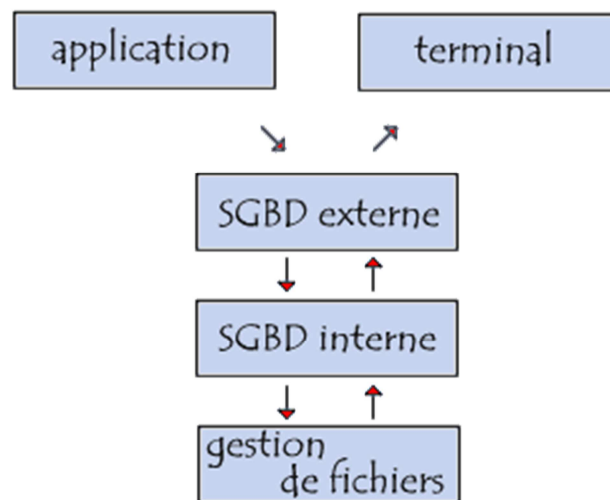
Afin de pouvoir contrôler les données ainsi que les utilisateurs, le besoin d'un système de gestion s'est vite fait ressentir.

La gestion de la base de données se fait grâce à un système appelé SGBD (système de gestion de base de données) ou en anglais DBMS (Database management system).

Le SGBD est une suite de programmes qui manipule la structure de la base de données et dirige l'accès aux données qui y sont stockées. Une base de données est composée d'une collection de fichiers; le seul moyen d'accéder aux données est par le SGBD, qui sert alors d'intermédiaire entre la base de données et ses usagers. Celui-ci reçoit des demandes de manipulation du contenu et effectue les opérations nécessaires sur les fichiers. Il cache la complexité des opérations et offre une vue synthétique sur le contenu. Le SGBD permet en outre à plusieurs usagers de manipuler simultanément le contenu, et peut offrir différentes vues sur un même ensemble de données [3].



Le SGBD peut se décomposer en trois sous-systèmes :



- **Le système de gestion de fichiers :**

Il permet le stockage des informations sur un support physique.

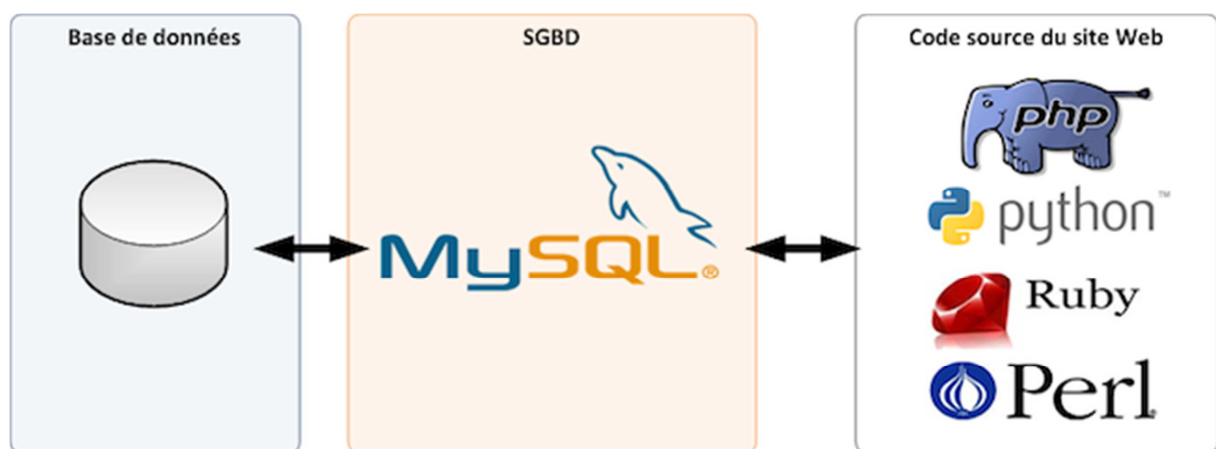
- **Le SGBD interne :**

Il gère l'ordonnancement des informations.

- **Le SGBD externe :**

Il représente l'interface avec l'utilisateur.

I.4. Quelques principaux SGBD :



Les principaux systèmes de gestion de bases de données sont les suivants [4] :

- Borland Paradox
- Filemaker
- IBM DB2
- Ingres
- Interbase
- Microsoft SQL server
- Microsoft Access
- Microsoft FoxPro
- Oracle
- Sybase
- MySQL
- PostgreSQL
- mSQL

I.5. Typologie des bases de données :

L'usage qui est fait des données diffère d'une base de données à une autre, elles peuvent être classifiées en fonction du nombre d'utilisateurs, du type de contenu, notamment s'il est faiblement ou fortement structuré, ainsi que selon l'usage qui est fait de la base de données, notamment l'utilisation opérationnelle ou à des fins d'analyse :

I.5.1. Selon le nombre d'utilisateurs :

Un seul, un petit groupe, voire une entreprise. Une base de données de bureau est installée sur un ordinateur personnel au service d'un seul utilisateur. Tandis qu'une base de données d'entreprise est installée sur un ordinateur puissant au service de centaines d'utilisateurs. Une base de données centralisée est installée dans un emplacement unique, tandis qu'une base de données distribuée est répartie entre plusieurs emplacements [5].

I.5.2. Selon l'usage :

C'est la manière la plus populaire de classer les bases de données, on distingue deux types :

- **Les bases de données opérationnelles :** en anglais OLTP (*Online Transaction Processing*)

Elles sont destinées à assister des utilisateurs à tenir l'état d'activités quotidiennes. Elles permettent en particulier de stocker sur le champ les informations relatives à chaque opération effectuée dans le cadre de l'activité: achats, ventes, réservations, paiements. Dans de telles applications l'accent est mis sur la vitesse de réponse et la capacité de traiter plusieurs opérations simultanément.

- **Les bases de données d'analyse :** en anglais OLAP (*Online Analytical Processing*)

Elles sont composées d'informations historiques telles que des mesures sur lesquelles sont effectuées des opérations massives en vue d'obtenir des statistiques et des prévisions. Les bases de données sont souvent des entrepôts de données (Datawarehouse): des bases de données utilisées pour collecter des énormes quantités de données historiques de manière quotidienne depuis une base de données opérationnelle.

Le contenu de la base de données est utilisé pour effectuer des analyses d'évolution temporelle et des statistiques telles que celles utilisées en management. Dans de telles applications l'accent est mis sur la capacité d'effectuer des traitements très complexes et le logiciel moteur (le SGBD) est essentiellement un moteur d'analyse [5].

I.5.3. Selon le contenu :

En informatique, les données stockées se divisent en trois types : les données structurées, semi-structurées et les données non structurées (ce dernier sera l'objet d'étude des prochains chapitres de ce mémoire) [6].

- **Données structurées :**

Les données structurées sont celles dont l'ensemble des valeurs possibles est déterminé et connu à l'avance, Nous trouvons les données structurées dans les bases de données et les langages informatiques. Nous les reconnaissons au fait qu'elles sont disposées de façon à être traitées automatiquement et efficacement par un logiciel.

Par exemple : Dans une base de données rassemblant les résultats d'une enquête d'opinion, l'âge ou la catégorie socio-professionnelle des individus interrogés sont des données structurées, car les tranches d'âges ou la liste des catégories socio-professionnelles possibles sont déterminées a priori [7].

Plusieurs organisations utilisent des bases de données (donc des données structurées) pour gérer leur inventaire, leur comptabilité Etc. Souvent ces données sont disposées dans des tableaux, comme dans une base de données relationnelle.

- **Données semi-structurées :**

Les données semi-structurées peuvent se voir comme une relaxation du modèle relationnel classique, un des fondements des bases de données traditionnelles, dans lequel on autorise une structure moins rigide et homogène des « champs de données » [8].

Dans les documents textuels, nous trouvons plusieurs niveaux entre l'absence complète de structure et les données structurées.

Exemple : La transmission d'un courrier électronique sur internet.

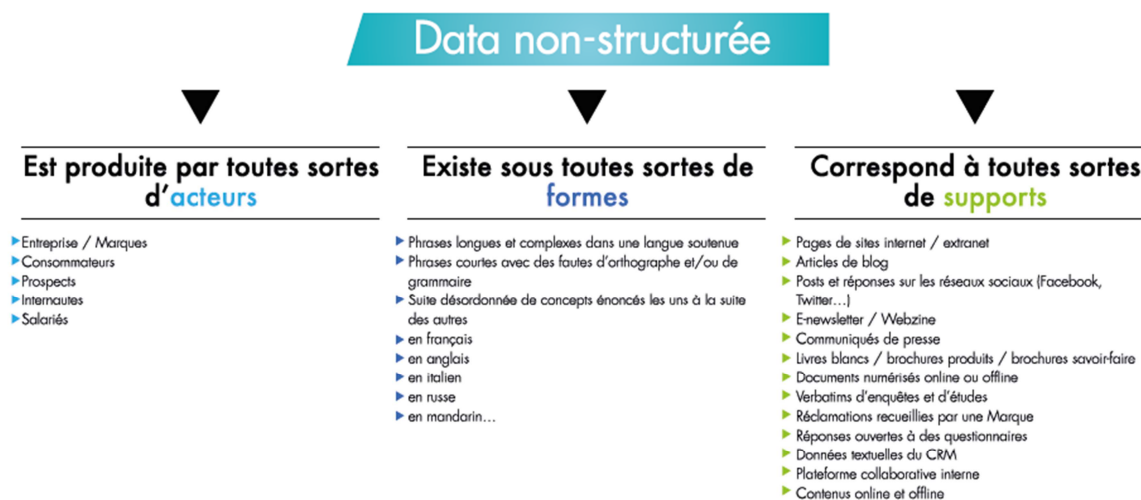
Un courriel contient une combinaison de données structurées (date, auteur, destinataire, ... etc.) et de données non structurées (corps du message).

Dans une page web, on retrouve également cette caractéristique : une partie de son contenu s'adresse à l'humain, comme le texte (données non structurées), alors qu'une autre partie est destinée à la machine, comme les balises (données structurées) [6].

- **Données non structurées :**

Les données non structurées sont des données irrégulières dont les informations ne peuvent pas être simplement rangées dans une base de données classique, par le fait qu'elles contiennent des textes, pages web, e-mails, fichiers multimédia (sons, images, vidéos ... etc.), statuts de réseaux sociaux, livres, rapports ... etc.

Le terme « **non-structurées** » ne fait pas référence à l'absence de structure, mais au fait que ces données ont des structures très complexes et non standards, où l'information ne peut pas s'obtenir avec des requêtes simples auxquelles on est habitué [9].



Les réponses libres aux questions ouvertes sont des données non structurées, car ces réponses sont potentiellement toutes différentes et impossibles à catégoriser.

D'une façon générale, les données non structurées sont des **données textuelles**.

Un film ou un enregistrement audio est destiné à l'humain, et c'est pourquoi ils sont des informations non structurées. Cependant, on ajoute souvent des **métadonnées** aux documents multimédias. Par exemple, les fichiers de musique MP3 contiennent souvent une section ID3 qui permet de spécifier l'interprète, le compositeur, etc.

La manipulation relativement rapide des données structurées a été résolue en grande partie par les bases de données. Par contre, les données non structurées ou semi-structurées posent des problèmes plus difficiles, surtout lorsque la recherche porte sur la partie purement non structurée des informations [6].

Voici quelques exemples de problèmes rencontrés dans :

- **La recherche de données :** Trouver automatiquement dans une pile de CV, le nombre de postulants ayant un niveau d'étude recherché tel que le Baccalauréat, Licence, ... etc.
- **La recherche de documents :** Trouver un rapport écrit en 2009 et qui traite des stratégies et projets d'avenir envisagés dans une entreprise.
- **La recherche de textes :** Trouver un texte qui aborde un sujet précis tel que : la définition d'un terme ou d'une expression dans un ensemble de documents.
- **La recherche de contenus multimédia :** Trouver une séquence audio ou une chanson à travers les paroles ou une suite de termes qui s'y trouvent. Rechercher un film qui correspond à certains critères comme le genre, le réalisateur, certaines séquences ou actions ou même contenant certains textes Etc. Trouver une photo d'un ami prise dans un lieu précis à une date précise ... etc.

Dans tous ces exemples, nous voyons que la solution serait plus facile à trouver si les données étaient structurées. Nous pourrions demander aux postulants de cocher une case dans laquelle ils auraient à répondre à la question « Êtes-vous titulaire d'un baccalauréat ? » ; nous pourrions classer les rapports par sujet ; nous pourrions classer les images et vidéos par date et selon les personnes représentées ... Ce n'est malheureusement pas toujours possible.

I.6. Naissance d'un nouveau « Paradigme » :



Les informaticiens, physiciens, économistes, et beaucoup d'autres accordent un intérêt grandissant aux quantités massives d'informations produites au sein de notre société.

Fruit de la révolution numérique, l'explosion des données transforme en profondeur la manière d'opérer des entreprises.

Aujourd'hui, la majorité des données stockées par les entreprises sont des données brutes (non structurées), et donc non exploitables. Les données structurées, utilisées par l'entreprise, ne représentent qu'un petit pourcentage des données globales.

L'évolution de la technologie et la croissance exponentielle de la quantité d'information générée chaque année engendrent une nouvelle façon de travailler, de voir et d'analyser les données, donner du sens à ce qui avant n'en avait pas, apprendre à tirer un maximum de profit des informations qu'on manipule ... etc.

Les experts ont résumé tout ceci en deux mots « **Big Data** ».

Le phénomène **Big Data** est né. Il doit son développement à la convergence de deux événements majeurs : d'un côté, l'explosion des informations manipulées au sein des écosystèmes digitaux qui nous entourent et d'un autre la prise de conscience que certaines de ces nouvelles informations « brutes » pourraient être analysées et délivrer une nouvelle forme de valeur [10].

Aujourd'hui, les technologies sont assez évoluées pour tirer parti de tout ce déluge d'informations tout en simplifiant leur usage. Ainsi, le nouveau paradigme du **Big Data** est au cœur des nouveaux enjeux de la recherche technologique pour les systèmes du futur de plus en plus complexes au point d'être considéré comme l'un des plus grands défis informatiques de cette décennie (2010-2020) [11].

Dans ce premier chapitre, nous avons vu une description générale des bases de données. Les prochains chapitres seront consacrés au monde des **Big Data**.

Chapitre II : Big Data et bases de données non structurées



II.1. Introduction aux Big Data :

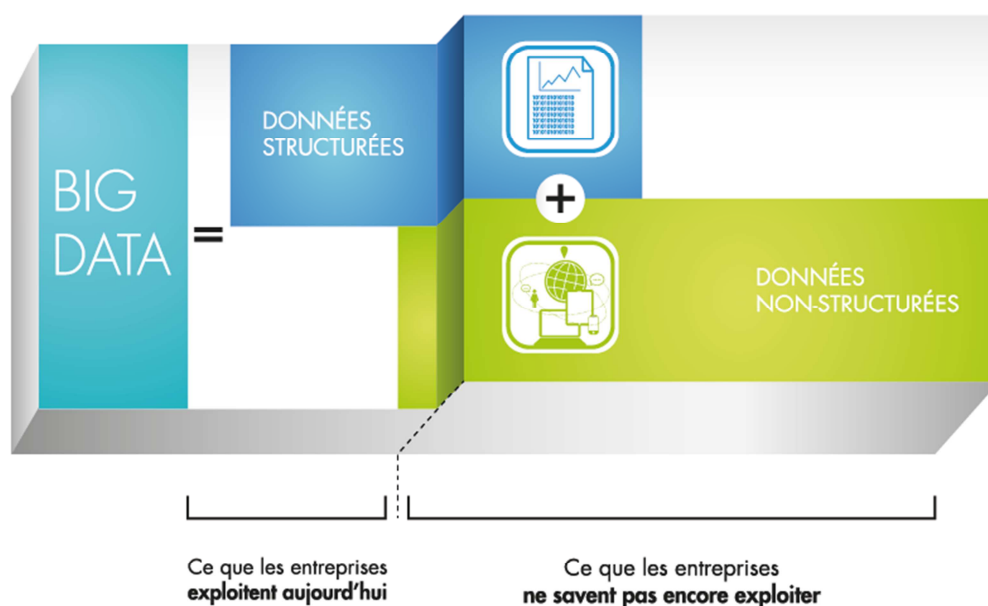
L'humanité produit aujourd'hui autant d'informations en deux jours qu'elle ne l'a fait en deux millions d'années. Des flots d'octets, un océan de données, un déluge de connaissances, à mesure qu'Internet tisse sa toile, le volume d'informations numérisées n'en finit plus d'exploser. Le cabinet d'études IDC prédit que D'ici huit ans, cette masse vertigineuse de données sera 50 fois supérieure à ce qu'elle est aujourd'hui et il faudra dix fois plus de serveurs informatiques pour espérer les gérer et exploiter [12].

Il y a vingt ans, nous stockions encore nos fichiers sur des disques durs de quelques mégaoctets. Aujourd'hui, la capacité des outils de stockage a dépassé le **téraoctet** et il n'est plus rare pour les entreprises et les organismes de recherche de manipuler des volumes supérieurs au **pétaoctet** (10^{15} octets).

Depuis quelques années, on entend partout le terme « **Big Data** », derrière ce terme se cache une myriade de technologies (certaines vieilles de plus de dix ans). Si sa popularité et son intérêt sont nouveaux, l'idée elle n'est pas nouvelle mais ce qui a fait que le Big Data soit aujourd'hui à la mode c'est l'augmentation drastique de la quantité de données, due entre autres à l'augmentation des sources de données (blogs, médias sociaux, recherches sur internet, réseaux de capteurs, etc.), qui permet de nouvelles utilisations des données.

En effet, lorsque l'on parle de manipulation de gros volume de données, on pense traditionnellement à des problématiques sur le volume des données et sur la rapidité de traitement de ces données. Mais aujourd'hui, il est aussi question de manipuler des données venant de sources diverses ; des données qui n'ont pas forcément beaucoup de valeur en elles-mêmes, mais qui, croisées les unes avec les autres, offrent une mine d'informations clés pour l'entreprise (marketing personnalisé, moteurs de recherches, surveillance, ... etc.).

Les données non structurées au cœur de l'information stratégique :

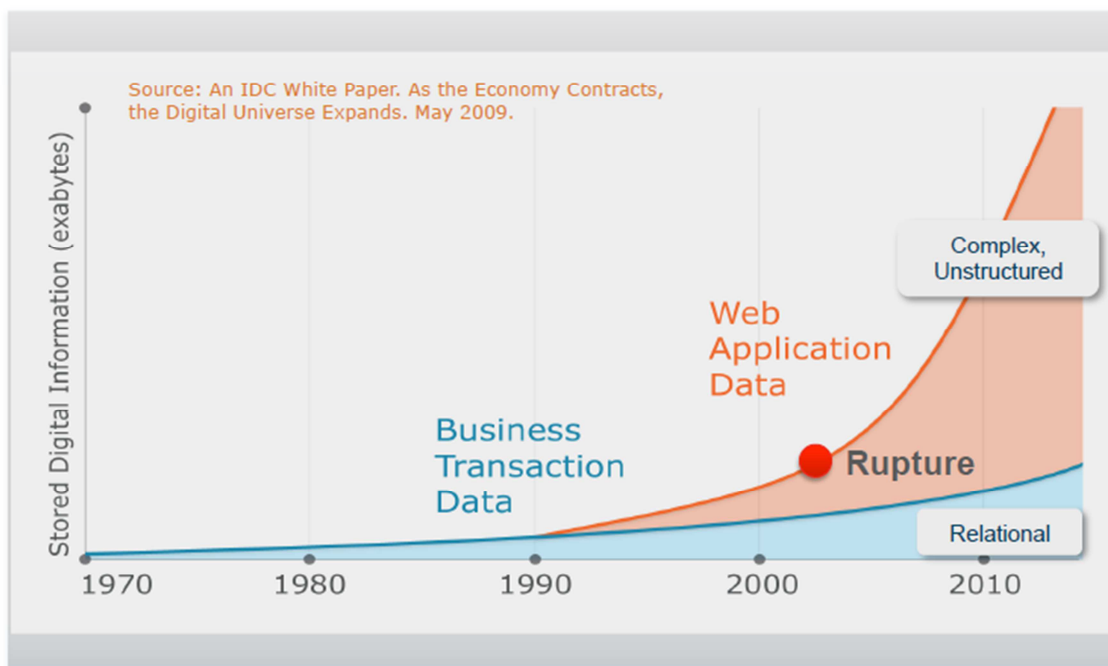


La Big Data est formée de deux grands ensembles (données structurées et non-structurées) qui sont très inégalement exploités par les entreprises.

Si le grand ensemble de la Big Data composé des données structurées est aujourd'hui assez clairement identifié et utilisé par les entreprises, il n'en est pas de même du second ensemble composé des données non structurées. Il est en effet beaucoup moins aisé d'exploiter ce type de data nécessitant des technologies qui, jusqu'à présent, n'étaient pas assez robustes et puissantes pour mener à bien cette lourde tâche. Par conséquent, les entreprises en sont encore aujourd'hui à un état très partiel d'exploitation de la Big Data [13].

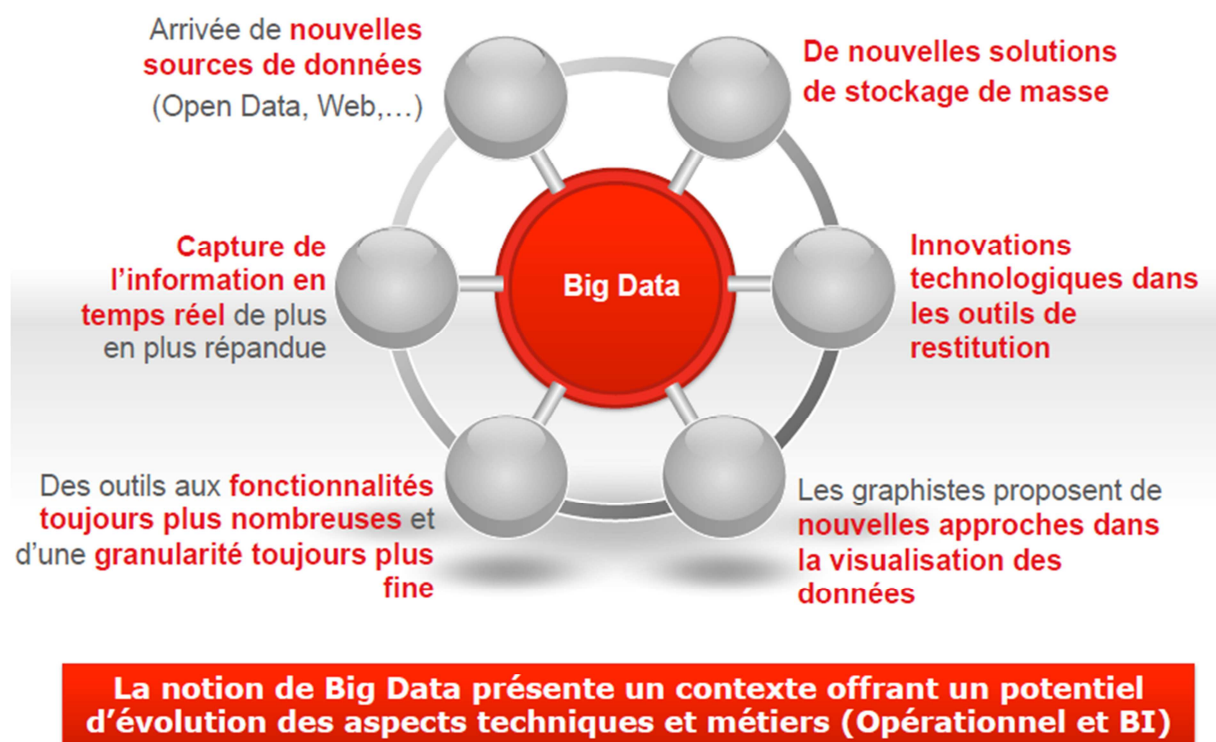
L'essentiel du volume d'informations généré aujourd'hui se fait par des humains (forums, blogs, réseaux sociaux, ...etc.) Mais dans les prochaines années, il sera produit par des capteurs. Caméras de surveillance, sondes météo, cartes bancaires, ...etc. constituent déjà des mines d'informations considérables pour les secteurs concernés. Mises en réseau ou rendues publiques, elles profitent désormais à bien d'autres domaines. La nouveauté, c'est la capacité à croiser toutes les données en provenance des capteurs, du Web et de l'open data (les informations mises à disposition par les pouvoirs publics) [12].

Le Web face à l'explosion

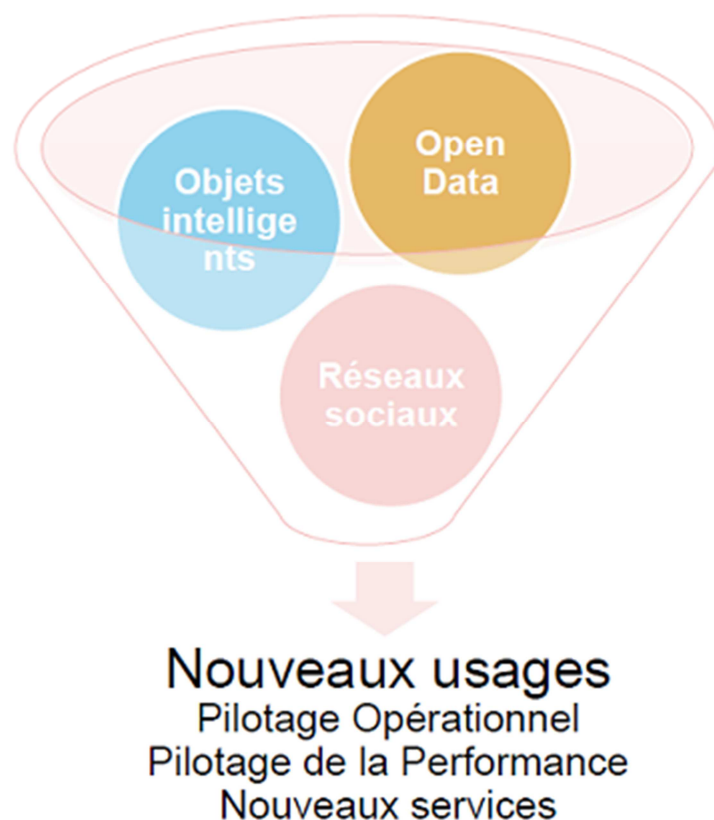


Il y a 10 ans, face à l'explosion du volume des données qu'ils généraient, les leaders du Web ont du trouver des solutions technologiques innovantes

Le Big Data : Une notion contextuelle



Nouvelles sources de données



Apporter de nouvelles possibilités



La capacité de travailler sur toutes les données détaillées permet de nouvelles possibilités d'analyse.

Voir les choses autrement



La prise de conscience d'éléments par leur mise en évidence déclenche de nouveaux comportements

II.2. Définition du Big Data :

Big Data est avant tout un terme marketing, il recouvre un ensemble de concepts et d'idées pas clairement définis. En se tenant à l'étymologie du terme, Big Data est une expression anglophone qui signifie « **Grosses données** » et qui est utilisée pour désigner des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données. **Or, la problématique des grosses données, n'est pas nouvelle** [14].

En effet, depuis plus de 30 ans, nous sommes confrontés à des volumes importants de données. Bien sûr, cela est difficilement comparable à la déferlante dont nous sommes témoins aujourd'hui mais néanmoins, à y regarder de plus près, les capacités de stockage de l'époque étaient bien différentes et une base de données de plusieurs dizaines de giga-octets pouvait paraître énorme. À court terme, nous nous dirigeons vers des bases de stockage de plusieurs péta-octets de données.

Le gros problème réside donc dans la gestion de la donnée et une fois encore, ce n'est pas nouveau. Cela fait presque dix ans que la problématique de gestion des gros volumes de données se pose dans les métiers de la finance, de l'indexation web et de la recherche scientifique. Pour y répondre, l'approche historique a été celle des offres de **Data Warehouse** (TeraData, Oracle, IBM, EMC ou HP). Ces dernières ont évoluées pour supporter et traiter de plus grandes quantités de données [15].

En somme, le Big Data serait plutôt des besoins et des envies nouvelles émanant de l'idée de mieux utiliser ces données qui commencent à s'entasser dans nos Data Warehouse. Et là, les réseaux sociaux et les moteurs de recherches sont parmi les nombreux facteurs qui ont mis à jour ces besoins. La quantité d'information que l'on peut obtenir directement grâce aux utilisateurs, que ce soit par des statistiques d'utilisation ou de recherches ou encore par des données mises directement à disposition des entreprises, est phénoménale.

Dans ces nouveaux ordres de grandeur, la capture, le stockage, la recherche, le partage, l'analyse et la visualisation des données doivent être redéfinis. Les perspectives du traitement des Big Data sont énormes et pour partie encore insoupçonnées, on évoque souvent de nouvelles possibilités en termes d'exploration de l'information diffusée par les médias, de connaissance et d'évaluation, d'analyse tendancielle et prospective et de gestion des risques (commerciaux, assuranciers, industriels, naturels). Il pourrait aider les entreprises à faciliter la prise de décision, ou créer la différence grâce à l'analyse prédictive et une « expérience client » plus personnalisée et contextualisée [14].

Fondamentalement, le **Big Data** s'approche beaucoup du **Data Mining** dans sa transformation de l'information stockée en information clé pour une utilisation future. Là où le Big Data marque une grosse différence, c'est dans le besoin émanant de ces données clés. Souvent, les entreprises ont une idée de ce qu'elles peuvent tirer de leurs informations, mais ne savent pas les rendre utile. Dans d'autres cas, la question est de savoir si on ne révèle pas de nouvelles informations en établissant des corrélations entre ces ensembles de données.

En résumé, parler de Big Data c'est faire allusion à un ensemble de technologies, d'architectures, d'outils, de procédures, et à tous les moyens mis en œuvre pour permettre à une organisation de rapidement capter, traiter et analyser de larges quantités et contenus hétérogènes, changeants et provenant de sources différentes, et d'en extraire les informations pertinentes, le tout à un coût accessible [16].

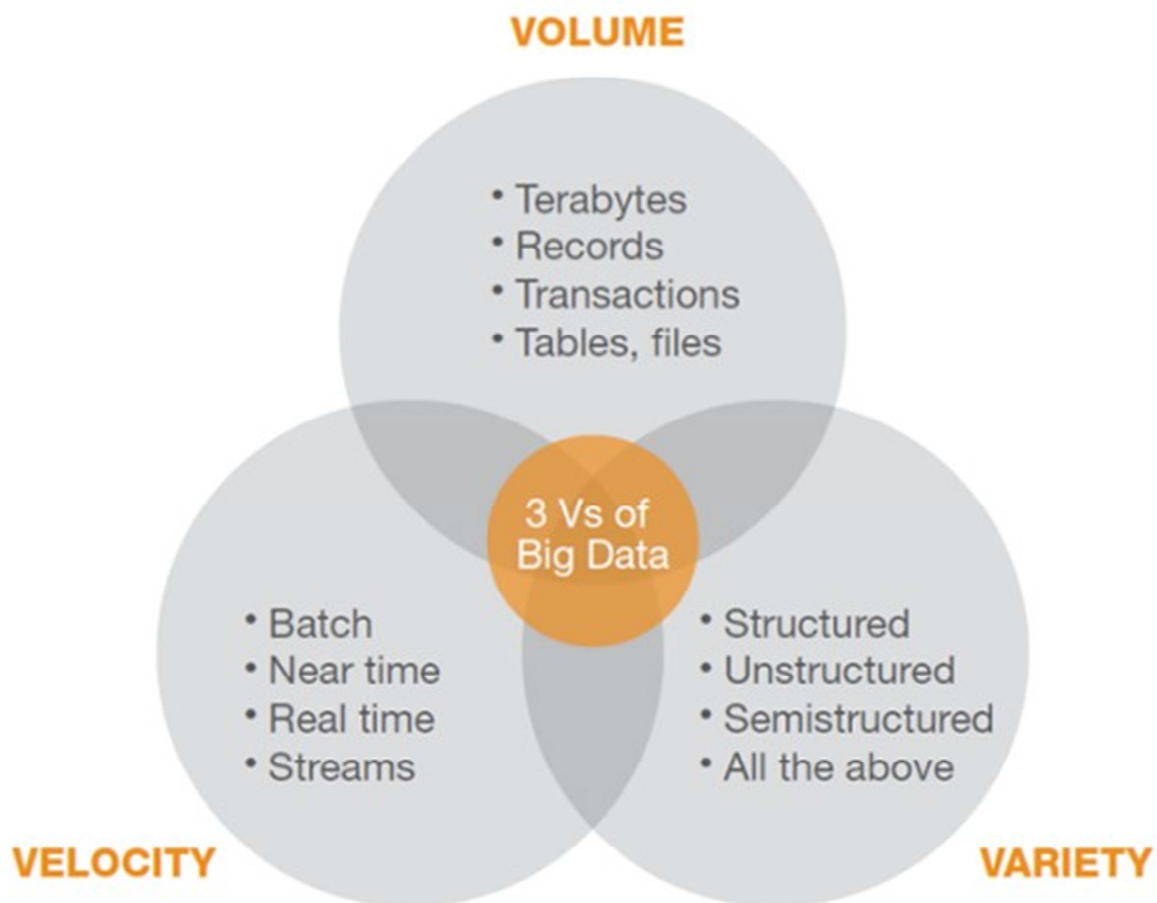


Big Data, le nouveau « Buzz word » et bien plus ...

Le terme « Big Data » a été utilisé pour la première fois en 2008 par l'institut **Gartner** pour faire référence à l'explosion des données numériques. Depuis il a connu un grand succès au point d'être élu mot numérique de l'année 2012, succédant ainsi à « **Cloud Computing** » [17], [18].

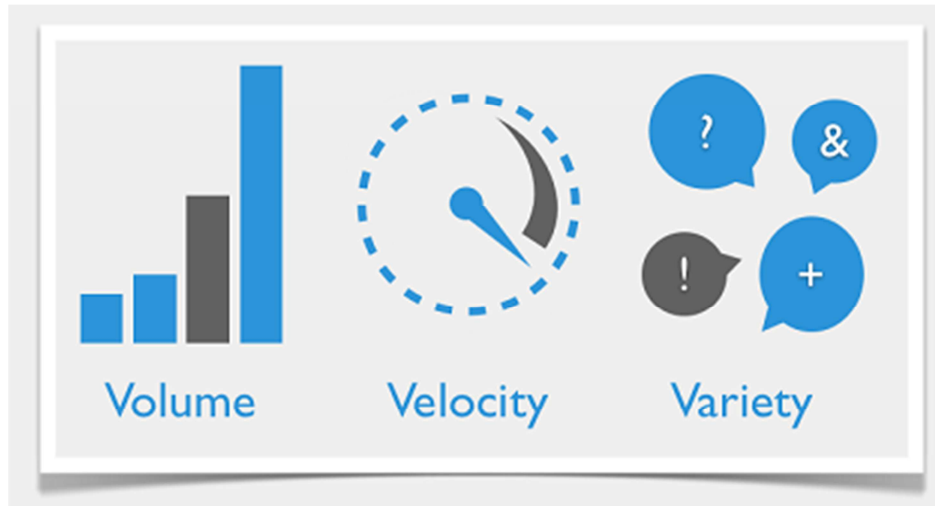
Derrière ce nom se cache un phénomène bien réel annonçant de grandes transformations, avec le Big data nous sommes sur le point de vivre une révolution majeure de notre économie au sens large, appuyée par des évolutions techniques, elles-mêmes alimentées par les progrès constants de la micro-électronique [11].

II.3. Les caractéristiques des Big Data :



Les grandes caractéristiques qui englobent les problématiques auxquelles le Big Data répond sont le **Volume** des données, la **Vitesse** d'acquisition et de traitement des données et la **Variété** des types de données : plus simplement, on parle des **trois « V »** ou les **trois dimensions** du Big Data.

Les premiers secteurs intéressés par le Big Data l'ont été pour tenter de résoudre leurs problématiques de Volume et de Vitesse, des secteurs comme les banques, le milieu des télécommunications ou les marchés financiers. D'autres s'y sont ensuite intéressés pour la capacité à donner de la valeur à des données variées, comme les services publics, le marketing ou la santé. Le fait est que le Big Data représente aujourd'hui un marché important, de plusieurs centaines de millions d'euros, où de nombreux fournisseurs de solutions Big Data existent [15].



II.3.1. Volume :

Le volume est évidemment la première caractéristique qui vient à l'esprit quand on parle de Big Data. Le monde crée de plus en plus de données chaque jour et les entreprises sont submergées de quantités impressionnantes d'informations de tous types, qui se comptent en **téraoctets**, voire en **Pétaoctets**. Suivant une courbe exponentielle, ce volume concerne non seulement les données produites, mais aussi celui des capacités de stockage des supports informatiques [19].

Voici quelques chiffres pour illustrer ce phénomène [16] :

- **90%** des données actuelles ont été créées ces dernières années seulement.
- Le cabinet Gartner prévoit une croissance de **800%** des données à traiter d'ici 5 ans.
- Le réseau social **Twitter** génère plus de **7 To** de données chaque jour.
- Les centres de recherche produisent des centaines de **Téraoctets** de données quotidiennement.



En **2012**, **Google** g rait **20 milliards** de pages web et **24 P taoctets** (24 millions de Gigaoctets) de donn es chaque jour.

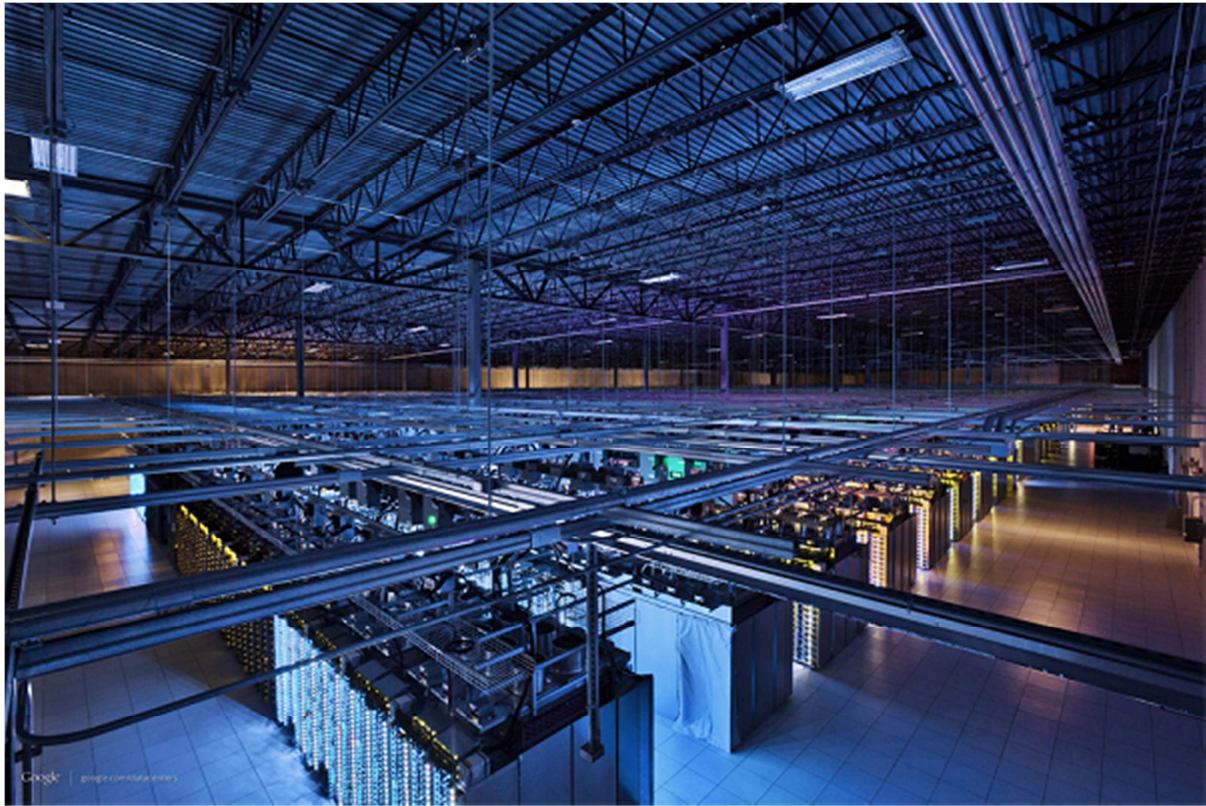
En **2013**, on a d pass  le nombre de **200 milliards** de requ tes par mois sur internet, ce qui repr sente plus de **77000** requ tes par seconde [20].



En 2012, Google a annonc  **4 milliards** de vid os vues chaque jour sur sa plateforme d'h bergement de vid os « **YouTube** » [21].

En 2013, il revendiquait **6 milliards d'heures** de vid os vues chaque mois et quelques **72 heures** de vid os sont mises en ligne chaque minute [22].

Pour gérer ce flux astronomique de données, **Google** a installé **36 Data Centers** à travers le monde, chaque site comprenant plusieurs centaines de serveurs et des dizaines de milliers de disques durs, c'est là que la notion de Big Data prend tout son sens.



Data center de Google de Council Bluffs en Iowa

Aujourd'hui, on ne peut pas parler de Big Data sans parler de réseaux sociaux, puisque ces derniers jouent un rôle important et participent grandement à la production de données dans le monde.



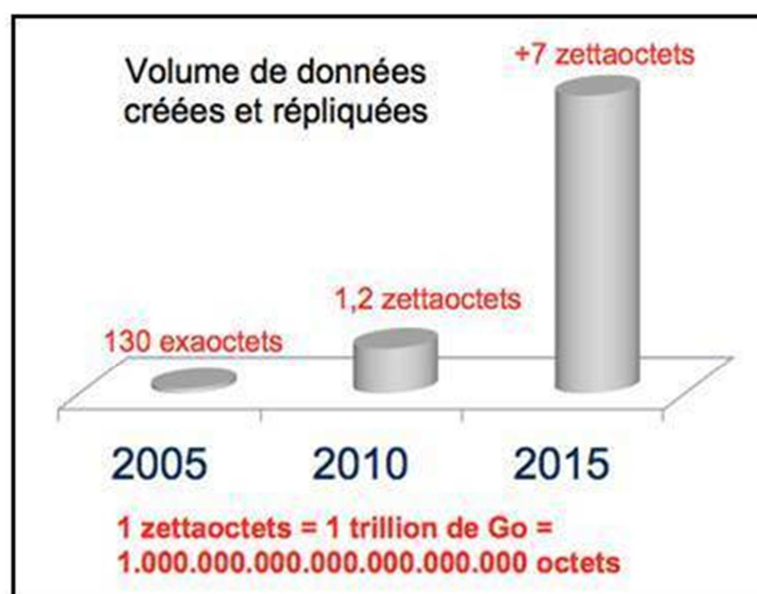
On prend l'exemple de **Facebook**, Leader des réseaux sociaux, il compte en **2014** environ **1,3 milliard** d'utilisateurs.

Facebook est amené à gérer une quantité impressionnante de données venant de ses utilisateurs. Chaque jour, le réseau social enregistre [23] :

- **2.5 milliards** de contenus partagés.
- **2.7 milliards** de mentions « j'aime ».
- **300 millions** de photos envoyées.
- **105 To** de données analysées en 30 min.
- **70.000** requêtes exécutées.

Au Total, Facebook Ingurgite au quotidien plus de **500 To** de nouvelles données.

Explosion de la quantité de données produites



Etude IDC Juin 2011 - The 2011 *Digital Universe* - Study: Extracting Value from Chaos.

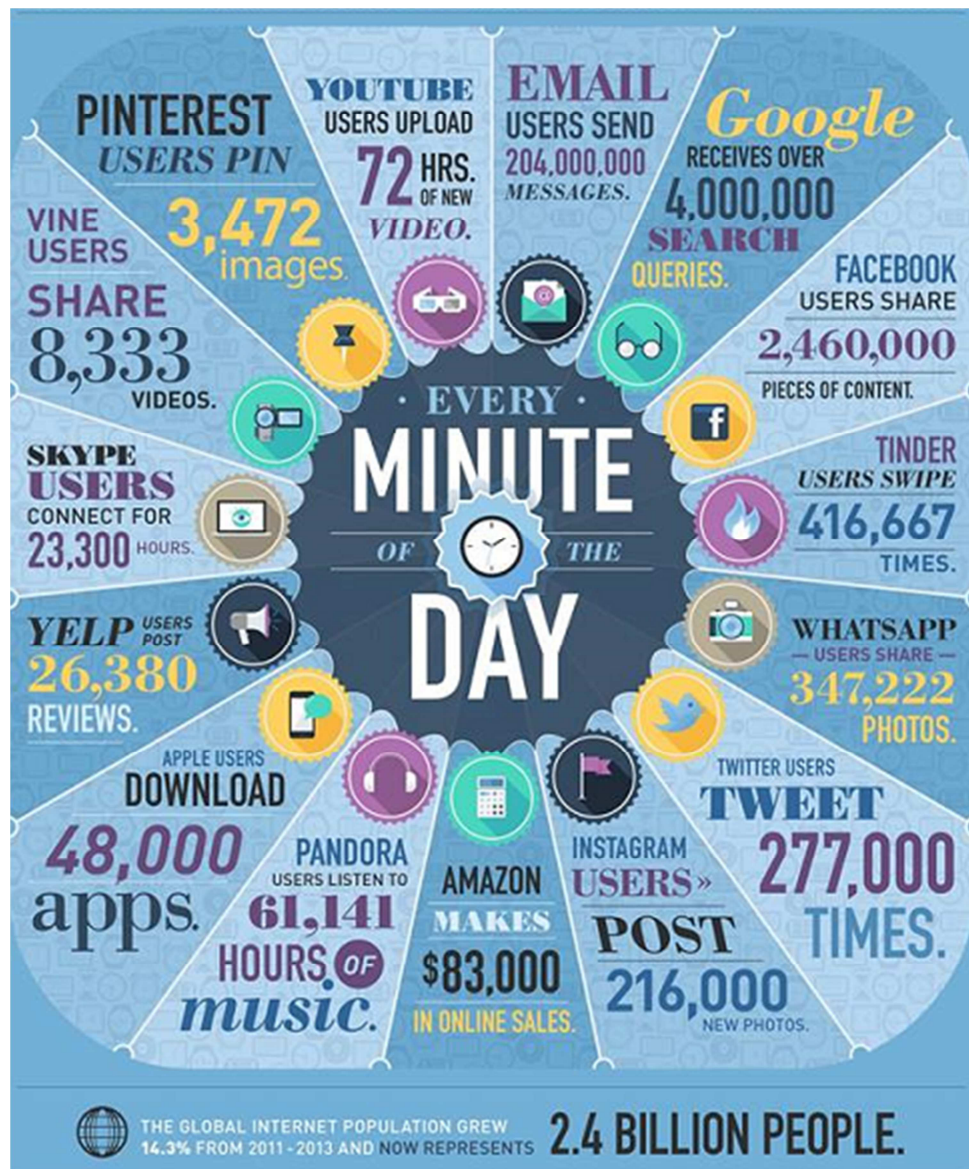
Le volume des données stockées aujourd'hui est en pleine expansion. Selon une étude **IDC** sponsorisée par **EMC Gartner**, les données numériques créées dans le monde seraient passées de **1,2 zettaoctets** par an en **2010** à **1,8 zettaoctets** en **2011**, puis **2,8 zettaoctets** en **2012** et s'élèveront à **40 zettaoctets** en **2020** [24].

Selon l'**IDC**, l'année prochaine, le monde générera **7,91 Zo** (7910 milliards de Gigaoctets) d'informations en une année.

Infographie mettant en avant la quantité énorme de données créée chaque minute dans la période 2010 - 2011



Infographie mettant en avant la quantité énorme de données créée chaque minute dans la période 2011 - 2013



La différence est flagrante, en à peine 2 ans les données ont augmenté de manière exponentielle, chez **Google** on passe de 2 millions de requêtes (2011) à 4 millions en 2013 soit le double, le nombre de contenus partagés sur les réseaux sociaux (facebook) a carrément Quadruplé passant ainsi de **684,478** à **2,460,000** ... etc.

Partant de ce constat, les outils actuels de gestion de bases de données et de traitement des données ne peuvent supporter une telle explosion de volume. Ils ne peuvent répondre aux nouveaux besoins dans des délais raisonnables et à des coûts raisonnables.

Les volumes de données à traiter et à stocker sont de plus en plus importants et peuvent rapidement devenir un handicap pour les entreprises (coûts exponentiels, incapacité d'analyser et d'accéder à toutes ces données dans un délai raisonnable, retard par rapport aux entreprises concurrentes). Mais Ils peuvent devenir une force par la mise en place de solutions capables de gérer et de valoriser ces volumes de données.

II.3.2. Variété :



Le volume des Big Data met les data centers devant un réel défi : **la variété des données**. Il ne s'agit pas de données relationnelles traditionnelles, ces données sont brutes, semi-structurées voire non structurées. Ce sont des données complexes provenant du web, au format texte et images. Elles peuvent être publiques (Open Data) ou relever de la propriété des consommateurs. Les données peuvent provenir de n'importe où, ce qui les rend difficilement utilisables avec les outils traditionnels [25].

La montée en puissance des données non structurées va de pair avec une diversification des formats et des types de données. L'entreprise doit donner du sens aux avis et propositions émis sur les réseaux sociaux, aux images, aux sons, aux vidéos, mais aussi aux informations émises par les terminaux mobiles ou issues des interactions M2M (Machine To Machine).

Dans le domaine du commerce, des solutions Big Data permettront de relier les données non structurées émises par un client (comportement, intonations de la voix, ...) aux données classiques enregistrées à son sujet (historique des achats, service après-vente, ...) pour développer en temps réel une offre adaptée à ses besoins [19].

TerraEchos, leader mondial en solutions de sécurité en temps réel, a ainsi développé un système très sophistiqué pour classer les sons. Celui-ci permet un contrôle intelligent et en direct pour un périmètre donné, équipé de milliers de capteurs. Ceux-ci recueillent les sons détectés qui sont organisés et analysés pour déclencher des actions appropriées (envoi d'un garde, activation d'une caméra, ...) sur une base de modèles à la fois prédéfinis et évolutifs.

Les données représentent la matière première de l'entreprise mais avant de les exploiter, il est intéressant d'identifier les sources de production des données.

On peut regrouper les sources de données en plusieurs catégories [19], [20] :

- **Les applications et services professionnels :**

Il s'agit de logiciels de gestion (Customer Relationship Management, Enterprise Resource Planning, Supply Chain Management,... etc.), des outils de production de contenu, des suites bureautiques telles que **MS-Office** et autre suites **Adobe** ...etc.

Même si ces outils sont connus et largement maîtrisés par les entreprises, Microsoft a reconnu en son temps que la moitié des contenus produits via la suite Office échappent à tout contrôle et ne sont donc pas valorisés. Ce phénomène a connu un nouveau rebond avec l'irruption du courrier électronique. Ainsi, 200 millions d'e-mails sont envoyés chaque minute dans le monde.

- **Le Web :**

Internet et ses volumes gigantesques de données stockées dans les sites d'actualités, scientifiques, gouvernementaux, de commerce en ligne ou issus du monde associatif. En investissant le Web, les entreprises et organisations ont créé un volume considérable d'informations et suscité des interactions toujours plus nombreuses, rendant nécessaires le développement des annuaires et moteurs de recherche, ces derniers créant eux-mêmes d'innombrables données grâce aux requêtes des internautes.

- **Les médias sociaux :**

Bien qu'accessibles via Internet, ils constituent une source distincte de données car ils proposent aux internautes de nouveaux outils d'expression (Crowdsourcing).

Le **Web 2.0** en particulier permet à chacun d'interagir et de collaborer sur internet dans le but de produire du contenu pour une communauté virtuelle.

C'est valable pour les grands réseaux sociaux tels que **Facebook** ou **YouTube** mais également pour les sites de partage tels **Flickr** ou **Instagram**, les blogs, les flux RSS, les réseaux professionnels tels **LinkedIn**, **Yammer**, ... etc. Chaque seconde se sont des mégabytes de données qui sont publiés sur ces sites depuis pratiquement tous les points du globe.

- **Le mobile :**

Pour une entreprise analysant les Big Data, un Smartphone disposant d'une connexion à internet n'est pas un terminal mais une source de données, Il y a 4 fois plus de téléphones mobiles en usage que de PC et Tablettes.

Aujourd'hui, Un utilisateur mobile établit 150 connexions par jour via son Smartphone pour consulter ses messages, se connecter aux réseaux sociaux et autres services en ligne.

Combiné aux médias sociaux et aux services de Cloud Computing, le mobile s'est imposé comme le premier média personnel de masse. À la fin 2013, **l'App Store** et **Google Play** ont dépassé les **50 milliards** d'applications téléchargées.

- **Les objets :**

Le mobile a ouvert la voie à l'Internet des Objets (IOT). Chaque objet du quotidien, équipé de capteurs, dans nos maisons ou dans l'industrie, est désormais un terminal digital potentiel, capturant et émettant des données en permanence.

Le géant industriel **General Electric** met en place des capteurs intelligents sur la plupart de ses produits, des équipements électriques de base aux turbines en passant par des scanners médicaux. Les données opérationnelles recueillies sont analysées pour permettre d'améliorer les services, d'en développer de nouveaux ou de minimiser les temps d'arrêts. Grâce aux analytics, le fonctionnement des machines fait désormais l'objet d'adaptations en temps réel. Le Big Data ouvre ainsi les portes de l'Internet de l'industrie.

II.3.3. Vitesse (Vélocité) :

La vitesse représente le temps nécessaire pour que les données soient collectées, traitées et activées par l'entreprise. "**Le monde digital est désormais plus rapide que le monde réel**" et les données n'échappent pas à cette tendance. Elles sont produites, capturées, traitées, et partagées à une vitesse inédite [19].

Les données non structurées et massives, affluent de plus en plus rapidement dans les systèmes d'information des entreprises. Le mode de traitement par lots qui est au coeur de nombreux systèmes et qui produit des données « rafraîchies » à fréquence régulière, semble donc désormais inadapté à la prise en compte et à l'exploitation rapide du flot constant d'informations [26].

Dans le marketing web par exemple, le phénomène d'enchères en temps réel (Real Time Bidding - RTB), s'appuie sur de la data en mouvement pour proposer une publicité spécifique en fonction de l'utilisateur qui se connecte au site.

La collecte et le partage des données en temps réel devient donc un prérequis absolu dans une approche Big Data.



Il n'y a d'ailleurs qu'à voir comment **Google Analytics** est en mesure de vous indiquer en temps réel le nombre de personnes qui consultent votre site, leur provenance géographique, les pages en cours de consultation et bien d'autres paramètres instantanés pour se rendre compte de l'ère de l'immédiateté dans laquelle nous sommes entrés, en imaginant toutes les exigences techniques et organisationnelles sous-jacentes [26].

Quand on pense que ces informations sont gérées en temps réel pour des millions de sites à travers le monde, on ne peut qu'être impressionnés par la performance.

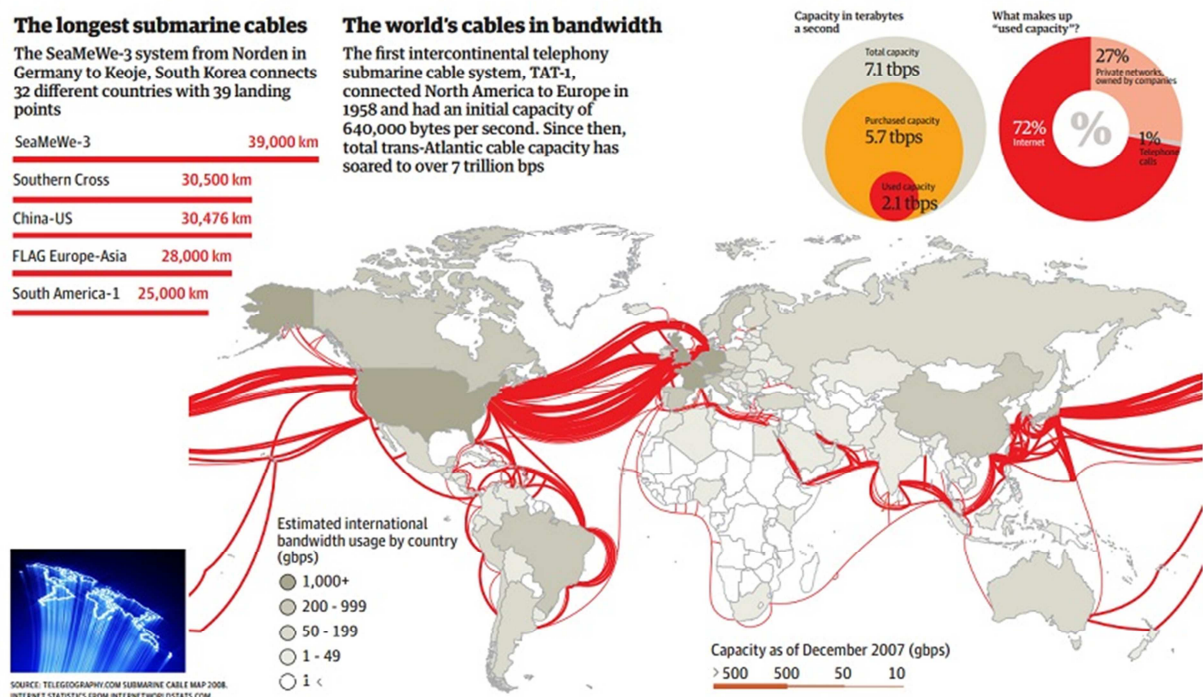
Trafic sur les réseaux numériques :

Pour satisfaire les entreprises qui gèrent d'énormes quantités de données dans leurs Data Centers et les clients qui utilisent de plus en plus internet, notamment les services multimédias et Cloud Computing, l'infrastructure des télécommunications a dû être adaptée aux nouvelles technologies. Les lignes coaxiales (cuivre) ont été remplacées par de la fibre optique, les modems asynchrones par des routers et les ordinateurs monoprocesseurs par des systèmes multi-cœurs équipés de cartes graphiques à haute performance [20].

Avec l'évolution de l'informatique et des télécommunications, en moyenne, le nombre de connexions à hauts débits double tous les 2 ans. Ce taux est plus lent chez les particuliers. Mais un jour ou l'autre tout le monde sera contraint de convertir son installation à la fibre optique car le cuivre ne sera plus supporté dans les années à venir.

Ce progrès étant continu, on constate que la bande passante d'Internet (le débit de transfert des données) augmente de manière exponentielle chaque année.

Sur les câbles de télécommunications en fibre optique installés à travers le monde, le débit des données en **2012** était de l'ordre de **125 Térabits/s** dont **82.4%** concernent le trafic Internet, **17.4%** les réseaux privés (notamment d'entreprises) et **0.2%** le téléphone.

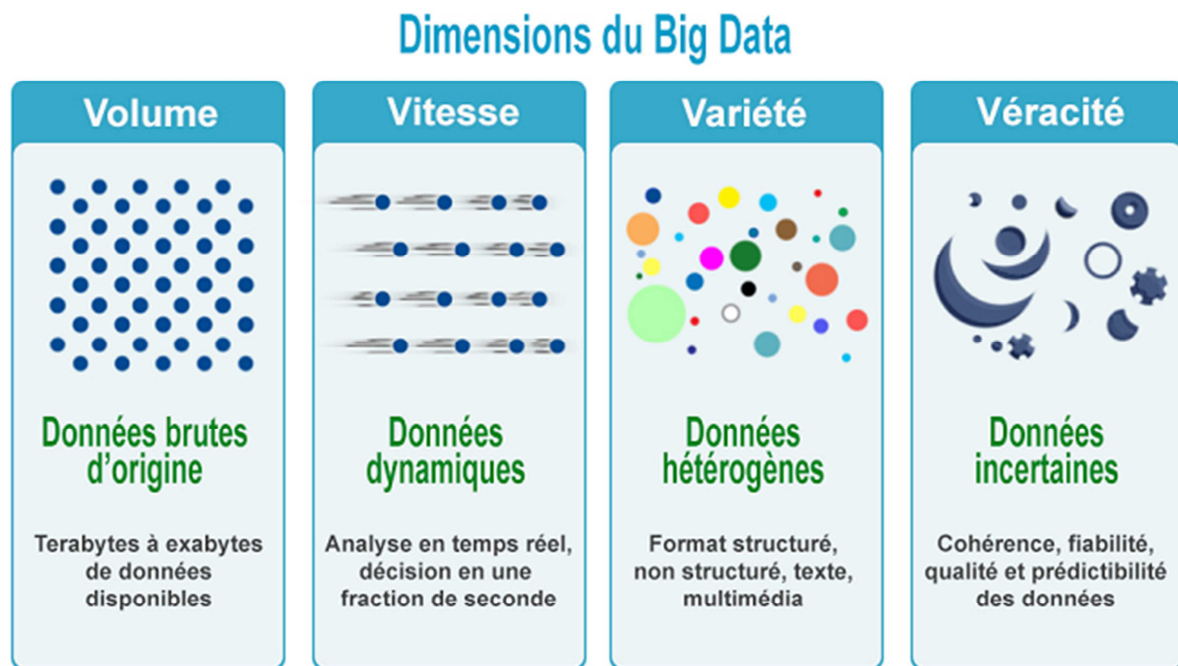


Vue globale du volume de données numériques transitant à travers le monde par fibre optique

Dans ce flux de 125 Tbits/s, 12.6 Tbits/s soit 10% sont destinés à l'Europe et 20.6 Tbits/s soit 16% sont destinés aux Etats-Unis qui représente le principal destinataire des communications.

Il est communément admis que pour désigner les caractéristiques d'un système Big Data, ce sont les 3 'V' de **V**élocité, **V**ariété et **V**olume qui prédominent. Néanmoins, avec l'arrivée des premières solutions métiers et des premières mises en exploitation de plateformes Big Data, les usages évoluent. Ainsi, L'approche exploratoire et la multiplicité des sources et des types de données implique de bien définir son besoin et de contrôler la **qualité des données**.

C'est pour cela que ce modèle, constitué de 3 attributs, fut complété par **IBM**, qui ajouta un 4^{ème} V, celui de la « **V**éracité ».



II.3.4. Véracité :

La Véracité traduit la mise en exergue du besoin de disposer de données fiables et de qualité dans un système Big Data. Comme la rumeur propage de fausses informations, de fausses données génèrent de faux résultats. Pour une entreprise, Il est donc stratégique pour sa pérennité et sa réputation que les données proviennent de sources fiables. Cette dimension est essentielle car de plus en plus de données proviennent de sources extérieures [20], [27].

En effet, Si les entreprises ont opéré de réelles avancées dans la structuration et l'analyse des données internes, une grande partie des données du Big Data provient de sources en dehors de notre contrôle, et doivent donc être triées et mises en cohérence avant de pouvoir être analysées et croisées avec des données internes.

Ces informations externes qu'elles soient structurées ou non constituent encore un vaste chantier auquel les entreprises ne sont que très peu préparées. Majoritairement générées par les consommateurs, ces données permettent de retracer leurs interactions avec une marque grâce à internet, aux réseaux sociaux, blogs et autres plateformes de partage : les clients ont maintenant la parole produisant ainsi de plus en plus d'informations non structurées et utiles à l'entreprise.

La qualité des données, dans une situation donnée, dépend de deux facteurs [28] :

- Les données peuvent être incomplètes ou inexacts, ou encore structurées d'une manière qui les rend difficile à analyser. Il faut donc préalablement vérifier la crédibilité de la source et la qualité du contenu.
- Dans l'immense nuage de données à la disposition des entreprises, il n'est pas toujours facile d'anticiper quelles sont les données qui seront utiles à l'entreprise, et la compréhension des « Business Drivers » passe souvent par une approche exploratoire qui permet de mettre en lumière des signaux faibles cachés dans des grands volumes de données. Il est donc important de bien étudier et définir son besoin, pour ne pas dépenser ses ressources à l'analyse de données « inutiles ».

Lorsque l'on traite ce type de données, on aura beau procéder à un tri, il subsistera toujours une part d'incertitude. Pourtant, malgré cette incertitude, ces données non structurées sont une source d'informations précieuses. La nécessité de reconnaître et d'accepter cette incertitude comme une caractéristique du Big data est indispensable. Pour gérer l'incertitude, les analystes ont besoin de créer un contexte autour des données. Une façon d'y parvenir est de combiner les données avec des sources moins fiables [29].

Selon des études, 1 décideur sur 3 ne fait pas confiance aux données sur lesquelles il se base pour prendre ses décisions. Comment s'appuyer sur une information si on n'a pas confiance en elle? Etablir la confiance dans les Big Data représente donc un défi d'autant plus que la variété et le nombre de sources ne cessent d'augmenter [30].

Plus récemment, certains éditeurs ont enrichi la définition du Big Data en ajoutant deux nouveaux attributs qui sont : **la valeur** et **la visibilité**. Désormais le modèle Big Data est défini par les 6 'V'.

II.3.5. Valeur :



L'analyse des données donne du sens aux informations collectées, et si ces dernières prises individuellement n'ont pas grand intérêt, elles peuvent prendre une signification quand on les appréhende globalement et croiser avec certaines autres données, elles peuvent même générer de la valeur selon l'usage qui en est fait et l'importance des données. Ce qui nous conduit à la question : **quelle valeur économique la donnée initiale ou traitée recèle-t-elle ?** C'est tout l'enjeu entourant les Big Data.

Plusieurs pistes sont ouvertes, qu'on pourrait résumer sous ces quatre indicateurs de performance économique [31] :

- Une amélioration du chiffre d'affaires grâce à l'effort de ciblage marketing.
- Une réduction des coûts grâce à une optimisation des plannings et une détection des erreurs.
- Un développement vers des activités innovantes à forte valeur ajoutée (nouveaux usages liés à la géolocalisation et au temps réel).
- Des gains de parts de marché liées à l'avantage concurrentiel d'être le premier détenteur de ces données stratégiques.

Selon une étude publiée par le cabinet IDC, c'est le deuxième indicateur qui suscite le plus grand intérêt aux yeux des managers : la valeur des données se mesurerait donc aux économies générées dans les processus de production et de vente d'une entreprise.

Cette valeur va peu à peu faire la différence pour les entreprises qui savent l'utiliser dans leur marché et ainsi [32] :

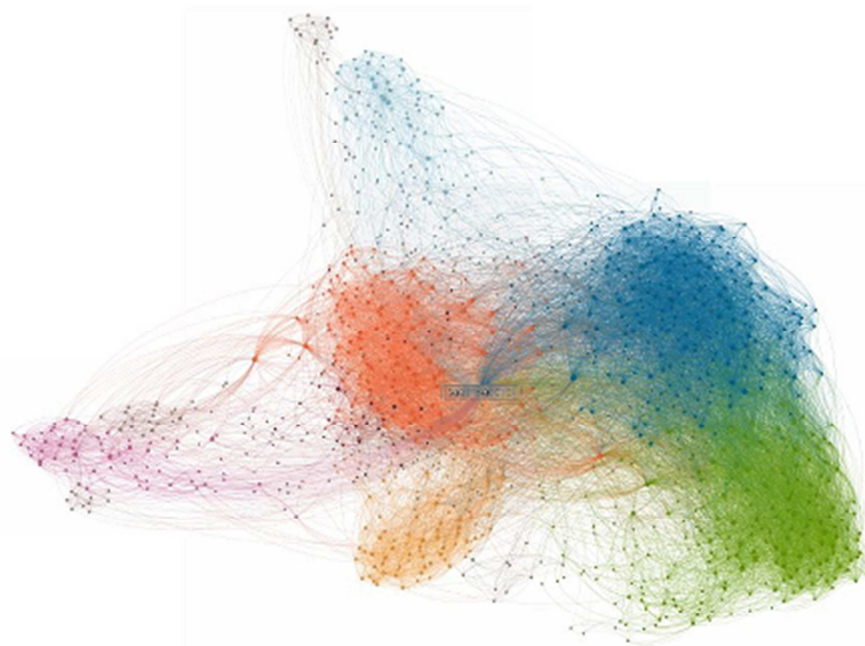
- Analyser la notoriété de sa marque sur les réseaux sociaux.
- Proposer des produits correspondant à tel ou tel profil de visiteur sur un site e-commerce.
- Prédire les flux de commandes en fonction de l'évolution météo.
- Détecter les fraudes de cartes bancaires Etc.

II.3.6. Visibilité :



La Visibilité offerte par une plateforme Big Data est la faculté de comprendre ce que l'on consulte et de trouver les bonnes données facilement. Disposer de données mais ne pas pouvoir s'en servir ou les visualiser est une perte de temps et d'argent. Il est nécessaire que les informations soient factuelles, disponibles et visuellement présentables, c'est-à-dire accessibles rapidement et de la façon la plus conviviale possible afin de les surveiller, de les comprendre quel que soit le support utilisé [20], [27].

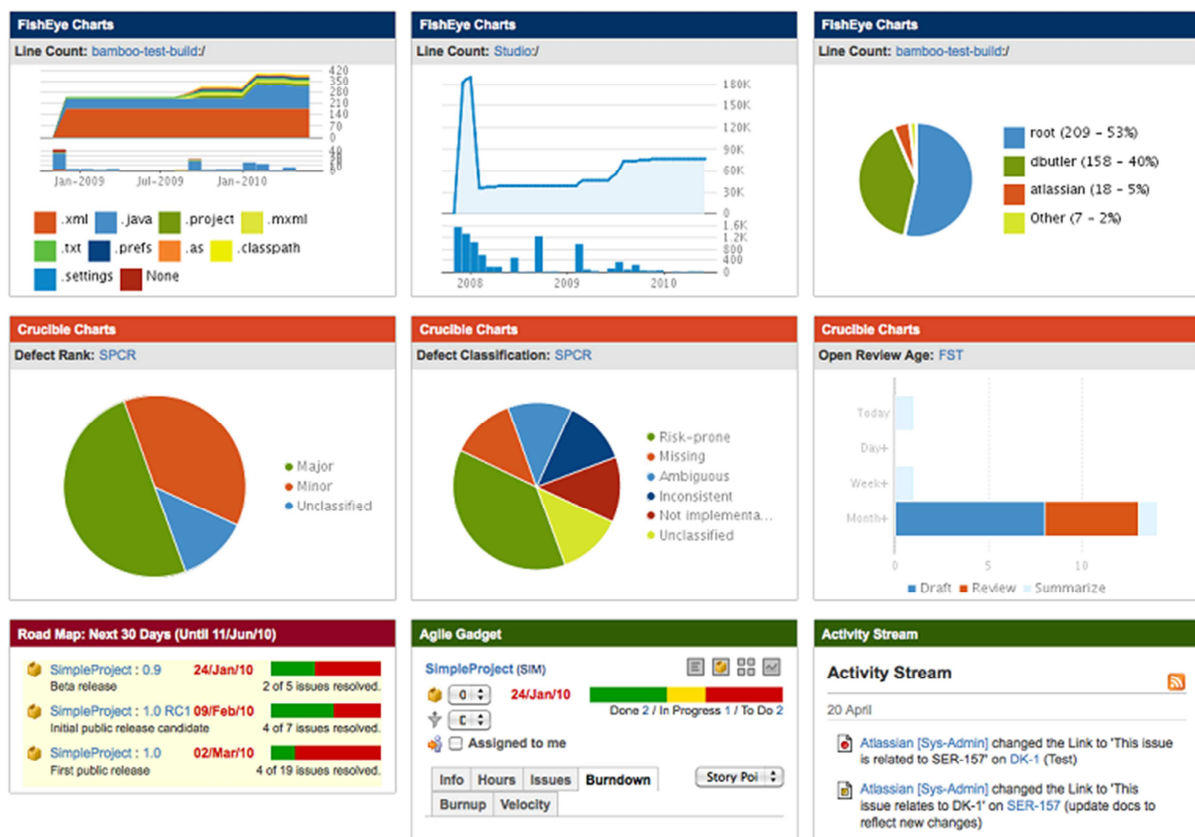
Tout cela est rendu possible grâce au recours à l'art de la **DataVisualisation** et ses techniques innovantes.



La **datavisualisation** concerne la représentation des données sous formes intelligentes, pratiques et interactives.

Il n'y a rien de plus pénible que de devoir lire un texte ou un tableau de nombres si on peut le résumer par un graphique.

Les visualisations dynamiques et intelligentes (**dashboards**) sont généralement utilisées par les managers, chefs de projets et autres dirigeants d'entreprises ainsi que par les équipes marketing pour prendre des décisions et de ce fait ont besoin d'une interface optimisée et intuitive.



Le **Big Data** s'adapte parfaitement à l'expression "une image vaut mille mots". Si l'exploitation des données permet de leur donner du sens, ce sont les images, les diagrammes, les graphiques, les cartes et les infographies qui leur donnent un sens global, révèlent les points forts et les détails.

Les datavisualisations permettent de :

- **Montrer "vraiment" les données :** Là où les tableaux de données sont rapidement ingérables, les diagrammes, graphiques ou cartes permettent une compréhension rapide et aisée des données.
- **Révéler les détails :** La visualisation des données exploite la capacité de la vue humaine à envisager une image dans son ensemble, tout en captant différents détails qui seraient passés inaperçus dans un format textuel ou dans une feuille de calcul.
- **Fournir des réponses rapides :** En éliminant le processus de requête, la visualisation de données réduit le temps nécessaire pour dégager des informations pertinentes pour les entreprises, par exemple au sujet de la fréquentation d'un site Web.
- **Prendre de meilleures décisions :** En permettant de visualiser les modèles, les tendances et les relations issues de l'analyse des données, l'entreprise peut améliorer la qualité de ses décisions.
- **Simplifier les analyses :** Les datavisualisations doivent être interactives. Les outils Webmasters de Google en sont l'illustration. En offrant des fonctionnalités simples et instinctives pour modifier les jeux de données et les critères d'analyse, ces outils libèrent la créativité des utilisateurs.

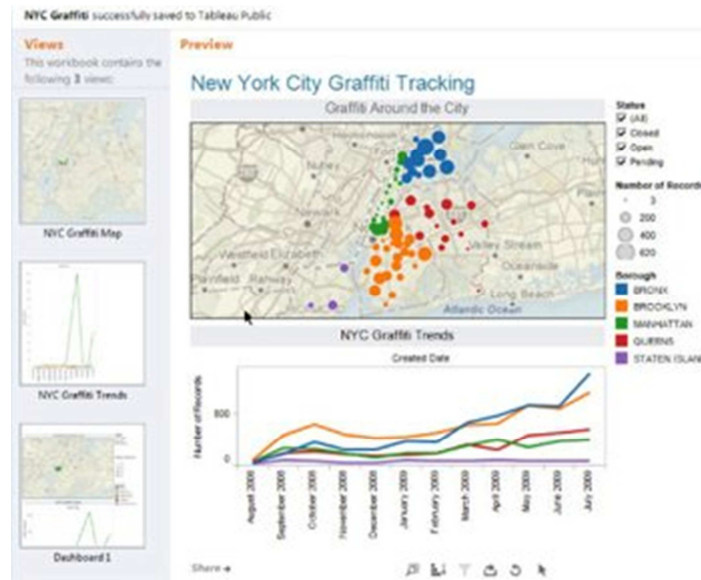
Les éditeurs de logiciels innovants occupent déjà ce terrain en proposant des applications pertinentes, c'est par exemple le cas de **CaptainDash** qui intègre de plus les technologies d'analyse prédictive.

CaptainDash propose un dashboard analytique en mode Saas (Software as A Service) qui combine visualisation des données, design attractif et mobilité et permet aux responsables et directeurs marketing de disposer de vues consolidées de la performance de leurs différentes activités [33].

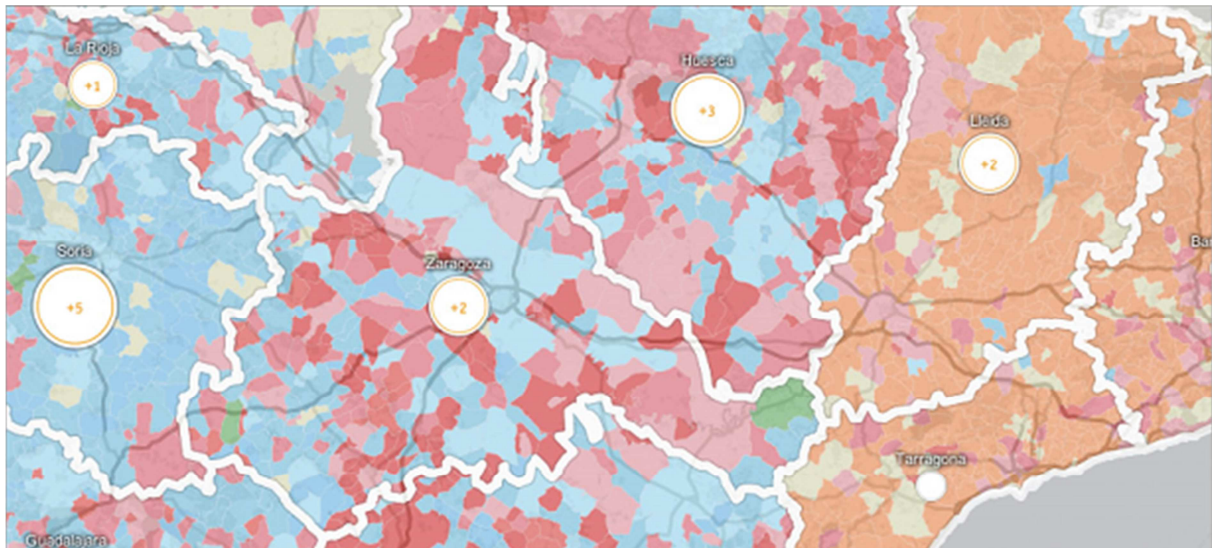
Ce dashboard présente de plus la particularité de ne nécessiter aucune connaissance technique particulière, il implémente les fonctions glisser-déposer des différents éléments pour visualiser, évaluer et mettre en corrélation les flux de données, de plus toutes les informations nécessaires sont accessibles via une interface mobile et véritablement intuitive.



Face à l'accroissement des volumes de données et au défi du Big Data, plusieurs solutions gratuites ont vu le jour, à l'initiative de **Tableau Software**, dans le domaine du social monitoring, qui propose un outil innovant gratuit compatible Smartphones et Tablettes. Il permet de créer des interfaces interactives et personnalisables.



CartoDB, basé également sur le modèle SaaS (Logiciel en tant que Service) est une solution Open Source et gratuite, développé par Vizzuality, qui propose d'associer des données à une carte qui évolue dans le temps et le tout sans aucune connaissance technique, un peu de temps et quelques données suffiront à créer une jolie carte [34].



II.4. Pourquoi Big Data, quels apports pour l'entreprise ?

D'après une étude de **McKinsey**, un distributeur qui utilise le Big Data de manière optimale peut augmenter sa marge opérationnelle de plus de 60% ! Le même rapport indique par exemple que 30% des ventes d'Amazon sont dus aux recommandations données aux internautes.

D'autres peuvent faire de même dans la mesure où chaque entreprise ou organisation a une forte probabilité de disposer aujourd'hui d'un gisement d'informations et d'indicateurs qui pourraient améliorer considérablement ses performances. Certaines de ces données n'existaient tout simplement pas dans le passé ou ne pouvaient pas concrètement être prises en compte et analysées. Les apports marketing, commerciaux et opérationnels d'une analyse des flux de données de type big data peuvent être précieux et divers. Une bonne capture, gestion, modélisation, analyse et exploitation des données issues de ces gisements de Big Data permettent de révéler des relations cachées, détecter des opportunités nouvelles, redéfinir certaines offres, induire des changements organisationnels, infléchir la communication.

Une bonne utilisation du Big Data renforce les performances et la compétitivité de l'entreprise en lui permettant de répondre plus vite et mieux aux besoins de ses clients. Elle permet notamment [35] :

- De remplacer ou d'assister les décisions humaines par des algorithmes automatisés seuls capables de gérer le volume et la vitesse des données.
- D'assurer plus de clarté et de transparence dans les mécanismes de décisions.
- D'expérimenter et d'évaluer plus vite voire instantanément des alternatives marketing.
- D'apporter des capacités prédictives nouvelles.
- D'anticiper plus facilement les mouvements et les tendances.
- De réduire de manière substantielle des coûts d'acquisition et d'exploitation de l'information par rapport aux procédés classiques (dont la recherche marketing).

Exploitation des données

Trouver, interpréter, visualiser, et comprendre toutes les données sont des étapes essentielles pour améliorer la prise de décision dans une entreprise.

L'exploitation des données est un défi auquel fait face toute grande organisation. Les données qui la concernent sont de nature très variable et sont stockées dans des systèmes différents : transactions financières, données sur la qualité de fonctionnement des services, informations sur le fournisseur, satisfaction des clients sur les réseaux sociaux, etc [36].

Les technologies liées au Big Data permettent d'héberger toutes ces données en un même endroit. Cela offre ainsi une image bien plus complète de l'organisation de l'entreprise, et lui permet d'acquérir de nouvelles connaissances en observant et analysant des données jusqu'alors indisponibles.

Plus les collaborateurs de l'organisation ont un accès simple à ces données, mieux ils pourront prendre des décisions importantes, mais aussi réaliser leurs plus simples tâches quotidiennes.

Par Exemple, les **newsletters** de l'entreprise : Une grosse structure peut envoyer des millions d'emails par an. Les intégrer dans ce processus d'exploitation des données permet d'obtenir des indications précieuses sur la performance de ces e-mails : quand les envoyer, à quelle fréquence, quel titre prendre pour objet. Ainsi l'entreprise optimise ses envois de newsletters et offre un meilleur feedback à ses clients.

Mais cela pose un problème, qu'il faut nécessairement résoudre : les sources de Big Data de l'entreprise sont-elles fiables ? Sans ce travail, l'analyse des données sera faussée.

Vision à 360° des clients :

Les entreprises utilisent le Big Data pour comprendre et mieux interagir avec le consommateur. En intégrant des sources d'information internes et externes supplémentaires suite à l'analyse des données, l'organisation bénéficie d'une vue à **360° du client** : ses motivations, les raisons d'un achat, les modalités favorites du geste d'achat, ses prochains achats, son taux de satisfaction, et les facteurs qui le conduisent à recommander une entreprise [36].

Les entreprises de télécommunications intègrent de nouvelles données, issues des appels des clients et de leur comportement sur les réseaux sociaux. Cela leur permet de créer une meilleure image de leurs clients. Certains ont ainsi commencé à prédire le taux de désabonnement et les comportements de fidélité de leur clientèle.

Ces pratiques soulèvent toutefois des difficultés liées à la notion de vie privée, que l'organisation doit régler en amont : quelles informations peuvent être collectées sans exposer l'entreprise à des contentieux juridiques ?

Sécurité de l'entreprise :

Pour une entreprise, le Big Data est un puissant outil de diminution des risques, de détection des fraudes, et de gestion de la sécurité internet en temps réel. L'analyse des transactions de carte de crédit en temps réel permet par exemple de diminuer le taux de fraude en laissant la possibilité d'arrêter des opérations douteuses ou irréalisables.

Le traitement des nouveaux types de données (réseaux sociaux et professionnels, e-mails, etc.) offre une vue beaucoup plus précise de l'environnement de l'entreprise. Elle peut ainsi augmenter et améliorer significativement la sécurité de ses infrastructures Internet et de ses plates-formes d'analyses, et renforcer avec un temps d'avance son cadre législatif en fonction d'évolutions futures qu'elle a déjà pu repérer [36].

Le Big Data n'est à lui seul ni une révolution totale de l'entreprise ni un simple projet informatique. Seules la démarche de transformation adoptée par l'entreprise et la prise de conscience de ses collaborateurs des enjeux stratégiques et décisionnels pourront en faire un outil efficace.

II.4.1. En quoi le Big Data concerne-t-il votre entreprise ?

Depuis la révolution numérique, la masse de données produites chaque jour dans ou en dehors d'un Système d'Information a pris de telles proportions qu'il est difficile de continuer à utiliser les outils traditionnels pour les manipuler de façon performante.

La plupart des entreprises se contentent d'exploiter une part réduite des informations générées et récoltées dans leur Système d'information. Mais comment savoir si votre entreprise est concernée par le Big Data ?

Voici quelques critères qui font qu'on est face à une problématique Big Data [37] :

- Votre Système d'Information est complexe, étendu, composé de plusieurs briques logicielles.
- Vous travaillez sur un secteur où des informations sont produites en temps réel par une multitude de capteurs, d'utilisateurs ou de clients.
- Malgré tous vos efforts d'optimisation, vos bases de données ne suffisent plus pour obtenir des résultats dans des délais raisonnables.
- Vos besoins de puissance de calcul peuvent varier de façon critique (scalabilité).
- Vos données ne sont pas forcément structurées.
- Vous avez besoin d'aller plus loin dans le croisement de vos données et donc de traiter des volumes plus importants.

Si votre entreprise répond à au moins deux de ces critères alors, vous devriez envisager un projet Big Data.

En quoi consiste un projet Big data ?

Ceci représente trois grands défis [37] :

- Le traitement de grands volumes de données, d'où est venu le nom du phénomène, c'est la capacité technologique à effectivement récupérer, centraliser, stocker et requêter ces données dans des environnements sécurisés et à les mettre à disposition de façon utilisable.
- La réflexion métier qui est différente. Puisque les données sont différentes, plus nombreuses et d'origines plus diverses, elles doivent vous permettre d'obtenir des informations inédites. Il serait dommage d'engager un projet Big Data pour obtenir des indicateurs classiques.
- La visualisation d'un grand volume de données est forcément plus complexe. Il faut pouvoir en donner une vue synthétique tout en gardant l'accès à la richesse de l'information disponible. De nouvelles initiatives apparaissent depuis 2010 pour améliorer les techniques et rendus de visualisation dans le Décisionnel, c'est la Datavisualisation.

II.4.2. Comment mettre en place un projet Big Data ?

Une démarche aussi prometteuse mérite qu'on s'y attelle rapidement. Mais comment s'y prendre concrètement ? Lorsqu'une entreprise aborde pour la première fois ce domaine et cherche à exploiter des types de données qu'elle n'a pas l'habitude de prendre en charge et de traiter, de nombreuses questions se posent [38] :

- Quelles sont les données pertinentes et significatives auxquelles on doit s'intéresser ?
- Quels indicateurs peut-on mettre en place ?
- Quels problèmes de qualité de données aurons-nous à régler ?
- Combien cela va nous coûter ?
- Combien de temps cela va nous prendre ?

Avec l'expérience, la réponse à ces questions devient plus facile. Mais comme le big data, par définition, nous soumet sans cesse à de nouvelles variétés de données, toujours plus vite et avec des volumes croissants, on peut être confrontés en permanence à résoudre des problèmes nouveaux.

Le premier projet big data est généralement initié lorsque le management se rend compte que l'entreprise est en train de perdre des opportunités en négligeant les données à sa disposition. La démarche consiste naturellement à solliciter la DSI (Direction des systèmes d'information) pour étudier le projet.

Les équipes marketing et informatiques se lancent alors dans un examen exhaustif des données qu'il serait possible de récupérer et d'exploiter, en se basant parfois sur des objectifs précis. A l'issue de cette phase d'étude préalable qui peut se révéler longue et complexe, le service informatique se lance généralement dans des développements pour répondre au cahier des charges établi, en capturant les données identifiées et en les organisant en vue de leur traitement. Fidèles à leurs approches et méthodes habituelles auxquelles ils ont été formés et qui constituent pour eux le standard à respecter, les informaticiens cherchent à intégrer ces données capturées dans des structures rigoureuses, qui permettront leur exploitation avec les outils de gestion de base de données usuels [38].

Les indicateurs ainsi développés sont ensuite déployés en test puis en production et mis à la disposition des équipes marketing. Dans l'univers du Big Data, la difficulté consiste à utiliser de nouvelles sources de données pour répondre à de nouvelles problématiques, et le tout d'une nouvelle façon, c'est pour cela qu'il est conseillé de définir, pour démarrer, des indicateurs relativement simples qui ne nécessiteraient pas beaucoup de données à collecter ni beaucoup de temps pour les obtenir.

Ainsi, un site de vente en ligne peut commencer par identifier les produits consultés par chaque visiteur dans le but d'envoyer, à ceux qui n'ont pas acheté, des offres promotionnelles qui pourraient les motiver à passer à l'acte. Pour limiter le processus et faciliter l'expérimentation, il est conseillé de ne pas s'intéresser à l'ensemble des données mais de limiter l'expérience à une partie bien délimitée. On peut par exemple prendre les connexions d'un mois donné et pour quelques produits seulement.

Ce focus permet de ne pas s'encombrer avec des volumes de données importants et des informations inutiles. Il facilite la manipulation des fichiers de données par les opérationnels du marketing qui peuvent définir plus facilement des sous-ensembles et effectuer des actions de tests en mesurant les retours obtenus. Des petits projets intuitifs comme celui-ci permettent à l'entreprise de se familiariser avec les possibilités qui s'offrent à elles et de cerner les problèmes éventuels, les efforts financiers et techniques à faire et les retours possibles. On ne démarre donc pas en listant tout le champ des possibles mais plutôt en entrant dans le Big Data pas à pas, à travers de petites initiatives que l'on pourra ensuite développer et étendre en connaissance de cause [38].

II.5. Les défis et enjeux liés aux Big Data :

Avec l'arrivée des technologies Big Data, il est fondamental de se poser plusieurs questions comme celle de l'impact sur le marché ou de la valeur que cela va créer, ou encore comment ces nouvelles technologies vont changer la stratégie, les processus et les performances de l'entreprise ?

II.5.1. Défi technologique :

Le défi le plus important porte sur la refonte intégrale des outils classiques de gestion de bases de données. Ces outils ne permettent, en effet, plus de traiter cette masse de données. Il faut donc repenser les bases de données traditionnelles et leurs langages.

L'un des enjeux du Big Data pour l'entreprise est donc de s'adapter en investissant dans les nouvelles technologies qui touchent au stockage et à l'analyse des données.

- **Investissement en capacité de stockage**

Traiter l'augmentation de stockage avec les systèmes actuels reviendrait à faire exploser les budgets d'investissement informatiques [39].

Même si cela était possible, il y aurait deux autres problèmes à résoudre :

- Celui de la place pour stocker les données car les entrepôts de données de l'entreprise seraient bien vite saturés. Devant le volume de données à traiter, il faudrait ajouter des quantités de disques et baies de stockage supplémentaires et on se retrouverait rapidement en limite de place en mètres carrés.
- De plus, la dépense en énergie électrique nécessaire au fonctionnement des entrepôts de données ajouterait des dépenses aux budgets de fonctionnement des entreprises.

Les nouvelles technologies Big Data permettent l'évolution du stockage vers des systèmes distribués, grâce à l'infrastructure du **Cloud Computing**. Cela correspond à répartir un même fichier sur plusieurs systèmes ce qui permet de rationaliser les volumes de stockage.

- **Investissement dans l'analyse**

La valeur des technologies Big Data provient également des croisements entre sources de données diverses, structurées ou non. L'entreprise doit investir dans ces techniques pour disposer de leur puissance d'analyse.

On distingue ainsi :

- o **Le web sémantique** qui permet d'utiliser toutes les ressources d'Internet et de croiser ces données avec n'importe quelle autre donnée du Web.

Il permet ainsi de simplifier la création de valeur à partir des données analysées. C'est par exemple ce que propose Google dans sa toute nouvelle fonction « **Knowledge Graph** », une immense base de données qui permet d'obtenir des réponses pointues.

- o **Les nouvelles techniques de DataMining** : Des nouvelles techniques sont en train de se développer, celles-ci améliorent les techniques traditionnelles (les associations, les classifications, les algorithmes) :
 - **Le Crowdsourcing** : remontée via les réseaux sociaux de données ou d'informations suite à un appel à contribution,
 - **L'analyse de sentiments** : recherche des opinions à partir de l'analyse textuelle,
 - **L'analyse de réseaux** : techniques permettant d'identifier des leaders d'opinion et leur pouvoir de prescription dans des réseaux sociaux en particulier.

Si la nécessité d'une nouvelle technologie représente le premier défi auquel les entreprises devront faire face, cela ne signifie pas que tous les autres types de données et de technologies soient obsolètes. La question qu'il faut se poser c'est **Comment faire en sorte que la nouvelle technologie de traitement du Big Data utilise les données et les technologies existantes ?** Ou même **Comment améliorer les données et les technologies existantes en y ajoutant le Big Data ?**

L'intégration des données constitue donc un défi de plus pour les gérants et chefs d'entreprises et devra être prise en compte dans la gestion des Big Data.

Les enjeux de l'intégration des données

Les passionnés de Big Data sont conscients des différences entre ce phénomène et les anciennes générations de données.

La difficulté consiste à trouver un référentiel capable de traiter d'énormes volumes de données. L'analyse de flux de données issues de machines, de serveurs et d'appareils mobiles, est problématique [41].

De plus, les données générées automatiquement exigent souvent de nouvelles techniques d'exploration et d'analyse, ce qui représente un défi supplémentaire. La majeure partie du Big Data est non structurée. Par ailleurs, des documents en texte brut et des vidéos s'ajoutent aux types de données. L'apprentissage automatique, l'analyse de textes ou de vidéos et une multitude d'autres techniques permettent de donner du sens à ces données désordonnées.

Du modèle en étoile à chaîne d'approvisionnement de données

L'intégration des Big Data dans le stockage et le traitement des données entraînera un changement conceptuel considérable. L'entrepôt de données ne sera plus au centre de l'univers. De nombreux référentiels spécialisés prendront en charge les applications ou les nouvelles formes d'analyse. En outre, les données proviendront de plus en plus souvent de sources externes à l'entreprise par l'intermédiaire d'API. Au lieu du modèle en étoile, au centre duquel se trouve l'entrepôt de données, l'infrastructure de traitement des données ressemblera davantage à une chaîne d'approvisionnement distribuée [41].

II.5.2. Un enjeu économique :

IDC évalue le taux de croissance des technologies Big Data annuel moyen de 39,4%, ce qui devrait le faire passer, en cinq ans, de 3,2 milliards de dollars en 2010 à 16,9 milliards de dollars en 2015.

Dans une étude récente le cabinet McKinsey qualifie le phénomène Big Data de « **nouvelle frontière de l'innovation de la compétition et de la productivité** » [42].

Cette étude montre que la maîtrise des données (capacité à les analyser) aura un impact sur l'aide que l'informatique va apporter aux métiers pour trouver de nouveaux axes de compétitivité. Par ailleurs, elle précise que l'exploitation du Big Data pourrait permettre :

- D'économiser plus de 250 milliards d'Euros sur l'ensemble du secteur public Européen (identification des fraudes, gestion et mesures de l'efficacité des affectations des subventions et des plans d'investissements, ...).
- Un surplus global de 600 milliards de dollars pour le secteur des sites marchands grâce à l'utilisation des données de géolocalisation (campagne de promotion destinée aux clients à proximité du magasin, ...).

Selon les analystes McKinsey, la gestion des nombreuses données avec les outils actuels aurait déjà un impact sur la productivité, ces dernières années. L'utilisation de façon effective des technologies Big Data, permettrait une augmentation potentielle des marges d'exploitation pouvant atteindre 60% [43].

Par ailleurs, le cabinet indique que les couches du système d'information sont désormais mobilisées et optimisées pour permettre une valorisation métier des données, qu'elles soient structurées ou non. Il prévoit ainsi qu'en 2016, 7% du Chiffres d'Affaires généré par les ventes de serveurs concernera des infrastructures destinées aux Big Data.

Avant, les entreprises travaillaient dans un contexte où la production de la donnée était coûteuse, elles devaient alors en limiter la production. Maintenant, ce problème est révolu car les données sont presque gratuites et abondantes. Ce qui est important, ce n'est plus de produire des données mais plutôt de les exploiter.

L'exploitation de ces données massives dépend de la stratégie et du besoin de l'entreprise. Par l'agrégation et le traitement de ces données, de nouveaux « **business models** » émergent en concevant de nouveaux services.

C'est le cas du marché du **Web Analytics** qui génère déjà plusieurs milliards d'euros sur des outils et des services. Cette technique d'analyse des comportements permet d'élaborer des micro-segmentations des scores d'**appétence** (le calcul qui donne la probabilité qu'un prospect devienne client) ou de **churn** (le calcul qui permet de connaître le taux de perte d'un client) qui sont nécessaires pour personnaliser la relation et les offres proposées au client. C'est le domaine le plus exploité aujourd'hui par les Big Data [44].

En effet, la collecte et l'analyse des informations du parcours client, de son comportement lors d'un achat sur les sites Internet et le temps passé vont permettre :

- Une optimisation des pages web et une amélioration de l'expérience utilisateur en apportant des modifications aux sites pour une meilleure rentabilité,
- L'accélération du développement de nouveaux produits et leur adaptation au plus près des attentes et des usages du client,
- De faire de la veille ou de la prévision en tirant parti des dialogues continus du web social pour surveiller des tendances.

En résumé, l'utilisation du Big Data améliorera la compétitivité d'une entreprise face aux concurrents en lui permettant d'être plus réactive face aux besoins, exigences et intérêts de ses clients.

II.5.3. Un défi organisationnel :



L'intégration des nouvelles technologies Big Data dans les entreprises va entraîner des changements en termes de gouvernance et d'organisation.

Ces changements seront abordés sous deux angles :

II.5.3.1. La conduite du changement :

L'intégration du Big Data dans une entreprise risque d'entraîner une rupture de son organisation : Elle devra modifier sa manière de travailler pour s'adapter à ce nouvel écosystème (arrivée de nouveaux acteurs, renforcement de la concurrence, importance des réseaux sociaux, changement de la relation client, ...) [45].

Pour réussir cette transformation, il est nécessaire d'adopter un changement de mentalité au sein de l'entreprise. C'est toute la réflexion métier qui est différente car les données, en permettant d'obtenir de nouvelles informations, deviennent le pilote de la stratégie de l'entreprise. La priorité réside bien dans la maîtrise de la donnée.

Les décideurs doivent être sensibilisés à la valeur de la donnée. Pour cela il faut arriver à les convaincre de l'intérêt qu'il y a à exploiter ces données, en développant deux axes :

- Le renforcement de l'efficacité des métiers et notamment la relation **Marketing-Client**.
- La capacité, pour les donneurs d'ordre, à tirer des synthèses d'aides à la décision (tendances du marché, ...)

Cela leur permettra d'adapter la stratégie de l'entreprise pour créer de la richesse à partir de ces données. Mais Pour atteindre un degré d'efficacité, il faut un partenariat renforcé entre les équipes **informatique** et **marketing** qui seront les acteurs du Big Data dans l'entreprise.

II.5.3.2. L'apparition de nouveaux métiers « Data Scientist » :



Les entreprises sont obligées de trouver et de mettre en place de nouvelles solutions analytiques complexes dédiées aux Big Data et pour cela, elles doivent investir en outils spécialisés mais aussi en ressources humaines.

Les nouvelles possibilités de collecte, de stockage et d'analyse sur un très important volume de données d'origines variées, ont fait émerger une nouvelle famille de métiers. Celui qui s'impose le plus est sans doute, le métier de « **Data Scientist** », qui aujourd'hui est considéré comme l'un des profils les plus recherchés, dans le secteur de l'informatique et des technologies, mais également les mieux rémunérés [46].

Le Data Scientist, doit utiliser des techniques statistiques et des outils informatiques spécialisés afin d'organiser, de synthétiser et de traduire les informations dont les entreprises ont besoin pour faciliter les prises de décision. Il doit non seulement donner du sens aux données mais aussi de les rendre intelligibles [47].

Selon les experts, un **Data Scientist** doit disposer d'une triple compétence, à savoir [48] :

- La maîtrise des techniques du Data Mining et des Statistiques.
- La maîtrise de l'outil informatique (connaissances en Programmation, Gestion de bases de données, ...etc.).
- Un savoir-faire métier dans le domaine d'application des données étudiées (le Marketing, ... etc.).

Ces spécialistes de la donnée sont très rares car il n'existe pas encore d'écoles proposant des formations académiques pour les former.

II.5.4. Un défi Juridique :

Le phénomène du Big Data désigne la croissance exponentielle du volume des données disponibles sous forme numérique aussi bien dans les entreprises que sur internet. Chaque individu participe au Big Data en dispersant des données sur ses faits et gestes accumulées par les réseaux sociaux, les applications, les mobiles, ... etc.

Du téléchargement d'applications mobiles aux achats sur des sites d'e-commerce, en passant par la recherche d'un travail, l'utilisation de cartes de fidélité en ligne, la géolocalisation dans des boutiques ... jamais les individus n'ont partagé autant d'informations et laissé autant de traces. Les ordinateurs, smartphones et tablettes génèrent une multitude de fichiers et d'informations, stockés dans des centres de données de plus en plus puissants.

Si le Big Data est l'avenir du e-marketing, en revanche les défis juridiques sont nombreux.

Les sites tels que : **Facebook, Google, Twitter, Amazon**, permettent aux internautes d'accéder à des services gratuits en cédant, en échange, des informations personnelles. Ils bâtissent donc leur modèle économique autour de ce trésor des données relatives à des millions d'internautes.

« Si vous utilisez un service gratuit, vous n'êtes plus le client, vous êtes le produit »

Du traitement des données découlent deux préoccupations : La **protection juridique des données traitées et l'usage qui est fait des traitements** [49].

Les données peuvent être personnelles et soumises à la réglementation de protection de ces données laquelle impose notamment d'accomplir certaines formalités et de respecter les droits des personnes concernées.

Titularité des droits sur les données :

Ici, il faut distinguer les données privées des données publiques, qui relèvent de l'open data. Les données privées, soit parce qu'elles sont le produit d'une entreprise ou qu'elles relèvent de la sphère privée des individus, ne peuvent faire l'objet, librement, d'une appropriation par un tiers. Pour éviter cela, Il existe des dispositifs juridiques, de lutte contre la concurrence déloyale, la violation de droits de propriété intellectuelle ou encore de protection de la vie privée, qui sont mis en place.

Quant aux données publiques, leur usage est régi par les dispositions de la licence ouverte qui autorise l'exploitation commerciale des données, y compris en combinaison avec d'autres données et par inclusion dans un produit ou une application [49].

II.6. L'analyse, le point clé du Big Data :



Le Big Data répond à de nombreux objectifs précis parmi lesquels on trouve **l'extraction d'informations utiles** des données stockées, **l'analyse de ces données**, **la restitution efficace des résultats** d'analyse ou encore, l'accroissement de l'interactivité entre utilisateurs et données [15].

La problématique de la Big Data relance celle du traitement des données non structurées

La combinaison de ce déluge d'informations et d'algorithmes logiciels intelligents ouvre la voie à de nouvelles opportunités de business.

L'analyse des Big Data permet tout d'abord de mieux écouter les usagers, de mieux comprendre leurs modes d'utilisation des services et d'affiner les offres. Dans le **domaine des transports**, on modélise les déplacements des populations pour adapter les infrastructures et les services (horaires des trains, etc.) [50].

Les analyses peuvent faciliter la recherche d'emploi. Il s'agit de combiner les qualifications des individus avec les offres d'emploi (issues des sites internet d'entreprises ou d'administrations).

Améliorer les performances gestionnaires

L'analyse de données massives peut accroître la transparence administrative, faciliter l'évaluation des services, assister la prise de décision, ou permettre d'économiser des ressources [51].

Une entreprise peut, par exemple, suivre ses ventes en temps réel pour mieux réapprovisionner ses stocks. De même, une administration publique peut suivre l'activité des agents, le versement de prestations, l'accroissement des demandes, etc.

Prédire et prévenir

L'analyse des masses de données permet plus spécifiquement d'anticiper, avec un certain degré de certitude, des comportements ou des besoins [52].

Dans le domaine de la santé par exemple, il est possible de mieux prévenir certaines maladies ou épidémies, ou d'améliorer le traitement des patients.

La prévention des crimes est l'une des applications possibles de l'analyse des masses de données. La police peut par exemple identifier les zones et les heures où des délits sont le plus à même d'avoir lieu, afin d'optimiser l'affectation des services.

II.6.1. Ce qu'est l'analyse Big data et ce qu'elle n'est pas :

L'analyse Big Data est [54] :

- Une stratégie technologique qui permet d'obtenir des informations plus riches et approfondies sur les clients, les partenaires et le marché – et au bout du compte bénéficier d'un avantage concurrentiel.
- Travailler avec des ensembles de données dont la taille et la diversité dépassent les capacités de capture, de stockage, de gestion et d'analyse des logiciels de base de données classiques.
- Le traitement en temps réel d'un flux continu de données afin de prendre plus rapidement des décisions urgentes.
- Des ressources décentralisées. Les processus d'analyse se dirigent là où sont les données pour plus de rapidité et d'efficacité.
- Un nouveau paradigme dans lequel l'informatique collabore avec les acteurs commerciaux et les « **Data Scientists** » afin d'identifier et mettre en œuvre des analyses qui permettront d'accroître l'efficacité opérationnelle et de résoudre de nouvelles problématiques commerciales.
- Déplacer la prise de décision vers la base de l'entreprise et permettre aux personnels de prendre de meilleures décisions, plus rapidement et en temps réel.

L'analyse Big Data n'est pas :

- Uniquement une question technologique. Au niveau des entreprises, il s'agit de déterminer comment exploiter des sources de données considérablement enrichies afin de développer de nouvelles perspectives.
- Qu'une question de volume. Il faut également prendre en compte la diversité et la vitesse. Mais le plus important est sans doute la valeur ajoutée issue des données.
- Générée ou utilisée que par de grandes entreprises en ligne comme Google ou Amazon. Même si les entreprises d'Internet ont été pionnières dans l'utilisation de Big Data à l'échelle du web, cette technique peut s'appliquer à tous les secteurs.
- Conçue pour fonctionner avec des bases de données relationnelles classiques et de taille standard, construites sur une architecture de disque et de mémoire partagés. L'analyse de Big Data utilise un réseau de ressources informatiques pour un traitement massivement parallèle (MPP).
- Destinée à remplacer les bases de données relationnelles ou l'entrepôt de données. Les données structurées continuent d'être d'une importance capitale pour les entreprises. Cependant, les systèmes traditionnels peuvent ne pas convenir pour les nouvelles sources et les nouveaux environnements du Big Data.

II.7. Quand le Big data rencontre le Cloud :



Le **Cloud Computing** (ou informatique en nuage) et le **Big Data** induisent une transformation majeure dans l'usage du numérique par les entreprises de tous les secteurs économiques. Les enjeux associés portent non seulement sur l'activité et la création d'emplois au sein des acteurs du numérique, mais aussi sur les gains de compétitivités pouvant être réalisés par les entreprises utilisatrices [55].

Le modèle de l'informatique en nuage consiste à pourvoir des ressources informatiques à distance et à la demande, qu'il s'agisse d'infrastructure, de plates-formes ou de logiciels d'application. Les avantages en termes de réduction de coûts et de facilité d'accès portent cette adoption rapide, qui se traduit par une mutation progressive mais déterminante des systèmes d'informations, des activités et des marchés qui s'y rapportent.

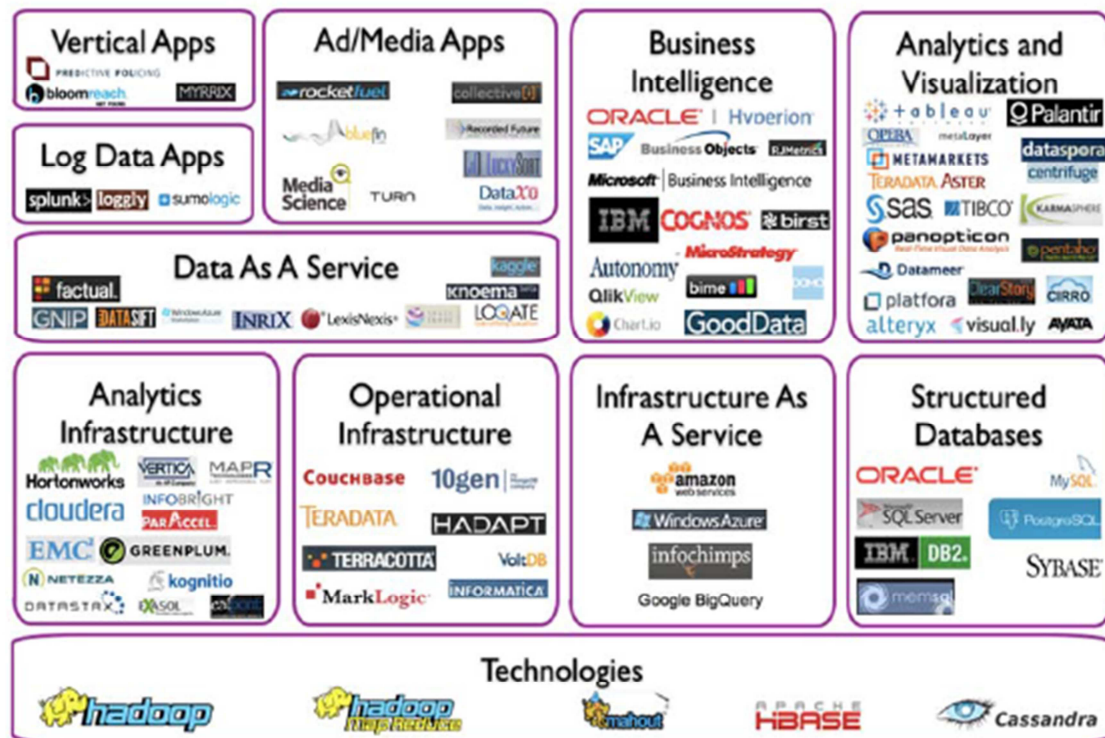
Le **Cloud Computing** facilite le mouvement vers le **Big Data**, lié au besoin de grande capacité de calcul et de stockage des flux de données issus de l'usage accru des nouvelles technologies numériques. Les entreprises doivent ainsi continuer à gérer une augmentation exponentielle du volume de données produites (structurées, semi-structurées ou non structurées) et les analyser le plus rapidement possible pour essayer d'en tirer de la valeur. Les outils associés représentent un champ en fort développement, offrant de nombreuses opportunités de création de valeur.

L'émergence du Big Data, couplé au développement des objets intelligents et connectés, accroît ainsi les besoins en calcul intensif, nécessaires pour l'analyse de certains flux de données. En outre, le recours à la simulation permet de concevoir des objets toujours plus intelligents et connectés, qui bénéficient des puissances de traitement du cloud computing et accèdent aux données de l'Internet.

Les fournisseurs de services cloud « **Cloud Provider** » lancent de plus en plus d'offres pour répondre à la demande du marché. Leurs développements sont cependant freinés par des difficultés identifiées notamment en matière d'interopérabilité et de passerelles entre la multitude d'offres proposées aux entreprises. Des questions se posent également sur les problématiques de standardisation, d'orchestration, de gestion et d'engagements de service pour « unifier » les différents services. L'adaptation des services à ces besoins spécifiques représente un potentiel important pour une adoption plus massive des entreprises déjà utilisatrices et l'atteinte de segments de marché encore hésitants dans leur migration vers l'informatique en nuage [56].

La montée en puissance du Big Data repose quant à elle sur la maturité de bon nombre de technologies. Disposer de la faculté d'extraire, stocker, visualiser, traiter et interpréter les données, hétérogènes dans leur forme, en grand volume et en temps contraint, demeure un défi technologique et un enjeu fort pour les acteurs du domaine numérique. De plus, le besoin d'agréger d'énormes volumes de données et d'opérer sur celles-ci des traitements jusque-là hors de portée doit pouvoir être proposé aux entreprises de toutes tailles.

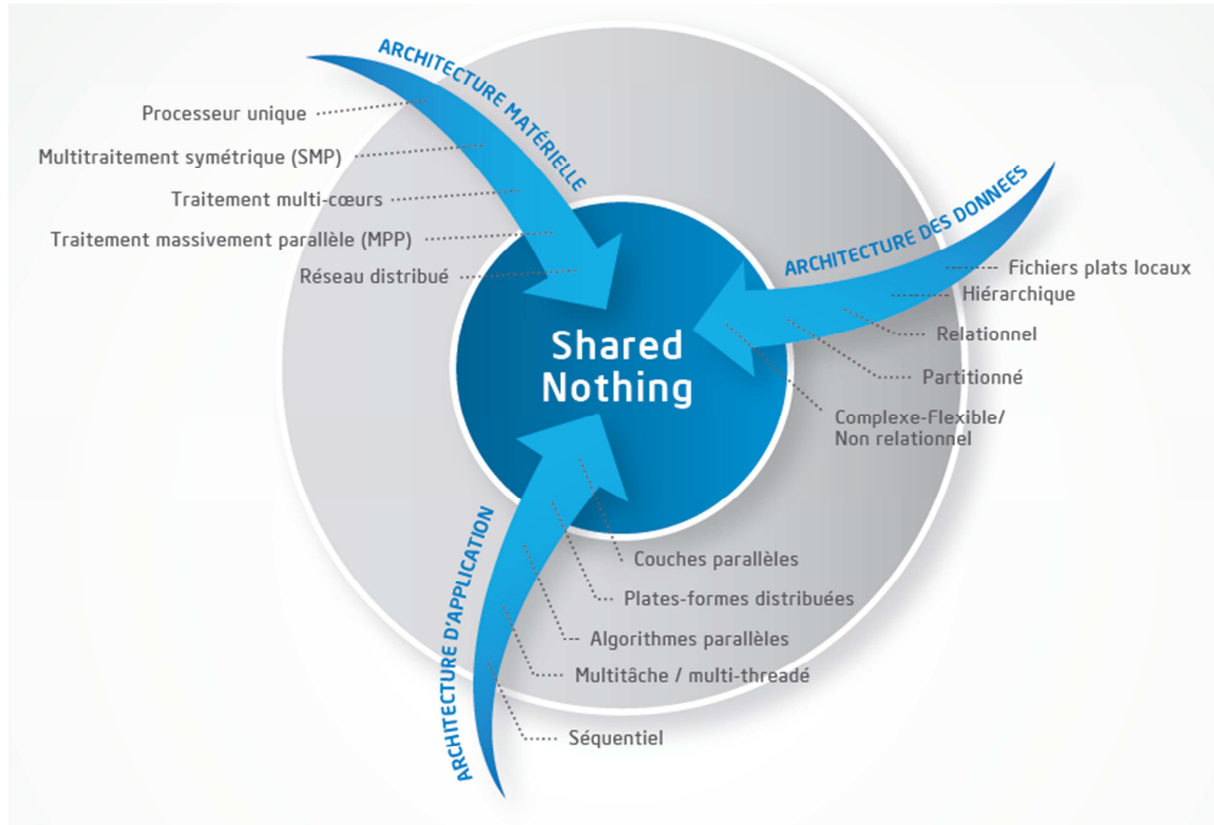
Chapitre III : Les technologies émergentes pour la Gestion des Big Data



Pour que les entreprises prennent la pleine mesure du potentiel du Big Data, elles doivent trouver une nouvelle approche pour la capture, le stockage et l'analyse des données. Les outils et infrastructures traditionnels ne sont pas aussi efficaces pour travailler avec des données plus importantes, plus variées et arrivant à grande vitesse.

En réponse au phénomène Big Data, une famille d'outils et de technologies ont vu le jour, dont **NoSQL** (Google, Amazon, LinkedIn, Facebook), le SGBD **BigTable** (Google), les frameworks **MapReduce** (Google) et **Hadoop** (Facebook, Twitter), certains services de Cloud Computing et systèmes de traitement massivement distribués. Les acteurs traditionnels des bases de données (Oracle, IBM, Microsoft, etc.) et du stockage (EMC, IBM, etc.) commencent tous à proposer des solutions Big Data [57].

III.1. L'Architecture « Shared Nothing » pour les systèmes évolutifs :



La nouvelle architecture **Shared Nothing** peut évoluer en parallèle avec les gigantesques volumes, la variété, et les exigences de vitesse du Big Data en répartissant le travail sur des dizaines, des centaines, voire des milliers de serveurs de grande diffusion qui traitent les données en parallèle. Chaque nœud est indépendant et sans état, de sorte que l'architecture shared nothing évolue facilement, il suffit d'ajouter un autre nœud permettant aux systèmes de gérer des quantités de données croissantes [58].

Le traitement est poussé vers les nœuds où résident les données. Cela est radicalement différent par rapport à une approche traditionnelle, qui rapatrie les données pour les traiter en un point centralisé. Enfin, les données doivent être réintégrées afin d'apporter des résultats significatifs. Les plates-formes logicielles de traitement distribué permettent aux réseaux de calcul de fonctionner en organisant et en déplaçant les données entre les machines, et en envoyant aux serveurs en réseau des instructions pour travailler en parallèle, collecter les résultats individuels, puis les réassembler pour le résultat final.

III.2. Le framework « Hadoop » :



Depuis un an, l'engouement du marché pour le Big Data se manifeste principalement sur une de ses composantes : **L'analyse de données.**

Le phénomène résulte clairement de la montée en puissance d'**Hadoop**, emblème par excellence des Big Data, il est réputé pour sa puissance d'indexation, de transformation, de recherche ou d'élaboration de modèles sur de très gros volumes de données. Il incarne à lui seul le concept de Big Data [59].

III.2.1. Définition :

Hadoop est un framework open source, implémenté en **Java** et distribué par la fondation **Apache** (Apache software foundation).

Destiné à faciliter la création d'applications distribuées et échelonnables (scalables), Hadoop permet aux applications de travailler avec des milliers de nœuds et des pétaoctets de données [60].

Il a été mis en avant par des grands acteurs du web tels que **Yahoo!** et **Facebook**.

III.2.2. Histoire et origines :

Hadoop, une histoire vieille de dix ans ...

Une des raisons pour lesquelles Hadoop concentre tant l'attention de l'industrie tient à la légitimité qu'il a acquise tout au long de cette décennie. Pour la petite histoire, le succès de Google lui est en partie attribuable. En 2001, alors qu'il ne pesait rien sur le marché des moteurs de recherche, le futur géant du web développa ceux qui deviendront par la suite les composants phares d'Hadoop : **Map/Reduce**, **Google BigTable** et **Google BigFiles** (futur - Google File System) [61].

« Pour stocker, traiter et indexer 5 milliards de pages web, il construisit **Map/Reduce**, **Google Big Table** et **Google File System**. Il exploite ces éléments pendant trois ans, et en fait la description dans une publication académique » [62].

En 2004, un certain Doug Cutting, fondateur du moteur de recherche open source « **Lucene** », qui, à l'époque, travaillait sur une méthode relativement économique pour indexer en masse des pages web, s'empare des publications de Google, et crée le premier prototype d'Hadoop.

Devant le succès de Google, et sa capacité à « avaler » si facilement le web, **Yahoo**, de son côté, cherche à investir cette technologie. Elle embauche Doug Cutting en 2006, promeut activement Hadoop, le stabilise, et va jusqu'à créer, en juin 2011, une filiale dédiée à Hadoop : **Hortonworks**.

En 2009, le père fondateur d'Hadoop rejoint **Cloudera**, une Start-up créée la même année. Concurrent principal d'**Hortonworks**, elle se place comme l'un des plus gros contributeurs au projet Hadoop. Entre-temps une autre Start-up a vu le jour sous le nom **MapR**. Cette dernière enrichit Hadoop d'une gestion de stockage propriétaire [61].

Ces trois Start-up ont levé plus de 150 millions de dollars en deux ans. Aujourd'hui, plusieurs géants du web tels que **Facebook**, **Twitter**, **Linkedin** ou **eBay** exploitent Hadoop.



Depuis le 15 Octobre 2013, **Apache Hadoop** est proposé dans sa version finale stable v2.2.0 [63].

III.2.3. Concepts fondamentaux d'Hadoop :

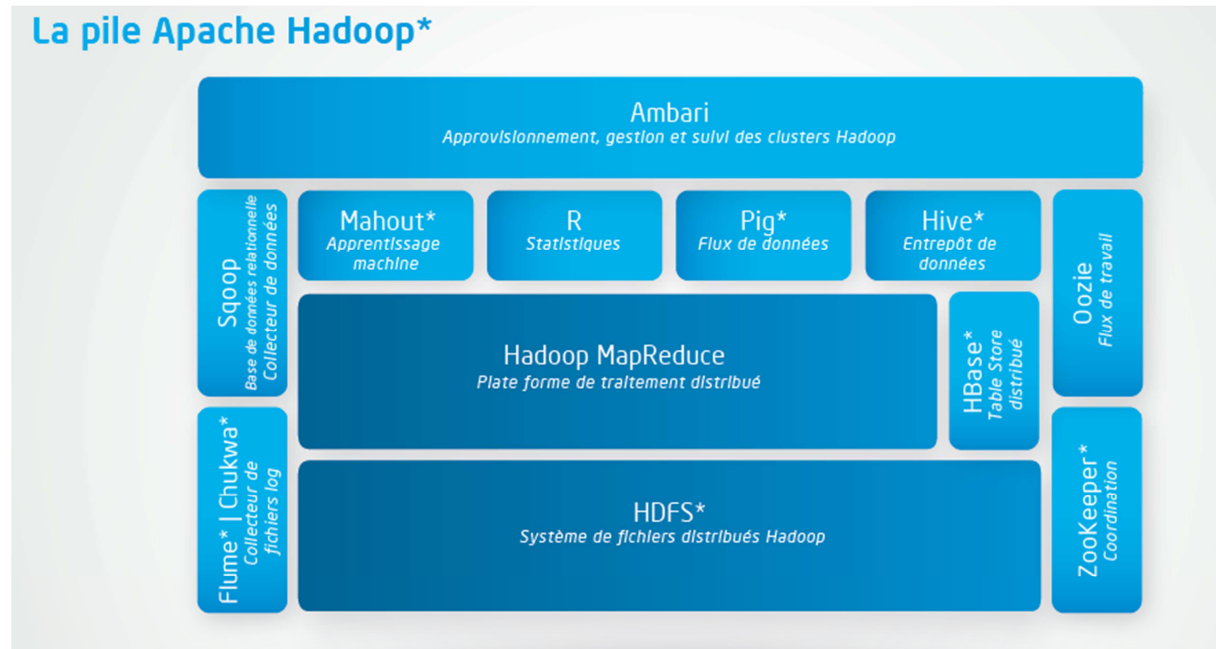


Hadoop se base sur deux concepts clés :

- **Map/Reduce.**
- **Hadoop Distributed File System** (qui s'inspire du **GFS** ou **Google File System**).

Le **HDFS** permet de distribuer les données et de les traiter grâce au **MapReduce** qui distribue une opération sur plusieurs nœuds afin de la paralléliser.

Ces deux technologies phares forment l'**écosystème Hadoop**, un écosystème fortement convoité et qui se trouve au centre de l'univers Big Data.



La Pile Logicielle : Apache Hadoop

III.2.3.A. Hadoop Distributed File System (HDFS):

HDFS est un système de fichiers distribué, extensible et portable, écrit en Java. Inspiré du **Google file System (GFS)**, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs banalisés. Il permet l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique. Le HDFS est Structuré en blocs où chaque fichier est divisé en blocs de taille prédéterminée. Ces blocs sont stockés dans un **Cluster (grappe)** d'une ou de plusieurs machines [64].

L'architecture **HDFS** est basée sur un modèle **Maître/Esclave**, dans lequel un cluster se compose d'un seul **NameNode** (nœud Maître) et un certain nombre de **DataNodes** (nœuds esclaves). Bien que l'on puisse exécuter plusieurs datanodes sur une seule machine, dans un environnement de production, ces datanodes sont le plus souvent répartis sur différentes machines [65].

Des sous-rôles relatifs au système de fichiers et à l'exécution de tâches distribuées sont associés à chaque machine.

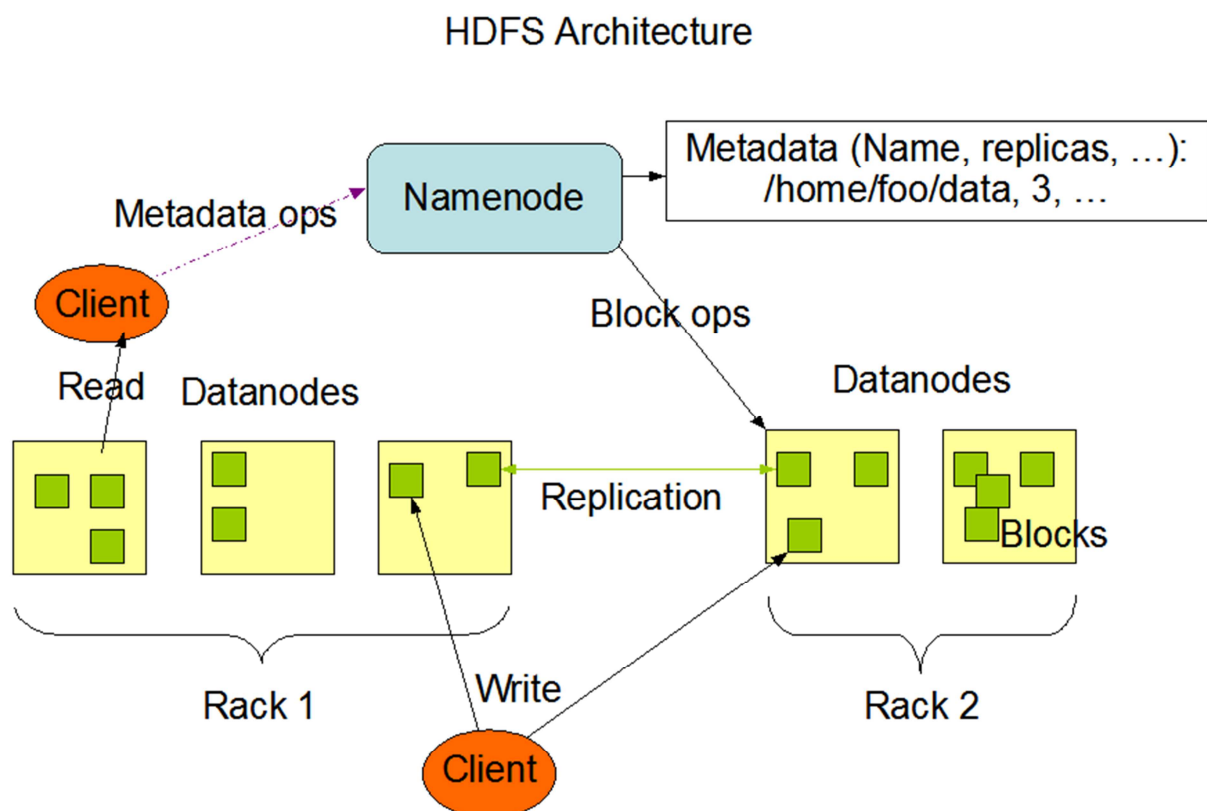
Dans le cadre des machines maîtres, trois principaux sous-rôles sont associés :

- **JobTracker** : Il s'agit de la responsabilité pour une machine maître de lancer des tâches distribuées, en coordonnant les esclaves. Le **JobTracker**, planifie les exécutions, gère l'état des machines esclaves et agrège les résultats de calculs dans le cadre de Map/Reduce [66].
- **NameNode** (Nœud de noms) :

Le rôle de NameNode intervient dans le fonctionnement du système de fichiers distribué HDFS. Une machine maître associée à ce rôle a pour responsabilité de répartir les données sur les machines esclaves et de gérer l'espace de noms du cluster. Elle contient les méta-données lui permettant de savoir sur quelle machine chaque fichier est hébergé [66].

Le NameNode agit comme le maître du système et peut être considéré comme son organisateur. Il entretient et gère les blocs présents sur les DataNodes, maintient l'ordre dans les méta-données, distribue les tâches et stocke l'emplacement de tous les fichiers présents dans le système local de fichiers.

L'architecture HDFS est construite de telle manière que les données d'un utilisateur ne sont jamais stockées dans le **NameNode**. A chaque fois qu'un client doit lire ou écrire une donnée, il communique avec le NameNode lequel répond à la demande en retournant la cartographie des blocks de données ainsi que les informations du **DataNode** [65].



Le **NameNode** est unique dans un cluster mais dispose tout de même d'une instance secondaire appelée **NameNode secondaire** qui gère l'historique des modifications dans le système de fichiers (rôle de sauvegarde).

- **Secondary NameNode :**

Ce rôle intervient dans le cadre de la redondance du NameNode. Généralement assumé par une autre machine physique que le NameNode, il permet, en cas de panne de celui-ci, la continuité de fonctionnement du cluster Hadoop [66].

NameNode est un conteneur d'informations vitales liées aux métadonnées de tous les blocs stockés dans HDFS. Ces données sont stockées non seulement dans la mémoire principale, mais également sur disque. Les deux fichiers associés sont :

- **Fsimage :** Une image du système de fichiers au démarrage du NameNode.
- **EditLogs :** Une journalisation des modifications effectuées sur le système de fichiers après le démarrage du NameNode.

Le **Secondary NameNode** lit en permanence tous les fichiers et métadonnées à partir de la RAM du NameNode et les écrit sur disque. Il est chargé d'appliquer les modifications consignées dans **Editlogs** aux blocs de données référencés dans **Fsimage**. Il télécharge les **EditLogs** du NameNode à intervalles régulières et applique les modifications à **Fsimage**. Le nouveau Fsimage est copié vers le NameNode et est utilisé chaque fois que le NameNode est lancé.

Cependant le **Secondary NameNode** ne peut être utilisé comme un substitut ou alternative au NameNode car il ne peut traiter les métadonnées. Donc, Si le NameNode tombe, c'est tout l'ensemble Hadoop HDFS qui tombe. Mais il reste un élément important dans le processus de récupération du système en cas de plantage du NameNode principal.

Fonctions principales d'un NameNode :

- Le NameNode maintient le système de fichiers et garde une trace de toutes modifications de ses propriétés.
- Il contrôle l'exécution des opérations d'E/S de bas niveau des Datanodes.
- Il gère la configuration du cluster et maintient une cartographie de l'emplacement et du découpage des blocs de données stockés dans le system de fichiers.
- Il trace l'emplacement d'un bloc de données dans un DataNode et affecte la cartographie d'un ensemble de blocs dans un fichier.

- Il enregistre les métadonnées de tous les fichiers enregistrés dans le cluster, comme par exemple l'emplacement, la taille des fichiers, les permissions, l'hierarchie, etc.
- A l'aide d'un journal de transactions (EditLogs), le NameNode enregistre le moindre changement des métadonnées du système de fichiers. Si par exemple, un fichier est supprimé de HDFS, le NameNode consigne ceci immédiatement dans ce journal.
- Le NameNode est également responsable de gérer le « facteur de réplication » de tous les blocs. Tout changement du facteur de réplication de l'un des blocs sera consigné par le NameNode dans le journal des transactions.
- NameNode reçoit régulièrement un Blockreport (rapport de bloc) de tous les DataNodes dans le cluster afin de s'assurer que les datanodes fonctionnent correctement.
- En cas de défaillance du datanode, le NameNode choisit de nouveaux datanodes pour de nouvelles répliquations de blocs de données, équilibre la charge d'utilisation des disques et gère également le trafic de communication des datanodes.

Dans le cadre des machines esclaves, deux sous-rôles sont associés :

- **TaskTracker** : Il s'agit du rôle permettant à un esclave d'exécuter une tâche Map/Reduce sur les données qu'il héberge. Le TaskTracker est piloté par le JobTracker d'une machine maître qui lui envoie la tâche à exécuter.
- **DataNode** (Nœud de données) :

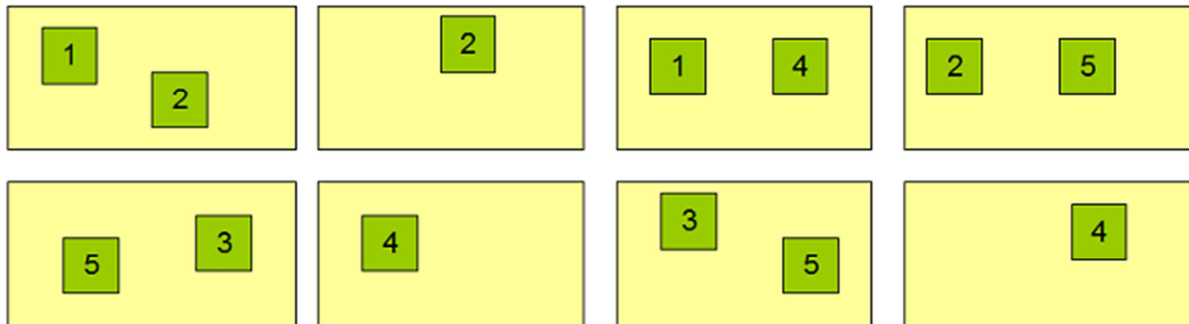
Ce nœud stocke et restitue les blocs de données. Les Datanodes sont les nœuds Clients (ou slaves, aussi appelés workers) dans HDFS. Un Datanode peut être considéré comme un serveur de blocs de données, il exécute les opérations demandées par le NameNode et représente la partie 'stockage' de HDFS. Contrairement au NameNode, un DataNode est installé sur un matériel de base, peu cher et qui ne n'offre aucune haute disponibilité.

Lors du processus de lecture d'un fichier, le NameNode est interrogé pour localiser l'ensemble des blocs de données. Pour chacun d'eux, le NameNode renvoie l'adresse du DataNode le plus accessible, c'est-à-dire le DataNode qui dispose de la plus grande bande passante. Les DataNodes communiquent de manière périodique au NameNode la liste des blocs de données qu'ils hébergent. Si certains de ces blocs ne sont pas assez répliqués dans le cluster, l'écriture de ces blocs s'effectue en cascade par copie sur d'autres [64].

Block Replication

Namenode (Filename, numReplicas, block-ids, ...)
/users/sameerp/data/part-0, r:2, {1,3}, ...
/users/sameerp/data/part-1, r:3, {2,4,5}, ...

Datanodes

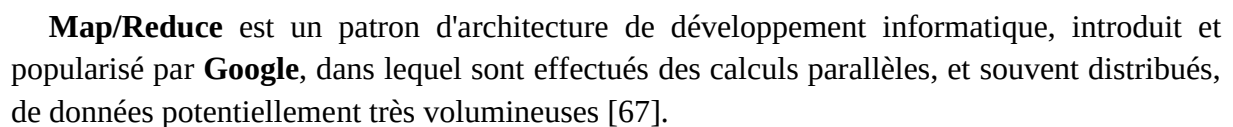


Chaque **DataNode** sert de bloc de données sur le réseau en utilisant un protocole spécifique au HDFS. Le système de fichiers utilise la couche **TCP/IP** pour la communication. Les clients utilisent le **Remote Procedure Call** pour communiquer entre eux. Le **HDFS** stocke les fichiers de grande taille sur plusieurs machines. Il réalise la fiabilité en répliquant les données sur plusieurs hôtes. Par défaut, les données sont stockées sur trois nœuds : deux sur le même support et l'autre sur un support différent. Les DataNodes peuvent communiquer entre eux afin de rééquilibrer les données et de garder un niveau élevé de réplication des données.

Fonctions principales d'un DataNode :

- Un Datanode effectue les opérations de lecture et d'écriture de bas-niveau émanant du NameNode.
- Il est responsable de la création de blocs, en les supprimant et/ou les répliquant en fonction des décisions prises par le NameNode.
- Envoyer régulièrement un rapport sur tous les blocs présents dans le cluster au NameNode.
- Créer un canal de communication et transmet des blocs de données à d'autres DataNodes.
- Envoyer périodiquement des pulsations au NameNode (en moyenne une fois toutes les 3 secondes), pour signaler l'état de santé global du HDFS.

- ### Schéma d'architecture Hadoop représentant les principaux rôles des machines



Les Big Data | Ahmed Senani

Map/Reduce permet de manipuler de grandes quantités de données en les distribuant dans un cluster de machines pour être traitées. Ce modèle connaît un vif succès auprès de sociétés possédant d'importants centres de traitement de données. Il commence aussi à être utilisé au sein du **Cloud computing** [67].

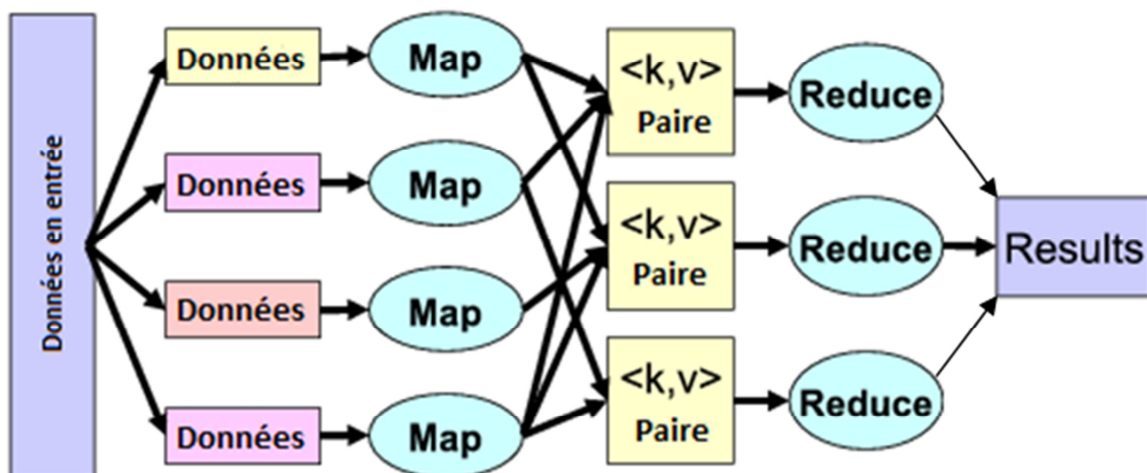
De nombreux **frameworks** ont vu le jour afin d'implémenter le **Map/Reduce**, mais le plus connu est **Apache Hadoop**.

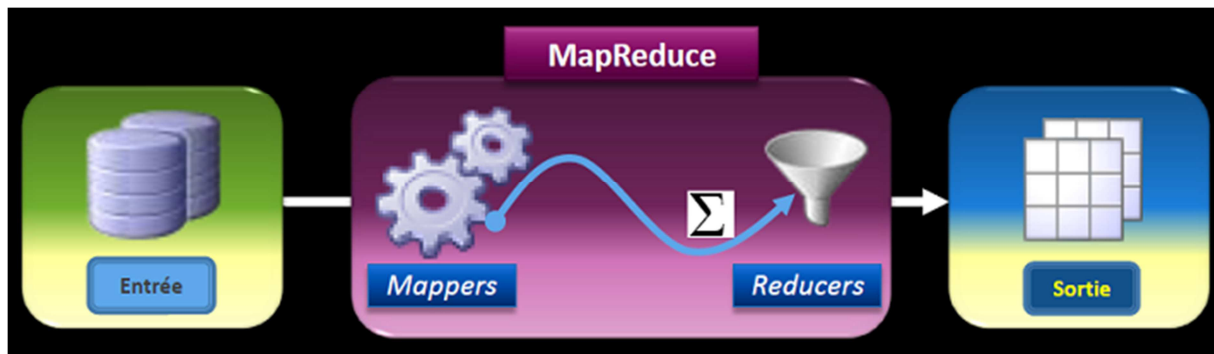
Le **Map/Reduce** possède plusieurs avantages, il permet de [68] :

- Traiter de grands volumes de données.
- Gérer des milliers de serveurs.
- Paralléliser et distribuer des traitements.
- Ordonnancer les entrées/sorties.
- Gérer la tolérance aux pannes.
- Gérer entièrement le cluster et la répartition de la charge, ce qui permet de faire des calculs distribués dans un environnement Cloud).

Ce qui rend ce **framework** si populaire c'est le fait qu'il possède plusieurs implémentations dans différents langages (**C++**, **Java**, **Python**, etc.).

Fonctionnement de l'algorithme Map/Reduce :





Le modèle **Map/Réduire** s'articule autour de **deux étapes** (fonctions) principales :

❖ La fonction "**Map**" :

Dans l'étape Map, le nœud à qui est soumis un problème, le découpe en sous-problèmes, et les délègue à d'autres nœuds (qui peuvent en faire de même récursivement). Les sous-problèmes sont ensuite traités par les différents nœuds à l'aide de la fonction Map. Le mapper récupère en entrée les valeurs de chaque ligne et les mappe au format «**key,value**» («**clé, valeur**»). Cette façon de procéder permet d'imposer une structure unique et simple aux enregistrements en entrée ce qui permet l'efficacité de Map/Reduce au niveau des entrées-sorties.

❖ La fonction "**Reduce**" :

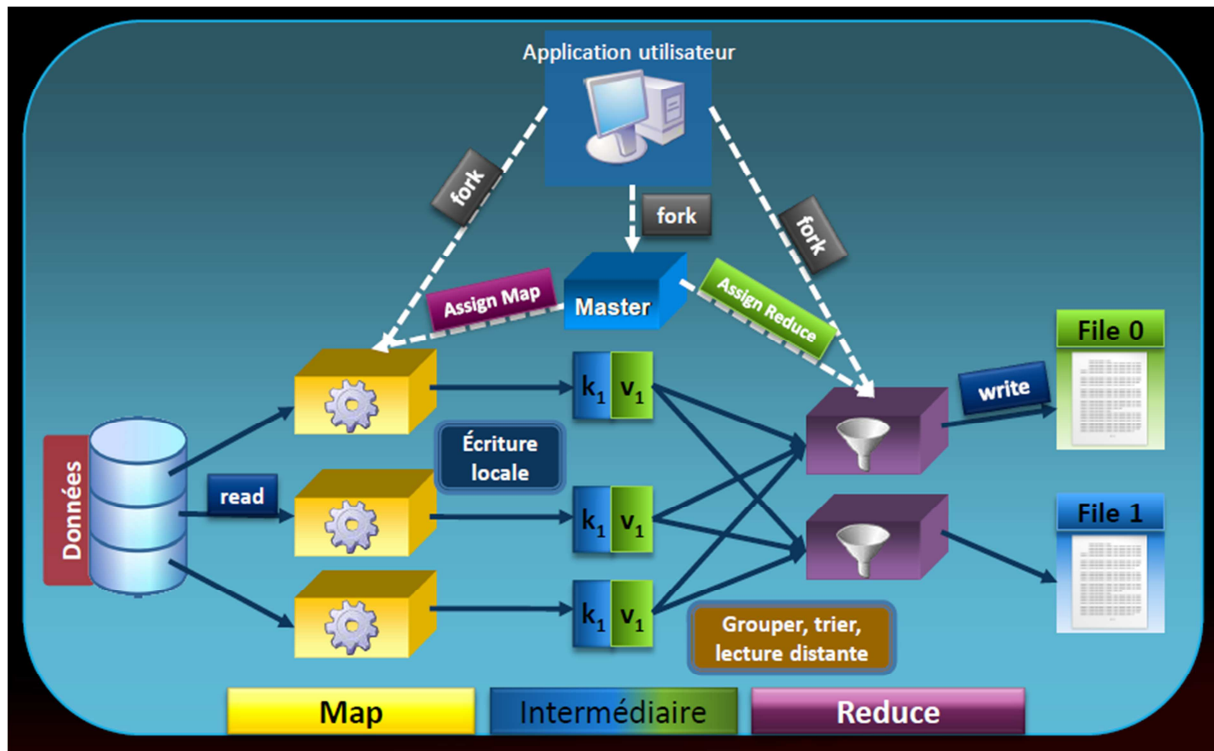
Dans cette étape, les nœuds les plus bas font remonter leurs résultats au nœud parent qui les avait sollicités. Celui-ci calcule un résultat partiel à l'aide de la fonction Reduce (réduction) qui associe toutes les valeurs correspondant à la même clé à une unique paire (clé, valeur). Puis il remonte l'information à son tour.

En résumé [70] :

- **Map**
 - Prend en entrée un ensemble de « **Clé, Valeurs** »
 - Retourne une liste intermédiaire de « **Clé1, Valeur1** »
 - **Map(key,value) → list(key1,value1)**
- **Reduce**
 - Prend en entrée une liste intermédiaire de « **Clé1, Valeur1** »
 - Fournit en sortie un ensemble de « **Clé1,Valeur2** »
 - **reduce(key1, list(value1)) → value2**

Il existe une phase intermédiaire appelée « **shuffle et sort** » qui se place entre la phase « map » et la phase « reduce ». Cette phase relie les clés similaires entre elles avec leurs valeurs associées, réarrange et fait le tri des éléments de la liste afin de préparer le reducing.

Mise en œuvre de Map/Reduce :



Répartition des rôles dans Map/Reduce :

❖ Le 'Master' [71] :

- Coordonne l'exécution des "unités de travail" (Workers) :
 - Attribue aux unités de travail les tâches "map" et "reduce".
- Gère la distribution des données :
 - Déplace les "**workers**" vers les données.
- Gère la synchronisation :
 - Regroupe, trie, et réorganise les données intermédiaires.
 - Détecte les défaillances des unités de travail et relance la tâche.

❖ Les ‘Workers’ [72] :

- Un ‘**Worker**’ est une “unité de travail” qui possède trois états :
 - **Idle**, indique qu’un worker est disponible pour une nouvelle planification.
 - **Completed**, indique la fin d’un traitement, le ‘worker’ informe le ‘master’ de la taille et de la localisation des fichiers intermédiaires.
 - **In-progress**, indique qu’un traitement est en cours.
- Les ‘**reducers**’ sont informés des états des ‘**workers**’ par le ‘**master**’.

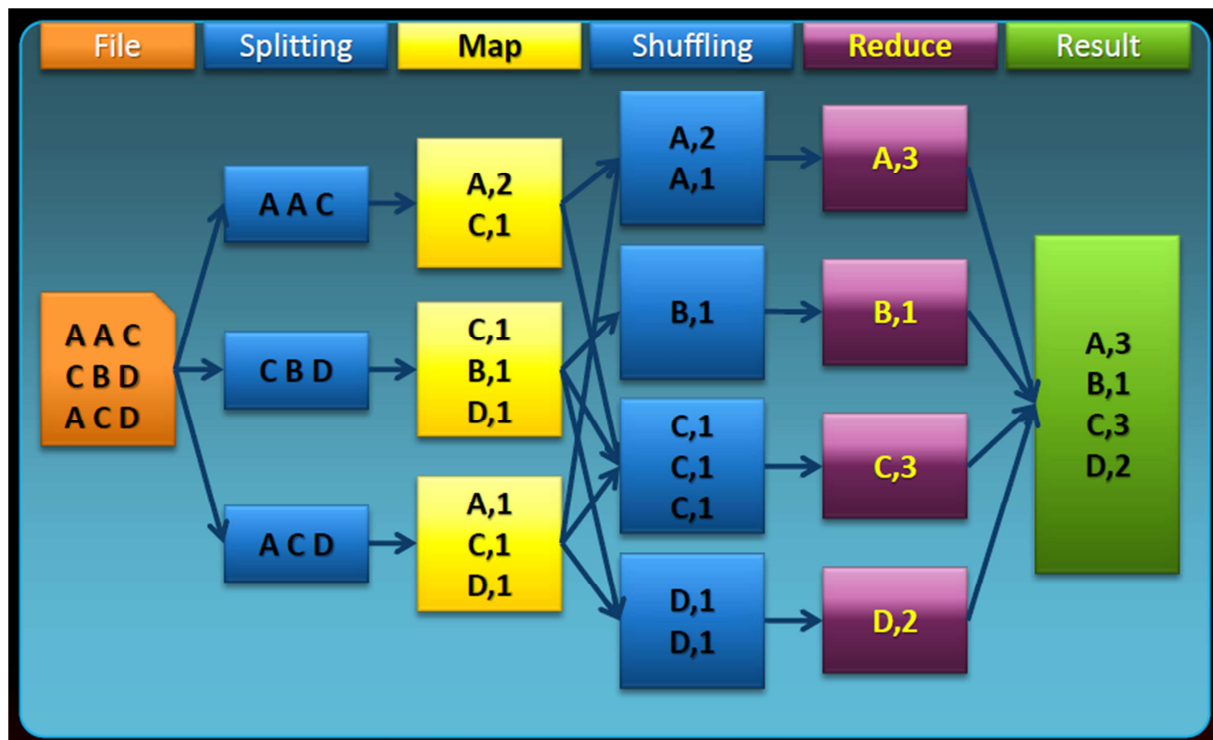
Gestion des données dans Map/Reduce [73] :

- Les données en Entrée et en Sortie sont stockées sur un système de fichiers distribué.
- Les données intermédiaires sont stockées sur le système de fichiers local des unités «**map**» et «**reduce**».
- Les données en sortie représentent souvent une entrée pour une autre unité **Map/Reduce**.

Gestion des défaillances [74] :

- Basée sur un mécanisme de réexécution.
- Le Master ‘**ping**’ régulièrement les unités « **map** » et « **reduce** ».
- En cas de défaillance d’une unité « **map** » :
 - Les tâches complètes ou en cours d’exécution seront réinitialisées.
 - Les tâches seront placées dans de nouveaux noeuds du système de fichiers.
- En cas de défaillance d’une unité « **reduce** » :
 - Les tâches complètes (état: **completed**) ne sont pas relancées.
 - Uniquement les tâches en cours d’exécution seront réinitialisées et relancées sur un autre nœud du système de fichiers distribué.
- En cas de défaillance du « **Master** » :
 - Les tâches **Map/Reduce** seront abandonnées et le client (utilisateur) est notifié.

Exemple de mise en œuvre :



III.2.4. Les Avantages d'Hadoop :

Hadoop offre plusieurs avantages clés pour l'analyse de Big Data [75] :

- **Stocker tout type de données dans leur format natif :**

Les données ne nécessitant pas de traduction vers un schéma spécifique, aucune information n'est perdue.

- **Adaptabilité au Big Data :**

Hadoop a déjà fait ses preuves en termes d'adaptabilité via des sociétés comme Facebook et Yahoo!, qui gèrent des implémentations gigantesques.

- **Offrir de nouvelles perspectives:**

L'analyse de Big Data révèle des interactions cachées qu'il serait difficile, long et coûteux voire impossible de résoudre en utilisant des approches traditionnelles de DataMining.

- **Réduire les coûts.**

Le logiciel open-source Hadoop fonctionne sur des serveurs standards et a un coût par téraoctet inférieur pour le stockage et le traitement. Le stockage peut être ajouté progressivement en fonction des besoins, et le matériel peut être ajouté ou échangé au sein ou hors d'un cluster.

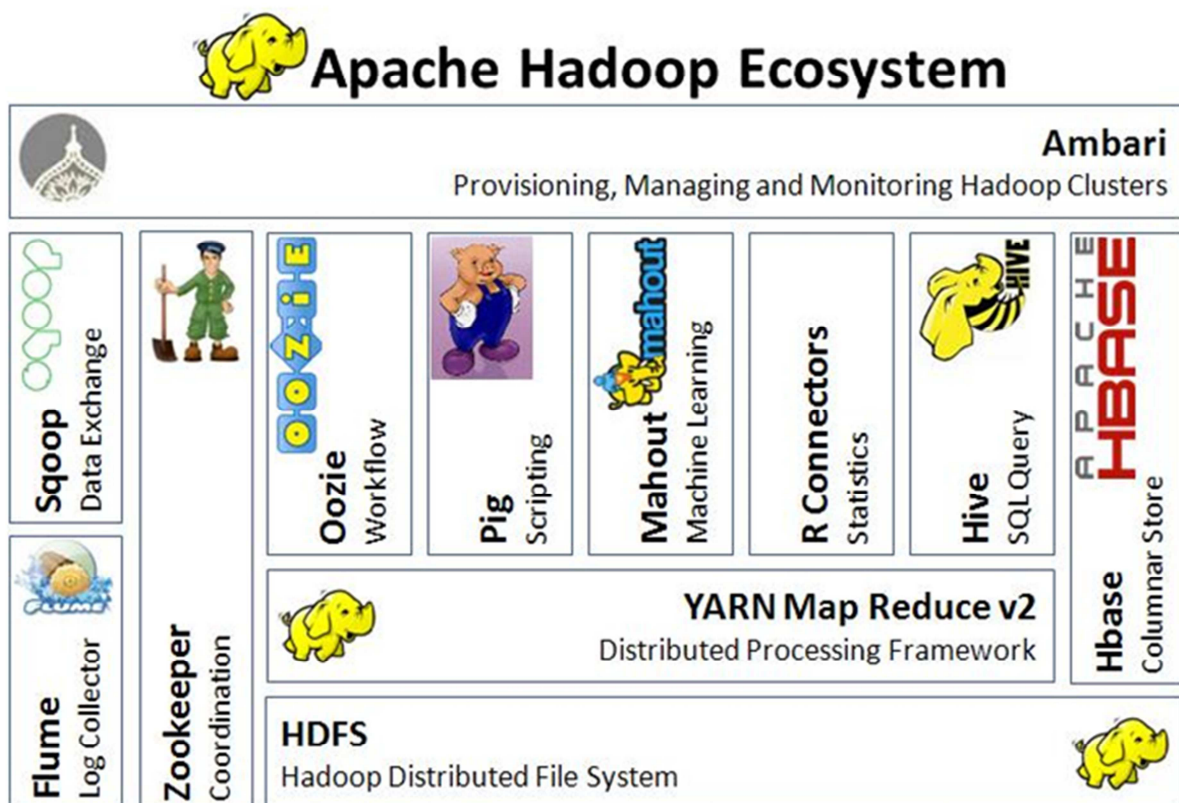
- **Une plus grande disponibilité.**

Hadoop se remet des erreurs matérielles, logicielles et du système en offrant une tolérance aux pannes grâce à la réplication des données et au basculement entre les nœuds de calcul.

- **Diminution du risque.**

La communauté Hadoop est active et diversifiée, avec des développeurs et des utilisateurs issus de nombreux secteurs à travers le monde. Hadoop est une technologie qui va continuer à se développer et est en constante évolution.

III.2.5. L'écosystème « Hadoop » :



Hadoop est un environnement d'exécution distribué, performant et scalable, dont la vocation est de traiter des volumes de données considérables. Parce que ce système est de portée générale, il a suscité le développement d'un vaste écosystème. Au-delà du **HDFS** et du **Map/Reduce** qui forment les briques fonctionnelles de base, l'écosystème Hadoop compte plusieurs projets **Apache** indépendants [76].

Parmi ces projets, on retrouve :



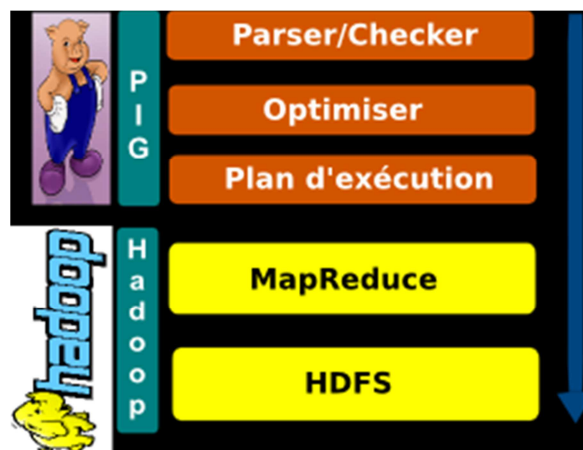
A l'origine d'un projet **Yahoo !**, Pig est un système de traitement des gros volumes de données en utilisant la plateforme **Hadoop Map/Reduce**.

Traiter des gros volumes de données et offrir des services à base de ces traitements nécessite des méthodes et des processus de traitement hautement productifs. C'est dans cette optique que le projet **Hadoop Pig** a vu le jour. Il démocratise les traitements et les rend accessible au non développeur, en proposant un langage d'abstraction simple au-dessus de Map/Reduce [77].

Pig est composé de deux modules qui sont :

- **Pig Latin** : langage procédural de haut niveau, il permet d'écrire des scripts de traitement de données qui sont traduits en jobs Map/Reduce et exécutés sur le cluster Hadoop.
- **Pig Engine** : Parse, optimise et exécute automatiquement les scripts **Pig Latin** comme une série de jobs Map/Reduce au sein d'un cluster Hadoop.

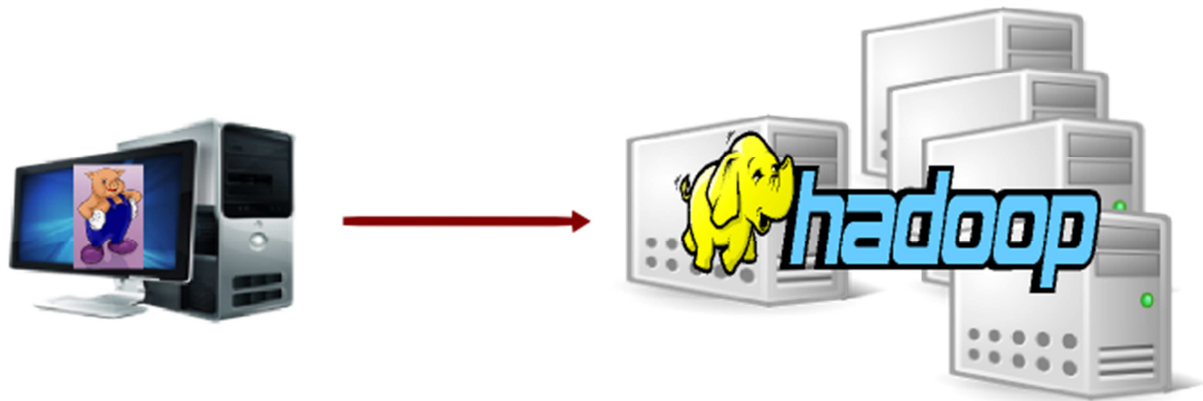
Il fournit les opérations standards pour la manipulation de données (filters, joins, ordering), des types primitifs, des types complexe (tuples, bags, maps). Simple à comprendre, il ouvre la voie vers Hadoop au non programmeur Map/Reduce.



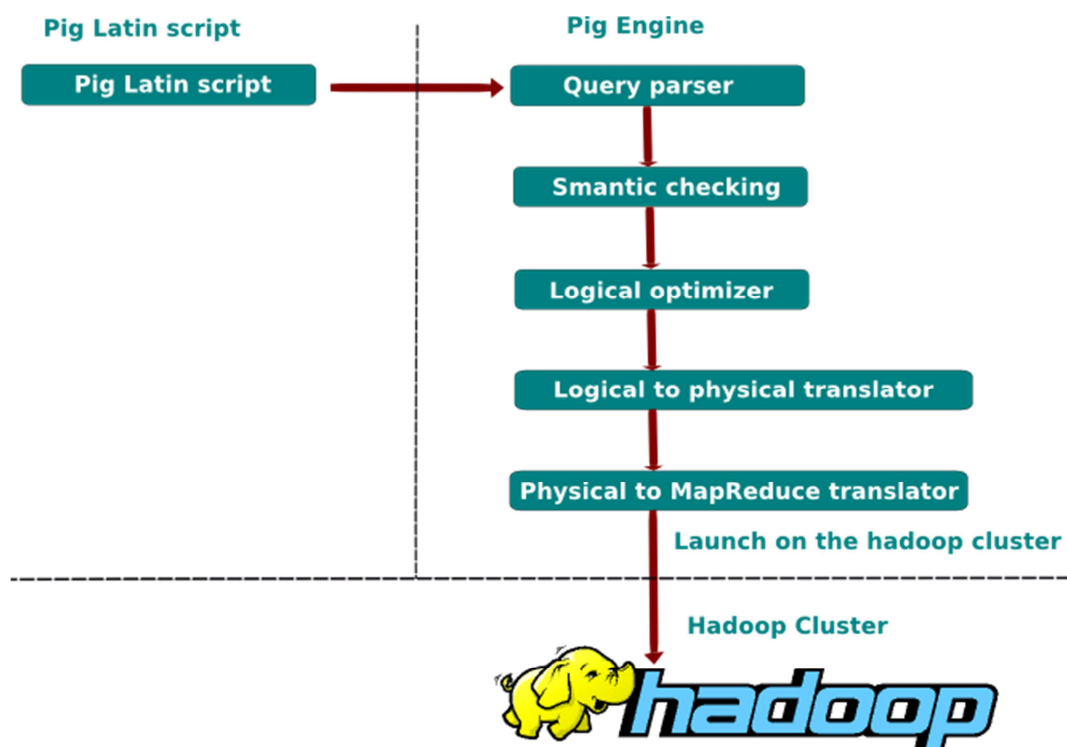
Architecture de Pig [77] :

Pig s'installe sur la machine locale de l'utilisateur et ne nécessite aucune installation sur le cluster Hadoop.

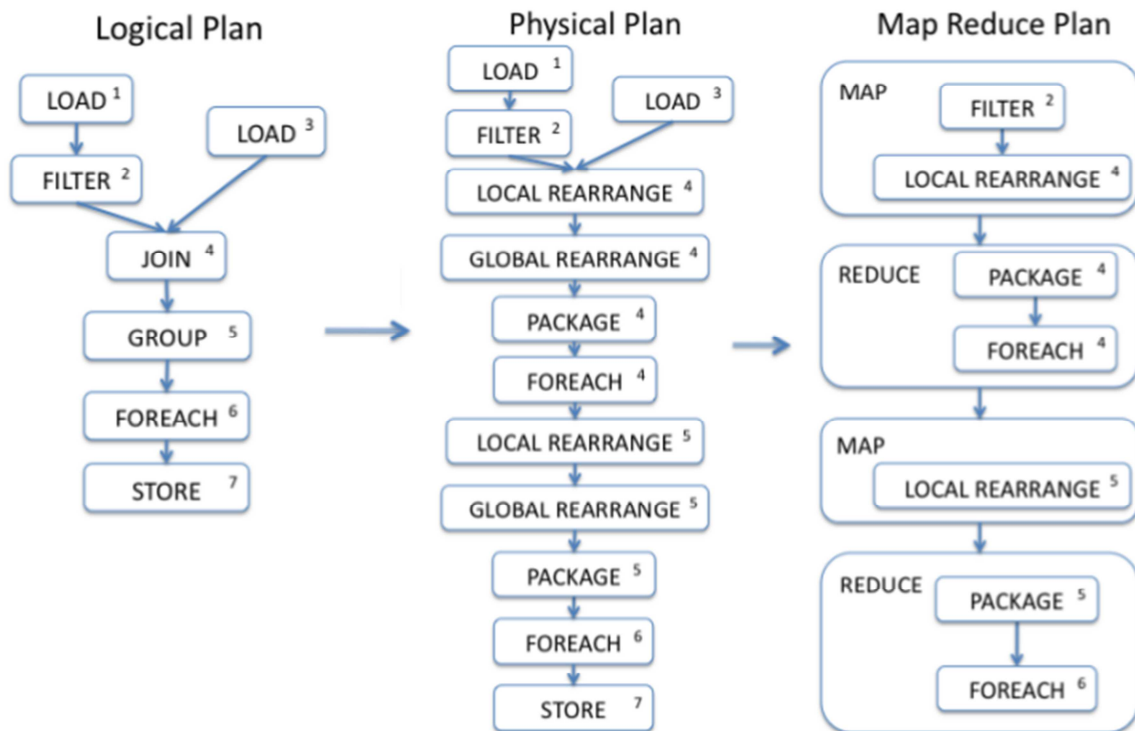
Interaction de Pig avec le cluster Hadoop



Processus de création d'un job Map/Reduce avec Pig



Processus de passage du plan logique au Map/Reduce



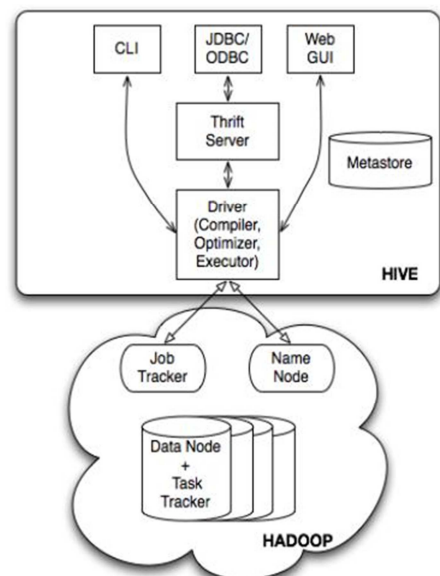
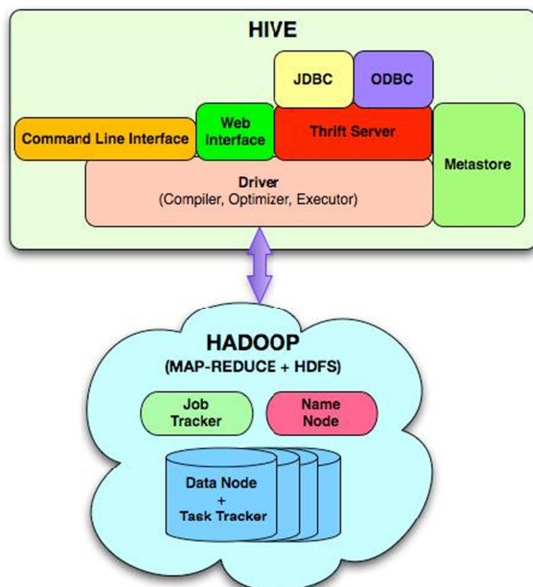
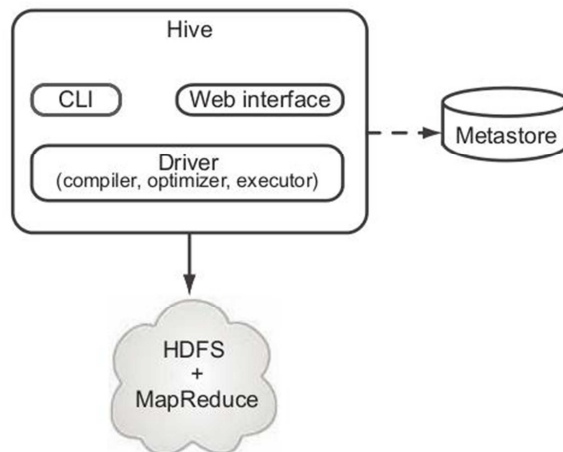
Hive :

Apache Hive est une infrastructure d'entrepôt de données (Data Warehouse) open source pour Hadoop. Initialement développé par **Facebook**, Hive fournit un langage de requête de haut niveau semblable à SQL, appelé **HQL** (HiveQL ou Hive Query Language), pour interagir avec un cluster hadoop, dans le but de faciliter les agrégations, le requêtage ad-hoc et l'analyse de gros volumes de données.

L'augmentation des sources de données et le volume généré rend les traitements BI classiques coûteux en temps et en argent. L'émergence du stockage distribué et des traitements parallèles avec le **framework Hadoop** en se basant sur des serveurs standards, valorise et donne une seconde vie aux données. Toutefois, le modèle de programmation **Map/Reduce** exige des développements bas niveau, personnalisés, qui sont difficiles à maintenir et à réutiliser [78].

C'est afin de lever ce type de complexité, que le projet **Hive** a été mis en place. Il permet de définir une structure pour les données non structurées, ce qui simplifie le processus d'analyses et de recherches.

Architecture de Hive :



1. Définition :

HBase est un système de gestion de base de données non-relationnelles distribué, écrit en Java, il est inspiré des publications de Google sur BigTable et dispose d'un stockage structuré pour les grandes tables [79].

HBase est généralement couplé au système de fichiers HDFS, pour faciliter la distribution des données sur plusieurs nœuds, et permet de gérer les accès aléatoires lecture/écriture en temps réel à un très grand ensemble de données.

Dans l'écosystème Hadoop, le stockage des données se fait sur le système de fichiers HDFS. L'accès à une donnée ponctuelle stockée dans un fichier, revient à faire un « full scan » sur tout le cluster ou du moins sur une partie. Une opération coûteuse en temps et en calcul.

Accéder à cette même donnée en temps réel ou presque, revient à structurer et indexer la donnée. **HBase** vient pour répondre à cette problématique et bien d'autres.

HBase est une base de données **NoSQL** orientée colonne, scalable et tolérante aux pannes. La charge de travail en termes de mémoire et de calcul (CPU) ainsi que le stockage, est distribuée sur toutes les machines du cluster HBase [80].

Les bases de données orientées colonnes, stockent de manière contigüe les valeurs de la même colonne sur disque. Une approche totalement différente des bases de données traditionnelles où le stockage se fait par ligne (les valeurs des différentes colonnes d'une ligne sont stockées d'une manière contigüe sur disque).

Ce type de base est plus adapté par exemple aux bases analytiques, où les traitements associées sont sur quelques colonnes de la ligne et non sur toute la ligne. Cette approche réduit drastiquement les I/O et par conséquent les temps de réponse. De plus, ce type de sauvegarde est un bon candidat à la compression des données par le fait qu'ils sont de même type.

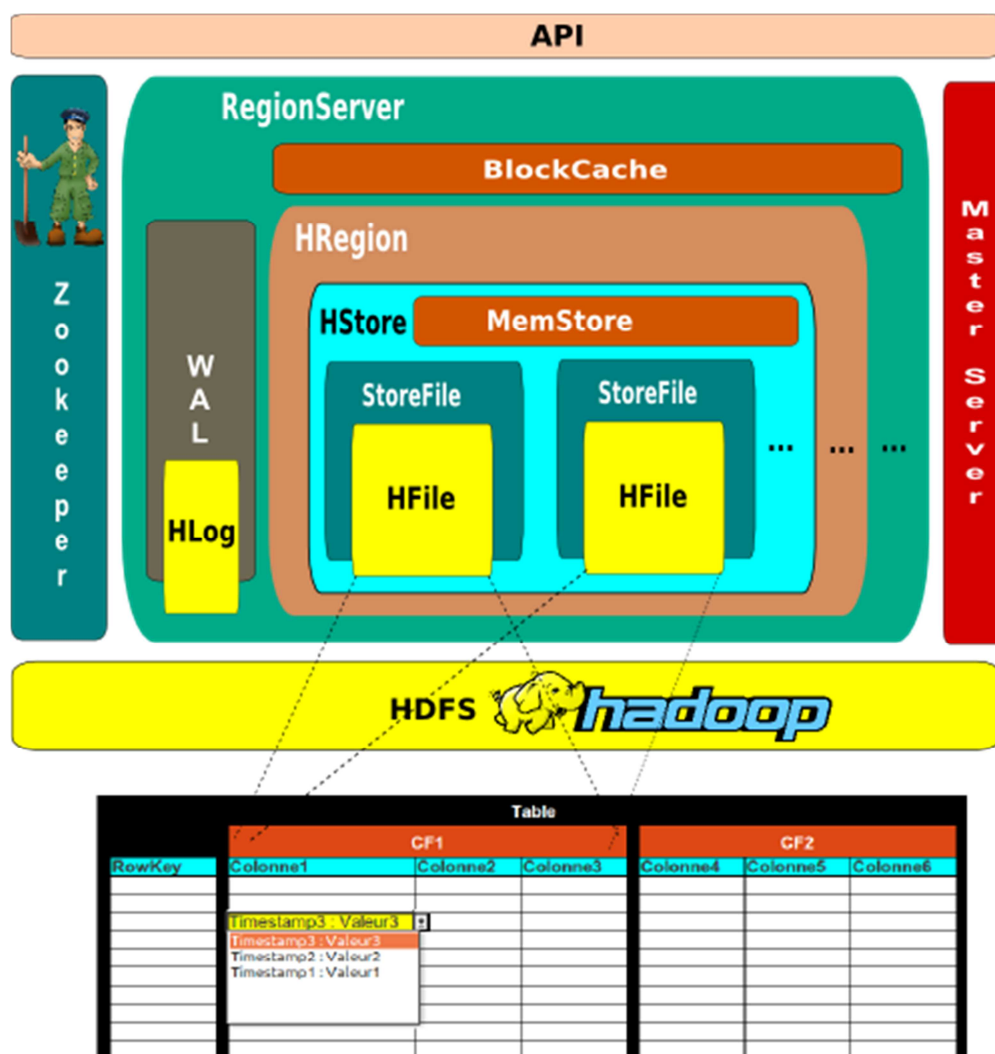
2. Concepts de HBase :

La base de données HBase est régie par les concepts suivants [80] :

- **Map** : Le stockage se fait dans une map. Cette dernière est basée sur le principe de clé/valeur. Chaque valeur (tableau de bytes) est identifiée par une clé (tableau de bytes). L'accès à une valeur par sa clé est très rapide.
- **Map triée** : La map est triée par ordre lexicographique. Cette fonctionnalité de tri est très importante car elle permet de récupérer les valeurs par intervalle de clés.
- **Multidimensionnel** : La clé dans la map est une structure composée de row-key, column family, column, et d'un timestamp.
- **Null (Sparse)** : Contrairement aux bases de données relationnelles, une colonne qui n'a pas de valeur n'est pas matérialisée (aucun stockage n'est nécessaire en cas d'une valeur nulle pour une colonne).
- **Persistence** : Les données stockées dans la map sont sauvegardées durablement sur disque.

- **Consistance (cohérence)** : Toutes les modifications sont atomiques et les lectures se font toujours sur la dernière valeur validée.
- **Système distribué** : Le système de base de données est construit sur un système de fichiers distribué afin que le stockage de fichiers sous-jacent soit réparti sur un ensemble de machines d'un cluster. Les données sont répliquées sur un certain nombre de nœuds permettant ainsi une tolérance aux pannes.

3. Les Composants de l'architecture HBase :



3.1. RegionServer :

Le RegionServer est le point d'entrée pour l'accès à la donnée. IL gère plusieurs processus et composants, et il expose plusieurs méthodes [80] :

- Data (get, put, delete, next, etc.).
- Region (splitRegion, compactRegion, etc.).

Un **RegionServer** contient un seul WAL, un BlockCache et plusieurs Régions.

3.1.1. BlockCache :

Le BlockCache est un cache LRU. Il est activé par défaut pour toutes les tables, ce qui signifie que toute opération de lecture sont chargées dans le cache LRU (Least Recently Used).

3.1.2. Write Ahead Log (WAL) :

Chaque ajout ou mise à jour dans un RegionServer est systématiquement écrit dans write-ahead log (WAL) en premier lieu. Avant d'être répliquer dans le MemStore. Ce mécanisme garantit la durabilité de la donnée en cas de défaillance du serveur. Le processus WAL écrit dans le fichier Hlog qui est placé dans HDFS. Pour chaque instance RegionServer il y a une instance de WAL.

3.1.3. Region :

C'est l'élément de base dans le stockage et la distribution de la donnée dans HBase. Chaque Region gère un sous ensemble d'une table HBase (une partition).

Une Region est composé de :

- **Store** : Chaque partition d'une ColumnFamily est gérée par un store. HStore est l'implémentation d'un Store, il est composé de plusieurs StoreFiles et d'un MemStore.
- **MemStore** : C'est un cache en mémoire. Il stocke toutes les écritures et les mises à jour relatives à une partition.
- **HFile** : fichier physique sur lequel les données sont sauvegardées.

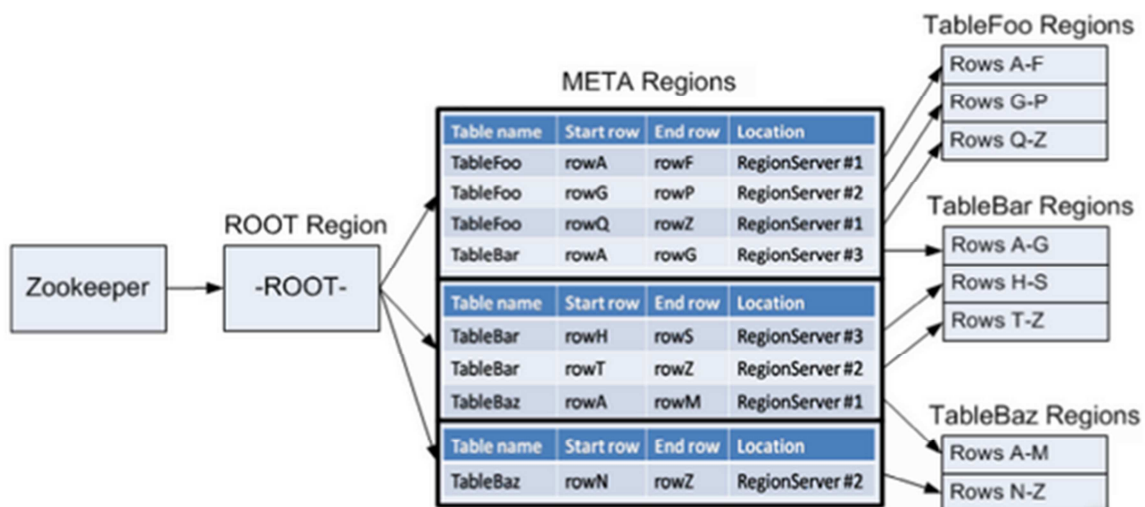
3.2. MasterServer :

HMaster est l'implémentation du MasterServer. Le MasterServer est chargé de coordonner et de surveiller toutes les instances de RegionServer du cluster. Il est en charge de la répétition des Region(s) sur les nœuds du cluster. Les changements dans les tables métadonnées passent par le MasterServer.

3.3. Zookeeper :

Il surveille l'état du cluster et informe régulièrement le MasterServer des différents états. De plus, il stocke les informations critiques du cluster, dont l'emplacement de la table système **-ROOT-**.

- **-ROOT-** : contient la liste des tables **.META.** et leurs emplacements.
- **.META.** : contient la liste des **Region** et leurs emplacement.



Gestion des métadonnées avec ZooKeeper.



Sqoop :

Sqoop est un projet développé par Cloudera et il fait aujourd'hui partie intégrante des projets Open-Source de l'**Apache software foundation**. Il a pour objectif de permettre une meilleure cohabitation des systèmes traditionnels de type SGBDs avec la plateforme Hadoop.

Il est ainsi possible de procéder à la collecte de données au sein d'applications traditionnelles n'ayant pas la capacité de se connecter au cluster. Inversement, il est possible d'exporter le résultat d'un traitement vers une base de données tierce afin qu'il soit exploité par une application [81].

Sqoop fournit un moyen rapide de transférer de grandes quantités de données entre la plate-forme de traitement de données Hadoop et les SGBD relationnelles, les systèmes d'entreposage et d'autres banques de données non-relationnelles. Il fonctionne avec la plupart des bases de données relationnelles actuelles, comme **MySQL**, **PostgreSQL**, **Oracle**, **Microsoft SQL Server** et **IBM DB2**, ainsi que les applications d'entreposage de données.

Sqoop place les données, soit directement dans un espace de stockage régi par le système Hadoop Distributed File (HDFS), ou peut les orienter vers d'autres applications Hadoop comme le système de gestion de base de données non relationnelle distribuée **HBase** et son stockage structuré pour les grandes tables, ou le logiciel d'analyse de données **Hive** de Hadoop [82].



Flume :

Flume est une solution de collecte, agrégation et transfert de gros volumes d'évènements (logs). Il a été pensé pour gérer des débits importants avec une fonctionnalité native d'écriture dans HDFS. Pour gérer ces gros volumes/débits, il se doit d'être très scalable, et donc distribué [83].

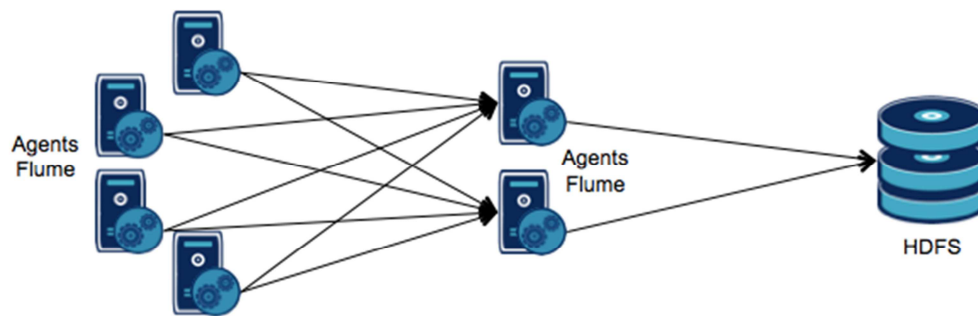
Flume a été développé au départ chez Cloudera (éditeur Hadoop) en open source sous license Apache. La première release stable de ce qu'on appelle aujourd'hui la version **Flume OG** (Old Generation, version 0.9.x) date de début 2010.

Flume OG a subi un changement profond pour plusieurs raisons, les principales sont les suivantes :

- **Architecture complexe** avec un master et un zookeeper pour centraliser la configuration.
- **Code endetté** avec beaucoup de fonctionnalités peu utilisées.
- Des limitations techniques notamment concernant la persistance des messages.

Flume NG (Next Generation) remplace donc **Flume OG** pour apporter une réponse à ces limitations, avec notamment une simplification de l'architecture, maintenant composée d'agents « pairs » distribués : plus besoin de **Master** et de cluster **Zookeeper** [83].

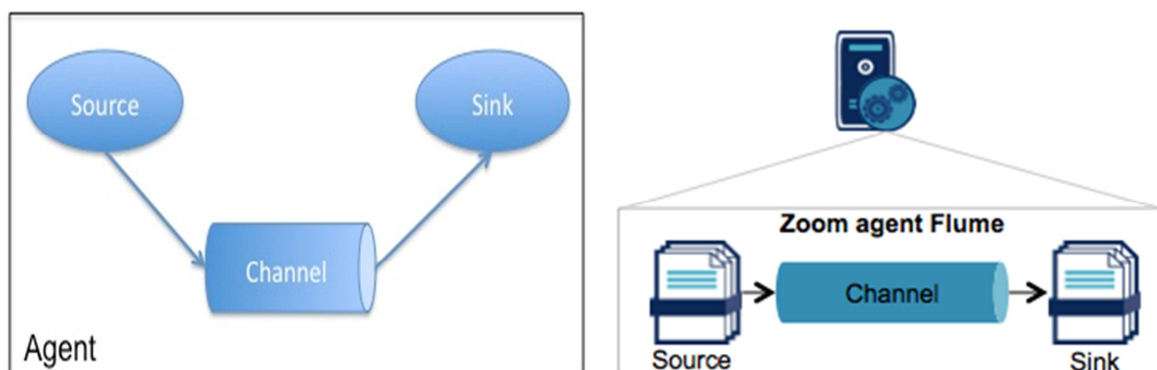
Architecture de Flume NG :



La figure ci-dessus représente une architecture typique d'agents Flume :

- Des agents collés aux sources d'événements.
- Des agents qui vont se charger de consolider les logs afin de les écrire dans un référentiel centralisé, typiquement un cluster HDFS ou une base HBase.

Flume est composé d'agents possédant les caractéristiques suivantes :



Chaque agent exécute des « **routes** ». Une route est constituée de :

Sources :

Une source consomme des événements provenant de sources externes (syslog, autre agent Flume, etc) et place ces derniers dans un ou plusieurs channels.

Il existe différents types de sources, les principales sont :

- **Avro** : Mettre l'agent Flume en écoute sur un port TCP et recevoir des logs au format Avro. Utilisé également pour communiquer entre différents agents Flume.
- **Spooling Directory Source** : Récupérer le contenu des fichiers de log qui arrivent dans un répertoire.

- **Syslog** (TCP ou UDP) : Capter les évènements d'un server syslog. Il est, par exemple, possible de configurer Tomcat pour envoyer les logs vers syslog et Flume les prendra en charge.
- **HTTP** : Un handler se charge de traduire les requêtes POST et GET (encore en phase d'expérimentation).

Channels :

Un channel (base de données, en mémoire, etc.) permet de stocker temporairement les évènements attendant d'être consommés

Il existe différents types de channels :

- **Memory** : garde les évènements en mémoire pour favoriser la performance.
- **JDBC** (Java Database Connectivity) : les évènements sont stockés sur une base de données.
- **File** : Persister les logs sur le filesystem pour garantir la non-perte de message en cas de panne et/ou redémarrage de l'agent.

Sinks :

Un sink va pousser les évènements, stockés temporairement dans la channel, vers la destination souhaitée (HDFS, Flume agent source, etc.).

Il existe différents types de Sinks :

- **HDFS** : pousse les évènements dans Hadoop.
- **AvroSink** : rediriger les logs au format Avro sur un port TCP distant. Le couple **AvroSource/AvroSink** est utilisé pour transférer les logs d'un agent vers un autre.
- **IRC** : pousse les évènements sur un channel IRC.
- **HBaseSink** et **ElasticSearchSink** : écrire dans une base NoSQL. Il existe aussi des sinks open source pour Cassandra, MongoDB, etc.
- **FileRollSink** : écrire dans le filesystem local.
- **NullSink** et **LoggerSink** : Utilisés pour tester la configuration et valider que les logs arrivent bien à destination.

Il existe une alternative à Flume, développée également par Apache, qui s'appelle **Chukwa**.

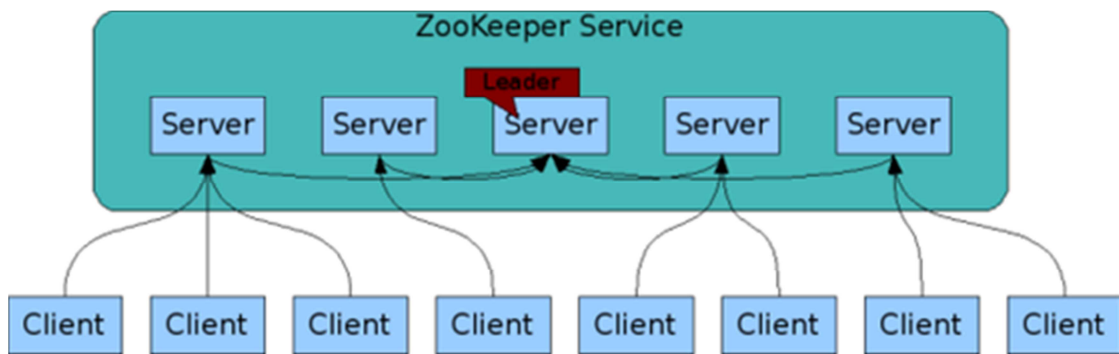


➤ Zookeeper :

ZooKeeper est un projet open-source de l'**Apache Software Foundation**, fournissant un service centralisé pour maintenir des informations de configuration, de nommage, assurer la synchronisation distribuée, et la fourniture de services de groupe. Tous ces types de services sont utilisés sous une forme ou une autre par des applications distribuées [84].

Le rôle de Zookeeper, en particulier, est de fournir aux composants Hadoop les fonctionnalités de distribution. Pour cela il centralise les éléments de configuration du cluster Hadoop, propose des services de clusterisation et gère la synchronisation des différents événements.

L'interface exposée par Zookeeper intègre l'aspect **wait-free** de messagerie de groupe, partage de registres avec un mécanisme de concurrence similaire au service de verrouillage, afin d'offrir une simple et puissante coordination [76].



Zookeeper a comme principale fonction de simplifier l'accès à la donnée en offrant un répertoire de classification des données existantes sur le cluster. Dans un aspect totalement virtuel, Zookeeper ordonne (virtuellement) les données et les classifie dans des registres tout en enregistrant leurs états mais aussi l'état d'avancement des clients sur chaque donnée. De ce fait, même si un client, pour une raison ou une autre, se trouve dans l'obligation de changer de serveur (mais reste dans le même service) pourrait continuer son travail, sans pour cela refaire toutes les transactions et requêtes déjà effectuées sur un premier serveur.

Apache ZooKeeper est un élément indispensable au bon fonctionnement de **HBase**.



Oozie est un système d'ordonnancement de flux de travail utilisé pour gérer et coordonner les tâches de traitement de données (jobs) à destination de Hadoop. Développé par **Yahoo!**, il est aujourd'hui distribué par la **fondation logicielle Apache** [85].

Oozie est implémenté en tant qu'application web Java exécutée dans un conteneur de servlets (servlet-Container) et s'intègre parfaitement avec l'écosystème Hadoop puisqu'il supporte les types de jobs suivants :

- **MapReduce** (Java et Streaming).
- **Pig**.
- **Hive**.
- **Sqoop**.
- Et autres ... tels que programmes Java ou scripts de type Shell.



Avro est un système de **sérialisation** des données faisant partie des projets **Apache**. Il utilise le format d'échange **JSON** (JavaScript Object Notation) pour définir les types de données et protocoles, et sérialiser les données dans un format binaire compact et léger. Il est utilisé principalement dans Apache Hadoop où il fournit à la fois, un format de sérialisation pour la sauvegarde et la persistance des données, et un format de transmission pour la communication entre les nœuds hadoop ou entre les programmes clients et les services hadoop [86].

Avro offre [76] :

- Une structure de données riche.
- Un format binaire de données compact et rapide.
- Un fichier conteneur pour les données persistantes.
- Appel de procédures à distance (RPC).
- Intégration simple avec les langages dynamiques. La génération de code n'est pas nécessaire pour lire ou écrire des fichiers de données, ni à utiliser ou à mettre en œuvre des protocoles RPC. La génération de code est comme une option d'optimisation, la mise en œuvre ne vaut que pour les langages typiquement statiques.

III.2.6. Les distributions d'Hadoop :



Une distribution contient différents projets de l'écosystème Hadoop et offre des packages, des outils et un support commercial, ce qui réduit beaucoup les efforts à mettre en œuvre, pas seulement pour le développement mais aussi pour l'opérationnel [87].

Les fournisseurs de distribution offrent des outils graphiques pour le déploiement, l'administration et le monitoring des clusters Hadoop. De cette façon, il est beaucoup plus facile d'installer, de gérer et de surveiller les clusters. L'investissement est beaucoup réduit.

Fournisseurs de distributions Hadoop

En plus d'**Apache**, Hadoop est distribué par trois grands acteurs, qui proposent des services de formation et un support commercial mais également des fonctions supplémentaires, à savoir : **HortonWorks**, **Cloudera** et **MapR**.



D'autres distributions Hadoop commencent à voir le jour. Par exemple, il y a **Pivotal HD** (Pivotal Hadoop Distribution) d'**EMC Corporation** ou **IBM InfoSphere BigInSights**. Ou même **Amazon Elastic MapReduce** (EMR) où Amazon offre une solution hébergée, pré-configurée dans son cloud.

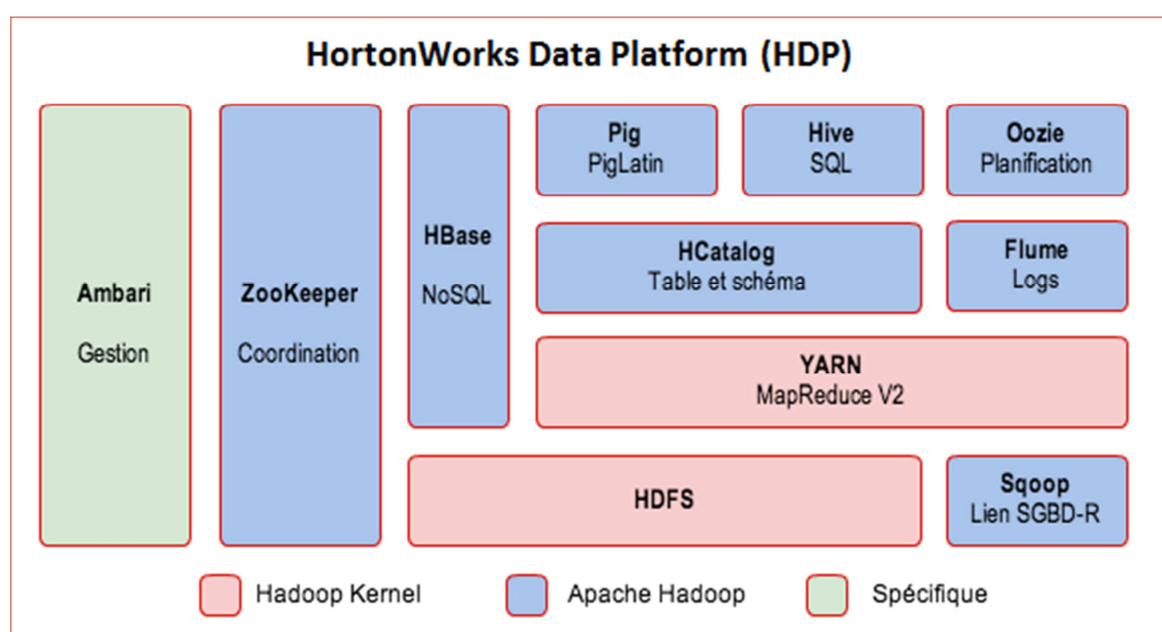
Beaucoup d'autres éditeurs de logiciels ne développent pas leur propre distribution Hadoop, mais travaillent ensemble avec l'un des éditeurs existant. Par exemple, **Microsoft** est partenaire d'**Hortonworks**, en particulier pour déployer **Apache Hadoop** sur son système d'exploitation **Windows Server** et sur leur service de cloud **Windows Azure**. Un autre exemple est Oracle qui offre une appliance big data qui combine matériels et logiciels d'**Oracle** avec la distribution **Cloudera** d'Hadoop. Certains fournisseurs tels que **SAP** ou **Talend** offrent du support pour plusieurs distributions [87].

1. HortonWorks :

HortonWorks a été formée en juin 2011 par des membres de l'équipe **Yahoo!** en charge du projet Hadoop. Leur but est de faciliter l'adoption de la plate-forme Hadoop, c'est pourquoi tous les composants sont open-source et sous licence Apache.

HortonWorks est le seul fournisseur qui utilise 100% du projet open-source Apache Hadoop sans ses propres modifications. Il est le premier vendeur à utiliser les fonctionnalités Apache HCatalog pour des services de méta-données. Il a effectué des améliorations au niveau du cœur Apache Hadoop ce qui le rend exécutable nativement sur les plateformes Microsoft Windows incluant **Windows Server** et **Windows Azure** [87].

Composants de la plateforme HDP :



Les éléments suivants composent la plateforme "**HDP**" d'HortonWorks [88] :

- Cœur Hadoop (HDFS/MapReduce).
- NoSQL (Apache HBase).
- Méta-données (Apache HCatalog).
- Plate-forme de script (Apache Pig).
- Requêtage (Apache Hive).
- Planification (Apache Oozie).
- Coordination (Apache Zookeeper).
- Gestion et supervision (Apache Ambari).
- Services d'intégration (HCatalog APIs, WebHDFS, Talend Open Studio for Big Data, Apache Sqoop).
- Gestion distribuée des logs (Apache Flume).
- Apprentissage (Apache Mahout).

HortonWorks a signé des partenariats importants avec **IBM**, **Microsoft**, **Teradata** et **Talend**.

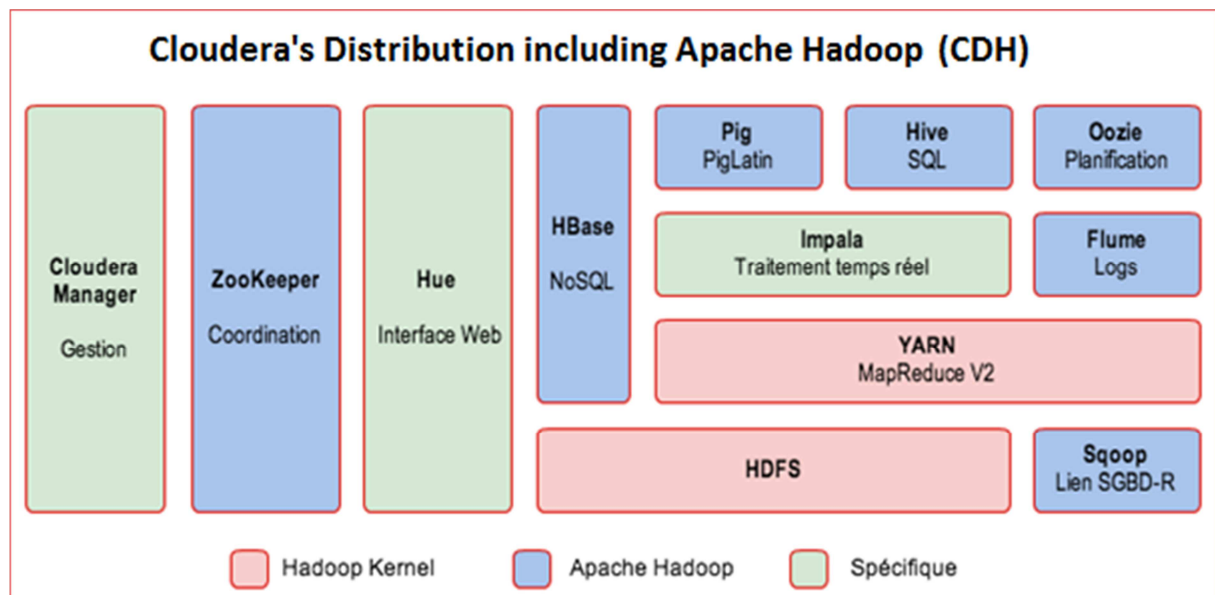
2. Cloudera :

Fondée par des experts Hadoop en provenance de Facebook, Google, Oracle et Yahoo, Cloudera se veut comme la compagnie commerciale d'Hadoop. Elle est à l'origine de la première distribution historique d'Hadoop et est actuellement la distribution Hadoop la plus utilisée dans le monde [88].

« Cloudera est la plateforme du big data pour le stockage de volumes massifs de données, en partenariat avec des sociétés qui font de l'analytique. Nous sommes la première distribution packagée de Hadoop open source. Et en 2012, nous avons annoncé Cloudera 4, la première plateforme big data analytique d'entreprise mature. » [89].

Si La plate-forme de Cloudera est en grande partie basée sur Hadoop d'Apache, elle est complétée avec des composants propriétaires de Cloudera. C'est le cas du projet **Cloudera impala** pour offrir des traitements temps réel de big data, ou encore tout récemment, **Cloudera Search**, un moteur de recherche pour données big data, intégré à la CDH.

Composants de la plateforme CDH :



La plateforme "CDH" de Cloudera est composée de [88] :

Composants 'Apache' :

- **HDFS** : Système de fichiers distribué.
- **MapReduce** : Framework de traitement parallélisé.
- **HBase** : Base de données NoSQL (accès read/write aléatoires).
- **Hive** : Requêtage de type SQL.
- **Pig** : Scripting et requêtage Hadoop.
- **Oozie** : Workflow et planification de jobs Hadoop.
- **Sqoop** : Intégration de bases SQL.
- **Flume** : Exploitation de fichiers (log) dans Hadoop.
- **ZooKeeper** : Service de coordination pour les applications distribuées.
- **Mahout** : Framework d'apprentissage et de datamining pour Hadoop.

Composants d'origine 'Cloudera' :

- **Cloudera Impala** : Moteur d'interrogation temps réel qui permet d'exprimer des requêtes au standard SQL sur des données HDFS (Hadoop Distributed File System) ou HBase.
- **Hue** : SDK permettant de développer des interfaces utilisateur pour les applications Hadoop.
- **Whirr** : Bibliothèques et scripts pour l'exécution d'Hadoop et de services liés dans le cloud.
- **Cloudera Manager** : Déploiement et gestion des composants Hadoop.

Cloudera a signé un partenariat avec plusieurs acteurs dont : **IBM, HP, Oracle**.

3. MapR :

MapR a été fondée en 2009 par d'anciens membres de Google. Bien que son approche soit commerciale, **MapR** contribue à des projets Apache Hadoop comme HBase, Pig, Hive, ZooKeeper et surtout Drill.

MapR se distingue de la version d'Apache Hadoop par sa prise de distance avec le cœur de la plate-forme. Ils proposent ainsi leur propre système de fichiers distribué ainsi que leur propre version de MapReduce : **MapR FS** et **MapR MR** [88].

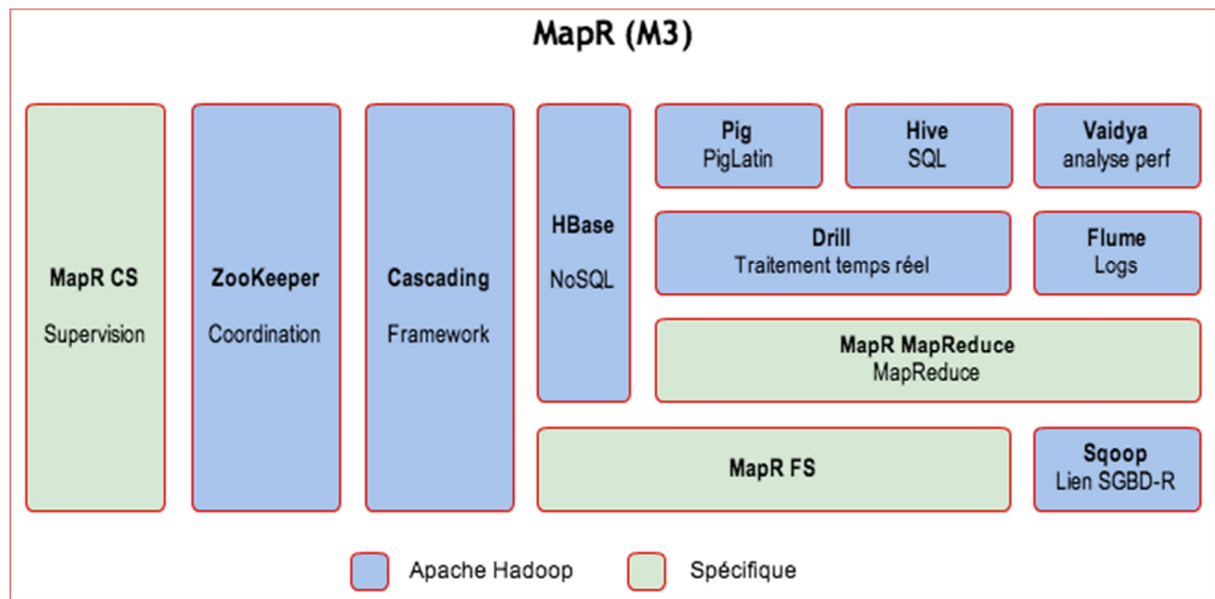
Trois versions de leur solution sont disponibles :

- **M3** : version open source.
- **M5** : Ajoute des fonctions de haute disponibilité et du support.
- **M7** : Environnement HBase optimisé.

MapR a remporté plusieurs succès commerciaux depuis sa création.

- Un partenariat avec EMC pour la création et le support d'une version spécifique de la plateforme Hadoop d'EMC.
- MapR est à l'origine de la version cloud de MapReduce d'Amazon : Elastic Map Reduce (EMR).
- Enfin ils ont été retenus par Google pour l'offre Big Data de Google Compute Engine (GCE).

Composants de la distribution MapR (M3) :



Composants 'Apache' :

- HBase,
- Pig,
- Hive,
- Mahout,
- Cascading,
- Sqoop,
- Flume

MapR propose son propre système de fichiers distribué (MapR FS) en remplacement de HDFS.

Avantages :

- Système plus adapté au mode read/write que HDFS.
- MapR intègre un serveur NFS (Network File System) pour l'intégration au SI de l'entreprise.
- Simplification de mise en œuvre (surcouche du File System de l'OS et non remplacement comme HDFS).
- Plus de Single Point Of Failure.

MapR FS reste compatible avec les API MapReduce/HDFS et HBase.

MapR propose son propre système (MapR MR) en remplacement de MapReduce d'Apache.

Avantages :

- MapR annonce de meilleures performances.
- Entièrement optimisé pour HBase.

MapR Control System (MCS) :

MCS permet la gestion et la supervision du cluster Hadoop. Il prend en charge à la fois les ressources du cluster (CPU, Ram, Disque) que les services et les jobs.

MCS permet de définir des alarmes sur des seuils ou des quotas ...etc.

La visualisation des informations est assurée par le composant HeatMap.

Autres spécificités de la distribution MapR (M3) :

Apache Cascading :

Cascading est un framework Java dédié à Hadoop. Il permet à un développeur Java de manipuler les concepts d'Hadoop avec un langage de haut niveau sans en connaître les API.

Apache Vaidya :

Hadoop Vaidya est un outil d'analyse de performances des jobs Map/Reduce. Son principe de fonctionnement est basé sur des règles qu'il confronte aux statistiques d'exécution des jobs et aux fichiers de configuration. Le rapport est produit au format XML.

Apache Drill :

Map/Reduce a la réputation d'être puissant mais complexe à manipuler. De plus, il est impossible de redéfinir les requêtes à la volée.

Drill est un projet open source réinventé à partir de Dremel de google et vient compléter Map/Reduce en permettant de créer plus rapidement des requêtes sur des données Hadoop en se basant sur le modèle SQL afin d'offrir des traitements temps réels.

MapR a Beaucoup de partenariats de haut niveau et très stratégiques sur le cloud tels que : Amazon Elastic MapReduce et Google Compute Engine.

III.3. La solution NoSQL :



III.3.1. Définition :

NoSQL (Not Only SQL) désigne une catégorie de bases de données proposant des alternatives au langage SQL et au modèle relationnel utilisés dans les bases de données traditionnelles [90].

Les bases de données **NoSQL** sont apparues il y a quelques années pour répondre aux besoins des sites à fort trafic comme Google, Amazon ou Facebook, de disposer de bases de données plus adaptées au stockage d'un volume massif de données. Elles offrent de meilleures performances en lecture/écriture que les bases de données relationnelles et possèdent une grande évolutivité (en terme de scalabilité horizontale). En contrepartie, elles ne possèdent pas toutes les propriétés ACID (Atomicity, Consistency, Isolation, Durability) qui garantissent la fiabilité d'une base de données. NoSQL ne remplace donc pas le modèle relationnel mais offre des solutions intéressantes dans certaines situations où les performances priment sur la cohérence des données.

III.3.2. Histoire et origines :

Les limitations présentes dans les systèmes de gestion de base de données de type relationnel ont obligé les acteurs du monde informatique à trouver de nouveaux moyens pour traiter des données dont le volume croissait de manière exponentielle tout en conservant des coûts de maintenance raisonnables. Le **NoSQL** est apparu comme une solution à ce problème. Apparu en 1998, ce terme acronyme de **Not Only SQL** désigne un nouveau paradigme dans lequel tout est orienté vers la performance et la simplicité. Le NoSQL abandonne donc le concept de modèle relationnel qui avait pourtant été considéré lors de sa création comme une évolution notoire dans le monde des systèmes de gestion de bases de données [91].

Son développement a été favorisé par l'émergence du Web 2.0. Auparavant, la création de contenus était réservée à une poignée d'utilisateurs expérimentés, ce qui limitait grandement la quantité de données à stocker. Le Web 2.0 a fourni des outils facilitant cette production de contenus en la rendant accessible à n'importe quel utilisateur. De ce fait, le volume de données a connu un essor très important. Cette évolution a permis l'apparition des sites Web à forte audience (plusieurs centaines de milliers d'utilisateurs), dans lesquels les bases de données pèsent plusieurs gigaoctets et où les accès se font de manière hautement concurrente. Il a donc également fallu être capable d'offrir un mécanisme de répartition de charge efficace et une scalabilité importante pour y faire face sans faire chuter les performances.

Les années 2000 ont été marquées par des travaux de développement des bases de données NoSQL. Les premières réalisations concrètes arrivent en 2004 lorsque Google commence le développement de BigTable à destination de ses différents produits. En 2008, Facebook développe Cassandra et le publie en open source. En 2009, une communauté commence à se former autour de NoSQL suite à une rencontre à San Francisco entre une centaine d'acteurs du développement logiciel. Au cours de cette conférence, de nombreux outils orientés autour de NoSQL sont présentés, la plupart étant open source et sous licence libre. On assiste alors à une croissance du marché des bases de données NoSQL, et à une diversification des solutions proposées [91].

Contrairement aux bases de données relationnelles, le NoSQL abandonne le principe des tables. On ne cherche plus à savoir de quel type sont les informations stockées. Les données peuvent donc se retrouver de manière complètement désordonnée dans la base de données. L'important est de pouvoir minimiser les temps de lecture et d'écriture pour conserver un système très rapide même lorsqu'il y a de gigantesques quantités de données.

III.3.3. Les Avantages des systèmes NoSQL :

Les solutions NoSQL offrent plusieurs avantages et fonctionnalités dont [92] :

- **Les schémas flexibles** : Les données de NoSQL sont stockées de manière semi-structurée. Il n'y a plus besoin d'avoir un schéma de données parfaitement optimisé pour avoir de bonnes performances. Ceci diminue donc les coûts de conception car la création de nouveaux types de données ne s'accompagne pas de changements de schéma. On peut donc amener de nouvelles fonctionnalités plus rapidement.
- **La gestion de données volumineuses** : Certaines applications génèrent un très grand nombre de données. Par exemple, tout ce qui concerne les logs, les analyses en temps réel, les statistiques. On arrive très vite sur de grands volumes pour lesquels il est nécessaire de conserver des temps de traitement restreints.
- **Un système distribué** : bien qu'il existe aujourd'hui des solutions de réplication et de clustering pour les modèles relationnels, ceux-ci n'ont pas été conçus à la base pour faire face à ce type de problématique. Le NoSQL a été conçu dès le début dans cette optique. Ceci facilite le travail des développeurs.
- **Une haute disponibilité** : la nature distribuée de NoSQL implique une meilleure disponibilité de service. En effet, on peut facilement démarrer de nouveaux serveurs pour faire face à une montée en charge du service, et au contraire une diminution du nombre de serveurs (une panne par exemple) n'affecte pas tout le système.
- **Des capacités de lecture/écriture élevées** : de nombreux procédés sont mis en place pour assurer des débits élevés. Il y a par exemple le système de clé-valeur, ou encore l'algorithme Map/Reduce, la réplication de données, et des mécanismes pour conserver de la consistance et éviter les fautes.

III.3.4. Fonctionnement :

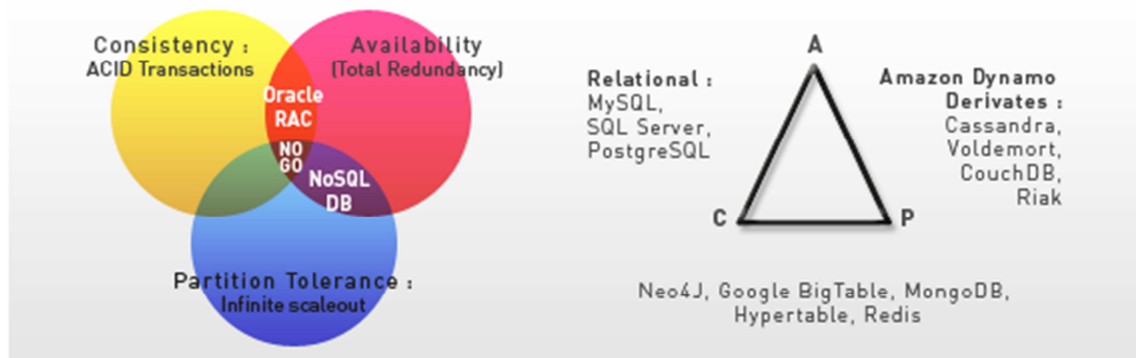
Les bases de données sont très largement utilisées dans les systèmes depuis de nombreuses années. La plupart de ces systèmes se basent sur les transactions pour garantir l'intégrité des données. On parle alors de caractéristiques transactionnelles connues sous l'acronyme ACID (Atomicité, Cohérence, Isolation et Durabilité) [93].

- **Atomicité** : La propriété d'atomicité assure qu'une transaction se fait au complet ou pas du tout : si une partie d'une transaction ne peut être faite, il faut effacer toute trace de la transaction et remettre les données dans l'état où elles étaient avant la transaction. L'atomicité doit être respectée dans toutes situations, comme une panne d'électricité, une défaillance de l'ordinateur, ou une panne d'un disque magnétique.

- **Cohérence** : La propriété de cohérence assure que chaque transaction amènera le système d'un état valide à un autre état valide. Tout changement à la base de données doit être valide selon toutes les règles définies, incluant les contraintes d'intégrité, les déclencheurs de base de données, et toutes combinaisons d'évènements.
- **Isolation** : La propriété d'isolation assure que l'exécution simultanée de transactions produit le même état que celui qui serait obtenu par l'exécution en série des transactions. Chaque transaction doit s'exécuter en isolation totale : si T1 et T2 s'exécutent simultanément, alors chacune doit demeurer indépendante de l'autre.
- **Durabilité** : La propriété de durabilité assure que lorsqu'une transaction a été confirmée, elle demeure enregistrée même à la suite d'une panne d'électricité, d'une panne de l'ordinateur ou d'un autre problème. Par exemple, dans une base de données relationnelle, lorsqu'un groupe d'énoncés SQL ont été exécutés, les résultats doivent être enregistrés de façon permanente, même dans le cas d'une panne immédiatement après l'exécution des énoncés.

On utilise les SGBDR (Oracle, DB2, MySQL, PostgreSQL) pour leurs nombreux avantages (notamment leur capacité à modéliser n'importe quelle structure de données sans redondances ni pertes d'informations). Cependant de nouveaux besoins apparaissent comme la montée en charge horizontale (parallélisation des traitements) ainsi que la haute disponibilité. Le modèle relationnel n'apparaît alors plus comme la meilleure des solutions pour gérer ce type de contrainte car l'énorme volume de données rend le parcours de base difficile avec tous les soucis d'intégrité référentielle et de jointures.

Pour faire face à des volumes importants de données, accédés de différents endroits du monde, il faut pouvoir répliquer ces données sur différentes machines physiques, c'est ce que l'on appelle un environnement distribué. Cependant, Le théorème **CAP** démontre qu'il n'est pas possible d'assurer des transactions totalement **ACID** dans un environnement distribué. [94].



Le théorème **CAP** ou **CDP**, aussi connu sous le nom de théorème de Brewer (en référence à Eric Brewer) dit qu'il est impossible sur un système informatique de calcul distribué de garantir en même temps les trois contraintes suivantes [95] :

- **Cohérence (Consistency)** : Les utilisateurs doivent pouvoir accéder à une même version de données même pendant les mises à jour de celle-ci.
- **Haute Disponibilité (Availability)** : une copie des données est toujours accessible dans le système malgré une défaillance d'une machine d'un cluster.
- **Résistance au morcellement (Partition Tolerance)** : l'utilisateur accède à la globalité du système de manière transparente malgré la segmentation des données sur une multitude de noeuds.

D'après ce théorème, un système de calcul distribué ne peut garantir à un instant T que deux de ces contraintes mais pas les trois.

On peut s'apercevoir que les bases relationnelles sont orientées **Cohérence** et **Availability** (Haute disponibilité). Cependant, elles ont des lacunes en partitionnement des données. Tandis que d'un autre côté, la plupart des systèmes NoSQL privilégient les deux derniers principes (**Availability & Partition Tolerance**), au prix d'un relâchement des contraintes de cohérence des données.

C'est pourquoi en général, les données pour lesquelles la cohérence est importante sont stockées dans des bases de données relationnelles, tandis que les données restantes, pour lesquelles elle importe moins, sont stockées dans des bases de données NoSQL [90].

Pour faire face à une perte de cohérence et maintenir la fiabilité du système, les bases de données NoSQL ont adopté une toute nouvelle approche sous le concept « **BASE** » (Basically Available, Soft-state, Eventually consistent), soit **essentiellement disponible, à un état « Léger », au final cohérent** [96].

BASE est conçu sur les principes suivants :

- **Basic Availability** : Le système doit toujours être accessible.
- **Soft-state** : L'état du système pourrait changer au cours du temps.
- **Eventually Consistent** : La cohérence des données à un instant T n'est pas primordiale. Le système finira par devenir cohérent une fois qu'il cesse de recevoir des entrées.

III.3.5. Types de bases de données NoSQL :

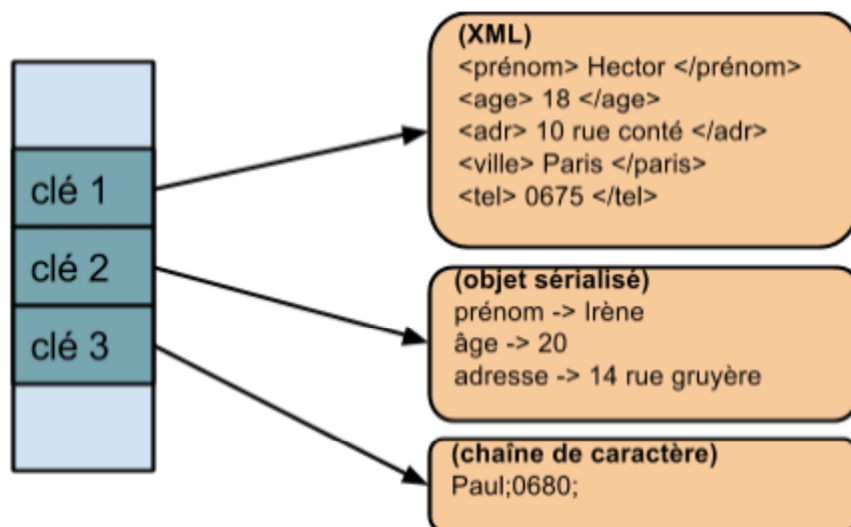
Les solutions NoSQL existantes peuvent être regroupées en 4 grandes familles (modèles) :

1. Le modèle Clé – Valeur :

Le modèle **NoSql clef-valeur** peut être vu à premier abord comme une simple association binaire entre deux colonnes, qui pourraient faire partie d'une table gigantesque dans une base de données relationnelle [97].

La valeur peut être une chaîne de caractère ou un objet sérialisé (objet XML, JSON, ou autres). Elle est identifiée par une clé unique. Par ailleurs, les bases orientées clé-valeur font abstraction du typage ou de la structure (on ne sait pas à l'avance ce qui se trouve dans l'association clé-valeur), il faut donc porter d'avantage l'intelligence de l'applicatif pour interroger la base à défaut du requêtage SQL traditionnel.

Exemple de clé-valeur :



Les particularités du NoSql sont nombreuses. Outre la perte des propriétés ACID, le schéma sacrifie les métadonnées et les laisse à l'appréciation de l'utilisateur. Celles-ci doivent donc être stockées dans un système externe, qui peut lui-même être du NoSql, ou bien du Sql classique (d'où l'appellation « not only Sql »). Une des caractéristiques du NoSql réside aussi dans la modularité de ses clefs. Celles-ci peuvent être composées à volonté suivant un schéma encore une fois prédéfini librement en externe, mais non spécifié dans le modèle. Connaissant ce schéma il est possible de reconstituer une clef.

Par exemple une clef peut être du type « **nom#prénom** ». Cette clef pourrait donc être associée à une valeur prédéfinie en externe, comme l'âge, la taille... A chaque fois donc qu'on veut retrouver la valeur associée à une personne, il nous suffit de concaténer son nom et son prénom puis de requêter la base de données pour obtenir la valeur associée.

Mais la modularité va encore plus loin car la constitution d'une clef est très libre. On peut donc imaginer des clefs du type « nom#prénom#age » et « nom#prénom#taille ». Toutes ces clefs uniques cohabitent donc ensemble dans la même base de données, et sont chacune associée à une valeur [97].

Les valeurs peuvent être vues comme un type complexe, cependant ce n'est qu'un niveau d'abstraction car généralement ces valeurs sont un enregistrement binaire direct. Du coup tous les types de données possibles peuvent être enregistrés directement dans une base de données NoSql, sans se soucier de conversion de données. Ainsi un texte, une image, un objet ou un fichier quelconque sont enregistrés dans la base de données sur un pied d'égalité.

1.1. Solutions NoSQL basées sur ce modèle :

a. Redis :

Redis est un système de gestion de base de données **NoSql** open source, implémentant le modèle **clef-valeur**. Outre son développement dissocié d'un enjeu économique particulier, il est intéressant à plusieurs égards. D'abord parce qu'il introduit la notion de valeurs complexes, mais aussi parce qu'il tourne physiquement différemment que la plupart des SGBD du marché. En effet les bases de données Redis sont exécutées directement en mémoire vive, sa vitesse d'exécution est donc excessivement augmentée par rapport à ses concurrents. Cependant **Redis** reste périodiquement persistant, c'est-à-dire qu'il est possible de déclarer des intervalles modulaires dans lesquels Redis enregistrera un clone de la base de donnée actuelle sur disque dur. Cet enregistrement se trouve d'ailleurs sous la forme d'un fichier unique pouvant atteindre une taille gigantesque. Cette fonctionnalité est intéressante car elle permet de réaliser des back-ups réguliers en interne, par exemple par l'intermédiaire d'un logiciel de versionning branché sur ce fichier [98].

Redis implémente aussi certaines fonctionnalités intéressantes comme la possibilité de donner une durée de vie à un couple clef-valeur, ou bien la possibilité de déclarer des pseudo-transactions par l'intermédiaire de scripts LUA. Ces scripts permettent de regrouper plusieurs opérations en une seule, pour limiter les aller-retour client-serveur. Le programme appelant délègue donc certaines opérations intermédiaires simples au SGBD Redis.

L'autre particularité de Redis est : les types de valeurs composés. Ainsi ce SGBD implémente la notion de listes, d'ensembles triés ou non, mais aussi de Hashes. Ce dernier permet donc par exemple de déléguer une composition de clef à un niveau supplémentaire, vertical cette fois. En reprenant les exemples précédents, la clef « **nom#prénom** » pourrait donc être associée à un Hash avec un champ « âge » associé à une valeur primitive, et un champ « taille » associé à une autre valeur primitive.

Une dernière caractéristique de Redis, qu'on retrouve dans la plupart des SGBD, est de pouvoir déclarer des serveurs maîtres et des serveurs esclaves. Généralement cette dissociation permet de n'autoriser que la lecture aux serveurs esclaves, et donc de déléguer à plusieurs serveurs la charge de répondre aux requêtes de lecture.

 b. Berkeley DB :

C'est une bibliothèque de logiciels qui offre un SGBD de haute performance. Berkeley DB est écrit en langage C avec des API (offrant un appel de fonctions simples pour l'accès et la gestion des données pour, Java, Perl, Python, PHP, C + +, Ruby...etc.) [99].

Les applications qui utilisent Berkeley DB peuvent choisir la structure de stockage qui convient le mieux. Le SGBD prend en charge les files d'attente persistantes, les tables de hashage et les arbres binaires. Les programmeurs peuvent créer des tableaux en utilisant une ou plusieurs de ces structures de stockage en une seule application.

Berkeley DB offre des services de gestion de données importantes, y compris la récupération, la concurrence et les transactions. La bibliothèque fournit des propriétés de transactions ACID strictes par défaut. Pourtant, les applications sont autorisées à les abandonner.

Plusieurs opérations peuvent être regroupées en une seule transaction, et peuvent être validées ou annulées atomiquement. Berkeley DB utilise une technique appelée verrouillage en deux phases pour être sûr que les transactions concurrentes sont isolées les unes des autres, et une technique appelée Write-Ahead pour garantir que les modifications validées vont survivre aux défaillances.

c. Autres :

Plusieurs autres systèmes de gestion de base de données NoSQL utilisent **le modèle clé-valeur** tels que : **Dynamo** (Amazon), **Riak**, **Voldemort** (Linkedin), **Tokyo Tyrant**, **MemcacheDB**, **FoundationDB**etc.

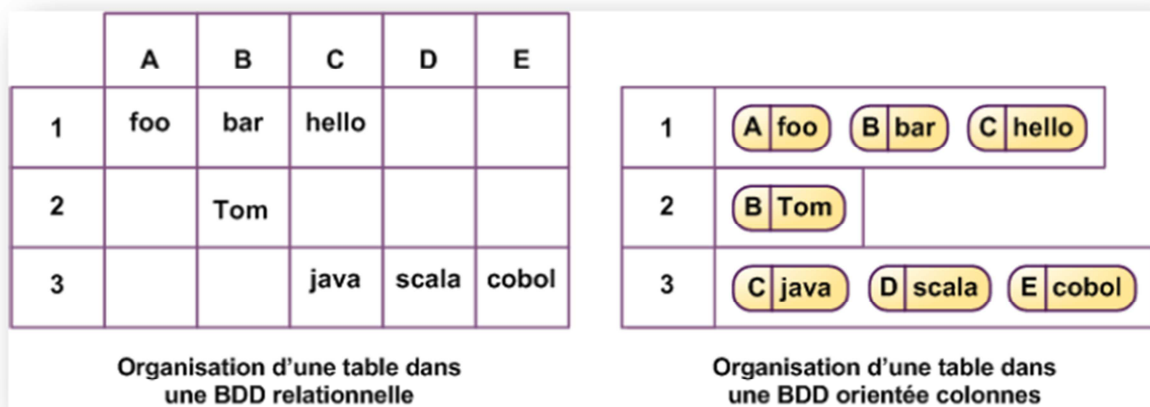
2. Le modèle orienté colonnes :

Le modèle orienté colonnes reprend bien sûr les principes de base du modèle clef-valeur, mais avec des métadonnées supplémentaires. En effet chaque ligne dispose toujours d'une clef pour l'identifier. Cependant une ligne peut disposer d'un nombre gigantesque de colonnes, chacune identifiée par une clef. Cette disposition permet notamment de représenter des relations de type one-to-many sur une seule ligne. La particularité de ce modèle est que des lignes différentes ne sont pas forcées d'avoir les mêmes colonnes [100].

Les caractéristiques faisant de ce modèle un choix intéressants sont nombreuses. Parmi elle on peut citer l'agrégation de plusieurs sous-ensembles similaires entre plusieurs lignes, qui se fait plus rapidement que dans des modèles classiques. En effet la structure de ce modèle permet d'accéder à un sous-ensemble de colonnes sans récupérer les colonnes adjacentes inutiles. De façon analogue la modification d'une valeur concernant la même colonne de plusieurs lignes se fait aussi de façon très rapide. Le choix d'un tel modèle se fait donc accordement aux besoins de l'application.

Ainsi lorsque dans un modèle relationnel classique une application aurait de nombreux champs remplis aléatoirement par un « null », on peut s'orienter vers un modèle orienté colonnes car ces champs vides ne prendraient aucune place mémoire.

Les deux schémas suivants illustrent bien à la fois la différence entre le modèle relationnel et le modèle orienté colonnes.



2.1. Solutions NoSQL basées sur ce modèle :

a. BigTable :

BigTable est un système de gestion de base de données NoSQL, haute performance, propriétaire, développé et exploité par Google depuis 2004 [101].

BigTable repose sur le **Google File System**, et elle permet de gérer des péta-octets de données répartis sur des milliers de serveurs. Les données sont enregistrées dans des tables multidimensionnelles, et identifiées par une clef de colonne, une clef de ligne, et un timestamp [102].

Ce SGBD définit aussi des familles de colonnes, qui regroupent généralement un ensemble de colonnes représentant le même type d'informations. De plus ces familles de colonnes sont compressées pour une meilleure performance. Le **timestamp** est quant à lui associé à une famille de colonne, il permet d'optionnellement récupérer une ancienne version de cette famille si l'application le désire.

Du point de vue infrastructure, un serveur maître gère un groupe de serveurs organisé en cluster. Le maître s'occupe de l'aspect redondance des données et de la haute disponibilité grâce à un service de verrous distribué appelé « **Chubby** ». Ce service multifonctions est par ailleurs critique dans le système car son indisponibilité peut stopper également le cluster BigTable.

Google et son BigTable ont très fortement orienté les pratiques et technologies du NoSQL, notamment avec l'implémentation libre de **Hbase** qui est devenu l'un des projets phares du monde NoSQL.

De nombreux acteurs majeurs du Web utilisent Hbase comme : Ebay, Yahoo, Twitter ...etc. Quant à BigTable, il est resté propriété de google et son code n'a jamais été publié.

BigTable est utilisé sur de nombreuses applications et services de google tels que : **Google Reader, Google Maps, Google Earth, YouTube et Gmail** ...etc.



b. Cassandra :

Cassandra est un moteur de base de données orienté colonnes, écrit en java, lancé par Facebook en 2008. Il s'inspire fortement du modèle **BigTable** de Google avec son mécanisme de sauvegarde et de restauration des données (persistance sur disque) mais emprunte aussi les avantages du modèle **Dynamo** d'Amazon avec son architecture distribuée. Il est depuis 2009 un projet open source appartenant à la fondation Apache [103].

Cassandra comme toute solution BigData permet de manipuler très rapidement un important volume de données. Son modèle orienté colonnes permet aux schémas de données d'être souple : chaque ligne peut avoir un nombre différent de colonnes tout en appartenant à la même famille de colonne.

Concernant son architecture, elle lui permet d'évoluer dans un environnement distribué, intègre des mécanismes de réplication de données et à la possibilité de combiné ensemble plusieurs serveurs Cassandra dans un même cluster.

En conclusion, **Cassandra** est une solution complète et Open source qui permet cumuler les avantages d'un moteur orienté colonnes avec une architecture décentralisée tout en acceptant les capacités de montée en charges automatique comme avec Dynamo. Cassandra est un choix intéressant pour un environnement critique.

Le **15 novembre 2010**, facebook a annoncé qu'il allait désormais utiliser **Hbase** en remplacement de Cassandra. Ce dernier reste tout de même utilisé par plusieurs sociétés dont : **Netflix, Digg, Cisco WebEx, IBM ...** etc.

c. Autres :

Plusieurs autres systèmes de gestion de base de données NoSQL utilisent **le modèle orienté colonnes** tels que : **Sybase IQ, Hbase, HyperTable, ...** etc

3. Le modèle orienté documents :

Les bases de données orientées documents sont une sous-catégorie des bases de données NoSQL qui reprennent le principe du modèle clé-valeur en lui ajoutant des améliorations. Une des limitations du modèle clé-valeur est en effet de ne pas pouvoir structurer la donnée contenue dans la valeur. Ceci rend compliqué la manipulation d'objets complexes, qui peuvent posséder de nombreux attributs. On est obligé de réécrire toute la valeur à chaque fois qu'on veut faire une modification [104].

Le modèle orienté documents définit la notion de document. Un document remplace la valeur transformant de ce fait le modèle **clé-valeur en modèle clé-document**. Il est composé de données structurées que l'on peut voir comme étant un ensemble d'attributs. On conserve la particularité de NoSQL qui est de ne pas avoir de schéma de données, ainsi les documents qui se côtoient dans la base de données peuvent avoir des structures très différentes.

La présence d'une structure rend plus aisée la manipulation des données : à chaque fois qu'on veut lire ou écrire une donnée, on la récupère grâce à sa clé mais ensuite on n'est plus obligé de la traiter dans sa globalité. Comme elle répond à une structure précise, on peut se contenter de travailler seulement sur certains de ses attributs. L'impact sur les performances est évident, particulièrement dans le cadre de la manipulation d'objets complexes. On n'a plus à stocker tout le document en mémoire, mais seulement les attributs sur lesquels on veut travailler.

Le modèle orienté documents permet par ailleurs de créer un système d'indexation des données plus précis que ce qui existe avec le modèle clé-valeur. En effet dans ce dernier cas, l'indexation ne peut se faire que par la clé puisque la valeur n'est vue que comme des données binaires non traitables directement. Avec un document, on peut faire une indexation sur certains champs en particulier, ce qui permet d'optimiser la base de données et les requêtes. On peut alors faire des requêtes NoSQL avancées, en y intégrant des prédicats sur les champs, et non plus seulement des prédicats sur la clé [104].

L'encodage utilisé pour stocker les documents varie d'un produit à l'autre. Parmi les principaux formats utilisés, on retrouve souvent le XML, le JSON et des langages dérivés de ces deux derniers.

Plusieurs éditeurs partagent le marché des bases de données orientées documents en proposant leurs implémentations, cependant les principaux moteurs sont **MongoDB** et **CouchDB**.

3.1. Solutions NoSQL basées sur ce modèle :



a. **MongoDB:**

MongoDB est une implémentation des bases de données orientées documents écrit en C++. Le développement du produit a commencé dès 2007 mais la première version a été publiée en 2009 en licence AGPL par l'entreprise 10gen. Cette entreprise était initialement spécialisée dans le développement de produits complètement open source pour le cloud computing [105].

En 2010, 10gen crée une nouvelle version de MongoDB qui va connaître un grand succès. Suite à cela, 10gen abandonne ses activités de cloud computing et devient MongoDB Inc. pour se concentrer uniquement sur le développement de MongoDB. Aujourd'hui, MongoDB est devenu le leader dans le domaine des bases de données NoSQL et le 6ème toutes plateformes de bases de données confondues.

Du fait de son appartenance aux langages NoSQL, MongoDB ne définit pas de schéma de données spécifique. On retrouve le concept de document pour stocker les données en base, mais on y ajoute la notion de collection. Une collection est un ensemble de documents qui facilite la hiérarchisation des données dans la base de données. On peut faire une analogie avec le monde relationnel, dans lequel une collection serait l'équivalent d'une table et chaque document serait une entrée de la table.

Les données sont encodées en BSON, une variante du langage JSON qui gère les données binaires. Cet encodage est utilisé de manière interne par le moteur MongoDB. La manipulation des données se fait au travers d'une interface présente sous forme de librairie pour de nombreux langages de programmation. Cette diversité permet d'intégrer MongoDB de façon optimale dans de nombreux projets.

L'architecture de **MongoDB** est basée sur un système de réplication de type maître-esclave. Chacun a un rôle précis. Pour les opérations d'écriture, le client s'adresse obligatoirement au maître qui va enregistrer les données et les copier vers tous ses esclaves. Pour les opérations de lecture, le client s'adresse indifféremment à un des esclaves qui va lui renvoyer les données appropriées. Ce système n'est pas sensible aux problèmes de type Single Point of Failure. En effet, dans le cas où un esclave tombe en panne, les données restent disponibles sur les autres esclaves. Dans le cas où c'est le maître qui tombe en panne, les esclaves sélectionnent automatiquement un nouveau maître pour remplacer celui qui est tombé en panne. La répartition de charge est quant à elle assurée par un système de sharding : on forme des groupes composés d'un maître et de ses esclaves. Chaque groupe ne stocke qu'une partie des données, et les requêtes sont réparties sur les différents groupes en fonction de ce qu'on recherche. On a enfin une haute scalabilité puisque des groupes peuvent être ajoutés à la volée pour faire face à une montée en charge [106].



b. CouchDB:

CouchDB est un SGBD orienté documents. C'est un projet sous licence libre qui a vu le jour en 2005. Le projet a ensuite été lié à la fondation Apache à partir de 2008, ce qui a permis d'accélérer son développement pour aboutir à une version stable dès 2010. Il est codé en Erlang, un langage de programmation développé par Ericsson, dont l'avantage est de permettre une très haute disponibilité grâce à des mécanismes avancés de tolérance aux pannes [106].

La particularité de CouchDB est d'embarquer un serveur Web utilisant des protocoles standards. Ainsi, les communications se font via le protocole HTTP supporté par de nombreux clients, tandis que les données sont encodées en JSON et les requêtes en JavaScript. CouchDB est doté d'une API REST (**Representational state transfer**) qui permet de faire toutes les opérations basiques sur les données via les méthodes HTTP standard.

En ce qui concerne la manipulation des données, CouchDB utilise l'algorithme MapReduce. Pour prévenir la perte de données et les éventuelles pannes, il utilise un système de réplication unidirectionnelle d'une base 1 vers une base 2. Pour faire une synchronisation complète entre deux bases, il faut donc faire deux réplications successives. Ce système de réplication est suffisamment avancé pour une utilisation hors ligne : deux bases qui ne peuvent pas être immédiatement connectées entre elles peuvent fonctionner de manière indépendante et se resynchroniser seulement quand la connexion sera revenue.

Pour gérer les éventuels conflits, un système de version est mis en place. Chaque document possède un numéro de version qui correspond au nombre de modifications qui ont été faites sur le document. La fusion entre deux éléments en conflits se fera en fonction des algorithmes choisis par le développeur.

c. Autres :

Plusieurs autres systèmes de gestion de base de données NoSQL utilisent **le modèle orienté documents** tels que : **CouchBase, RavenDB, HyperDex, RethinkDB... etc**

4. Le modèle orienté graphe :

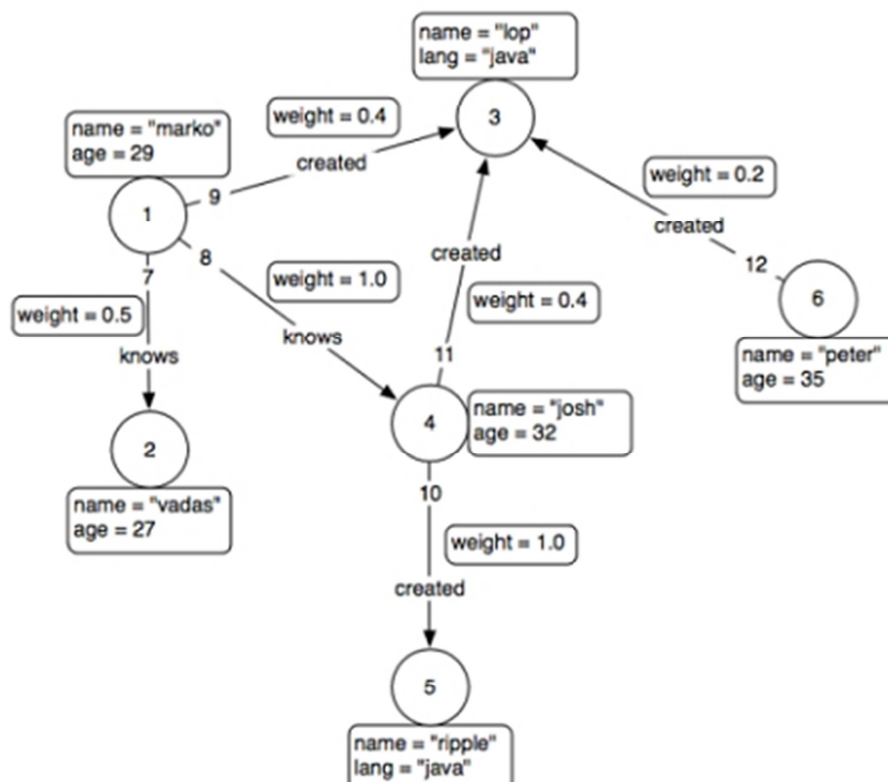
Les bases de données orientées graphe sont largement utilisées dans nos systèmes d'information, notamment avec la théorie des graphes pour le calcul du chemin le plus court, mais aussi dans les réseaux sociaux. Elles permettent de faire abstraction des jointures tout en liant les objets par un système de relation

Le modèle s'appuie sur les notions de :

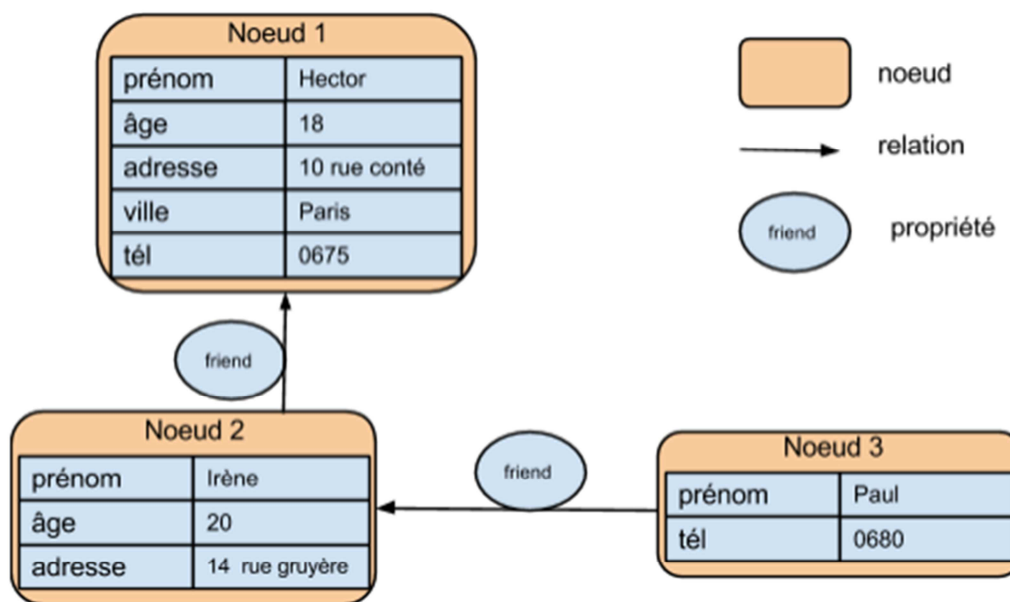
- **Nœud** ou **sommet** (node, vertex) pour définir un objet.
- **Relation** ou **arête** (relationship, edge) avec une orientation et un type (marqué et orienté).
- **Propriété** ou **attribut** (property, attribute) pour définir un nœud ou une relation.

Plus spécifiquement, le modèle est un multigraphe attribué, marqué et orienté. Un graphe marqué a une étiquette pour chaque arête qui est utilisée comme type pour celle-ci. Un graphe orienté autorise des arêtes avec une direction fixée, depuis le nœud source/queue vers le nœud destination/tête. Un graphe attribué autorise une liste variable d'attributs pour chaque nœud et arête, dans laquelle un attribut est une valeur associée à un nom, simplifiant la structure du graphe. Un multigraphe autorise plusieurs arêtes entre deux nœuds. Cela signifie que deux nœuds peuvent être connectés plusieurs fois par différentes arêtes, même si deux arêtes ont la même queue, tête et étiquette [95].

Le schéma suivant montre un petit graphe marqué-attribué:



Exemple d'implémentation dans le modèle graphe :



Cette structure est idéale pour des recherches du type "**partir d'un nœud et parcourir le graphe**" plutôt que "**trouver toutes les entités du type X**", plus adaptées aux bases de données relationnelles traditionnelles. Il est cependant possible en base de données orientées graphes d'effectuer des recherches de ce dernier type, en utilisant un système d'indexation qui pourra selon les cas être interne au graphe (super-nœuds servant à l'indexation) ou au-dessus du graphe (via **Apache Lucene** par exemple) [107].

Les bases de données orientées graphes sont utilisées aujourd'hui dans la modélisation des réseaux sociaux : **LinkedIn** utilise par exemple ce système avec un graphe représentant les personnes et leur relations, et parvient ainsi facilement à afficher le degré de séparation entre chaque contact, qui n'est finalement que la distance entre les nœuds.

4.1. Solutions NoSQL basées sur ce modèle :

La solution majeure pour la gestion des bases de données orientées graphe est **Neo4J**.

a. **Neo4j :**

Neo4j, est une base orientée graphes, complètement compatible avec les transactions ACID, écrite en Java. Les données sont stockées sur disque sous la forme d'une structure de données optimisée pour les réseaux de graphes. Le Noyau Neo4j est un moteur de graphes extrêmement performant, avec toutes les caractéristiques attendues d'une base de données de production telles que la récupération, les transactions par commit en 2 phases, ... etc.

Neo4j peut être utilisé à la fois comme base embarquée sans aucun travail supplémentaire d'administration et comme un serveur à part entière, avec une interface REST (Representational state transfer) pour une intégration facilitée dans les environnements basés sur PHP, .NET ou JavaScript [95].

Le développeur travaille directement sur le modèle de graphe via une API Java qui expose cette structure de donnée flexible. Il existe de bons connecteurs produits par la communauté pour d'autres langages tels que JRuby/Ruby, Scala, Python, Clojure et autres.

Neo4j a des composants optionnels qui viennent en complément du noyau. On peut ainsi structurer le graphe via un méta-modèle, obtenir une implémentation de RDF TripleStore compatible SAIL et SparQL ou encore des implémentations d'un bon nombre d'algorithmes de graphe communs, entre autres.

b. Autres :

Plusieurs autres systèmes de gestion de base de données NoSQL utilisent **le modèle orienté graphe** tels que : **AllegroGraph, Virtuoso, InfoGrid, Sones, HypergraphDB ... etc**

Let $\mathbf{D} = \mathbf{D}_1 \oplus \mathbf{D}_2 \oplus \mathbf{D}_3$ and $\hat{\mathbf{D}} = \hat{\mathbf{D}}_1 \oplus \hat{\mathbf{D}}_2 \oplus \hat{\mathbf{D}}_3$ be the decomposition of $\mathbf{D}$ and $\hat{\mathbf{D}}$ into the direct sum of the three components $\mathbf{D}_i$ and $\hat{\mathbf{D}}_i$ for $i = 1, 2, 3$. Then

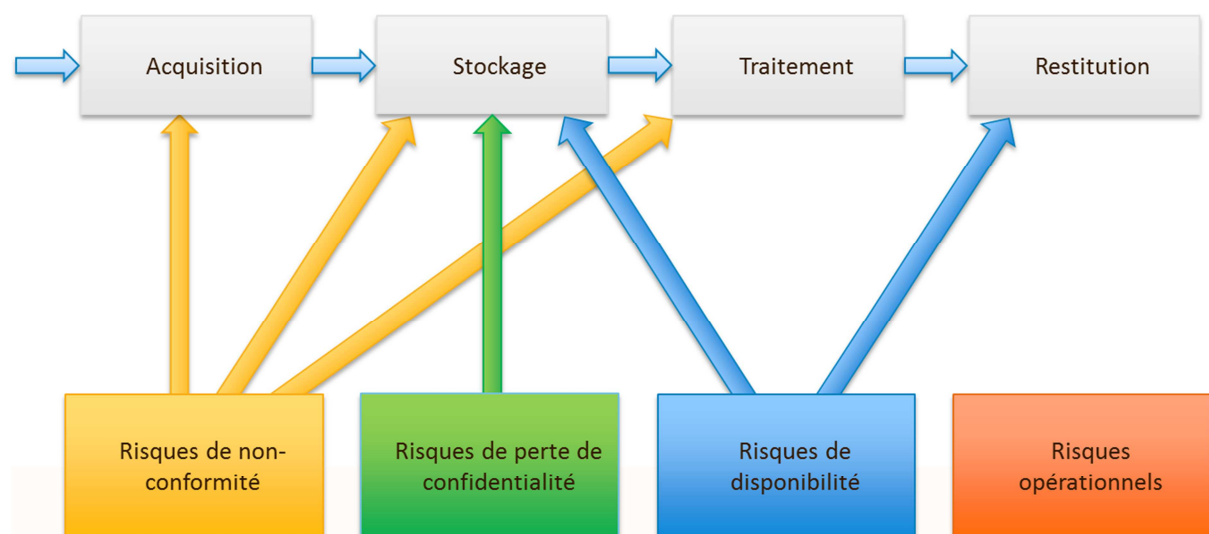
Ainsi, les Big Data vont avoir un impact majeur sur l'industrie de la sécurité et susciter l'avènement de modèles de sécurité intelligents.

Les bonnes questions à se poser :

Plusieurs questions méritent d'être posées et traitées afin de passer son projet Big Data au crible en matière de sécurité [110] :

- A qui appartiennent les données que je collecte, et quelle est leur nature (données spécifiques, à caractère personnel...)
- Quel sont mes fournisseurs externes et quelle est la structure de mon contrat avec eux ?
- Quel est mon objectif avec ces données ? Est-il déclaré ?
- Existe-t-il des contraintes réglementaires particulières ?
- Quels types de traitement vont-ils être réalisés, et par qui ?
- Les résultats seront-ils utilisés légitimement ?
- Sais-je où les données seront stockées, et suis-je capable de les modifier ?
- Combien de temps seront-elles stockées ?
- Les technologies sont-elles maîtrisées ?
- Sont-elles utilisées dans des versions à jour ?
- Sont-elles entièrement maintenues ?
- Chaque type de données est-il protégé de manière adéquate contre les différentes menaces (modification, accès illégitime, vol...) ?

Toutes ces questions seront étudiées à travers une analyse de risques, en partant de l'analyse de la chaîne de traitement des données et en dégagant les risques associés.



IV.1. Analyse des risques :

IV.1.1. Risques de non-conformité :

➤ Dus à la collecte d'informations non déclarées à la CNIL :

Origine : La collecte de données à caractère personnel implique une obligation de déclaration de la collecte et des traitements sur ces données à la CNIL (Article 226-16 du code pénal [111]).

Dans tout ce qui suit, l'autorité française CNIL (Commission nationale de l'informatique et des libertés), sera prise comme référence et représentera toute autorité, de n'importe quel pays, responsable de la protection des droits intellectuels.

En Espagne, l'équivalent de la CNIL est l'AEPD (**Agencia Española de Protección de Datos**).

En Allemagne, son équivalent est l'ULD (**Unabhaengiges Landeszentrum fuer Datenschutz**)

Scénario : Dû à la nature des données non-structurées du Big Data, des informations personnelles et supplémentaires se sont mêlées dans les enregistrements.

Impacts : Un contrôle de la CNIL pourrait résulter en une condamnation, avec une amende, et une perte d'image pour l'entreprise, due à la publication du jugement.

Contre-mesures [112] :

- **Filtrage a priori des données collectées** : Supprimer dès la collecte, les données détectée comme personnels via l'utilisation de patterns connus. Cela permet notamment de supprimer les adresses IPs, numéros de téléphone et numéro de sécurité social. Il est malheureusement impossible d'être exhaustif dans la suppression par "liste noire", ce qui ne fait que réduire le risque.
- **Anonymiser les données** : Anonymiser les données assez rapidement dans le processus de collecte et de stockage pour empêcher l'identification d'un individu à partir de ces données. Le processus est intéressant, il permet de conserver la relation d'appartenance des données sans en conserver le caractère personnelle, mais reste complexe et ne peut être généralisé à tous les types de données personnelles.

➤ **Dus à la durée de conservation des données personnelles :**

Origine : Les textes de loi, ainsi que les autorisations de la CNIL imposent une limite temporelle pour la conservation de données personnelles. Il est possible de garder les données plus longtemps si celles-ci sont anonymisées. (Article 226-20 du code pénal [113])

Scénario 1 : Dû à la nature de la base de données NoSQL utilisées, il se peut qu'il n'existe pas de système permettant d'ajouter des métadonnées pour enregistrer la date de collecte des données, et donc, pas de moyen de déterminer quand celles-ci doivent être supprimés.

Scénario 2 : Le système de stockage ne permet pas de rechercher les enregistrements par leurs métadonnées, et ne permet donc pas de faire des recherches par date d'insertion pour effacer les enregistrements.

Impacts : Un contrôle de la CNIL pourrait résulter en une condamnation, avec une amende, et une perte d'image pour l'entreprise, due à la publication du jugement [114].

Contre-mesures : Annoter les données avec la date de collecte et faire des passes de « suppression »

➤ **Dus à l'incapacité de rectifier/modifier les données - dus à l'incapacité d'accéder directement aux données :**

Origine : Les lois obligent le possesseur d'une base de donnée à mettre à disposition des utilisateurs du service la capacité de : connaître les informations récoltées à son propos, de les rectifier, et de les faire supprimer (Décret n°2005-1309 du 20 octobre 2005, Article 110).

Scénario : Dû à la nature des bases de données non structurées, le système ne permet pas de récupérer les données à partir du nom d'une personne, une recherche exhaustive est obligatoire. Le problème se pose par exemple avec le traitement des « logs » non anonymisés. La suppression ou la rectification est alors problématique.

Impacts : Un contrôle de la CNIL pourrait résulter en une condamnation, avec une amende, et une perte d'image pour l'entreprise, due à la publication du jugement.

Contre-mesures [115] :

- Supprimer un document tout entier si les informations sont contestées.
- Anonymiser les données : Anonymiser les données dans le processus de collecte et de stockage pour empêcher l'identification d'un individu à partir de ces données.

➤ **Dus à un problème d'attribution de responsabilité (incapacité à tracer les actions des administrateurs) :**

Origine : Afin de détecter les modifications anormales et de pouvoir attribuer les modifications d'un système à leur responsable, tracer les actions d'administration est primordial. En outre, certaines normes ou lois imposent de tracer les actions des administrateurs du système [116].

Scénario : Tracer les actions d'un administrateur pour savoir qui a fait des modifications sur les infrastructures.

Impacts : Amendes, Condamnation, Rupture de contrat.

Contre-mesures :

- **Tracer les actions des administrateurs** : Prendre une base qui produit des logs.
- **Stockage sécurisé des logs** : On peut choisir de sous-traiter le stockage des logs pour empêcher un administrateur de pouvoir modifier ou supprimer ceux-ci suite à une action malveillante. Il est aussi possible de stocker les logs et empêcher un administrateur seul d'accéder aux logs, limitant le risque.

➤ **Dus à une incapacité à protéger les données conformément aux réglementations et standards**

Origine : Certaines normes imposent le cloisonnement de certaines données : c'est le cas par exemple avec les numéros de cartes bleues.

Scénario : Les données sensibles à cloisonner sont stockées sur la même infrastructure de stockage que les données non sensibles. Des traitements utilisant ce stockage comme source peuvent accéder aux données sensibles.

Impacts : Non-conformité avec la norme ciblée, pouvant entraîner une rupture de contrat et fuite de données sensibles.

Contre-mesures [117] :

- **Authentification et séparation logique** : Instaurer une séparation logique des données avec authentification obligatoire pour l'utilisation des données sensibles, suivant le système de stockage utilisé.
- **Chiffrer les données** et les déchiffrer à la volée en cas de traitement sur ces données.
- **Séparer physiquement** les systèmes de stockage en infrastructures distinctes.

IV.1.2. Risques de non disponibilité :

➤ Politique de sauvegarde :

Origine : Une attaque, une erreur de manipulation ou un problème technique entraîne une perte de donnée, ce qui peut être un coup fatal pour une entreprise.

Scénario : Suite à une mauvaise manipulation, des données importantes ont été supprimées.

Impact : Perte de données.

Contre-mesures [118] :

La politique de sauvegarde doit intégrer les informations suivantes :

- **Technique de sauvegarde :**
 - o Sauvegarde brute des données sur les disques (sauvegarde peu performante).
 - o Système intégré et distribué de sauvegarde.
 - o Sauvegarde incrémentale basée sur les métadonnées du système de fichiers distribué.
- **La fréquence :**
 - o Fréquence des sauvegardes complètes.
 - o Fréquence des sauvegardes différentielles
- **Les tests de restauration et leur fréquence**
- **Stockage des sauvegardes :**
 - o Sur quel type de stockage ? (bande, disque).
 - o Où sont-elles localisées ? (sur site, hors site, dans le Cloud)

➤ Pannes matérielles :

Origine : Les problèmes de disponibilité suite à des défaillances matérielles sur les équipements réseaux ou ceux de l'infrastructure sont exacerbés par le Big Data du fait de sa mise à l'échelle sur de nombreux serveurs.

Scénario 1 : Suite à une panne matérielle sur équipement réseau non ou mal redondé, une partie des serveurs de stockage est inaccessible.

Scénario 2 : Suite à une panne matérielle sur l'un des serveurs de stockage (mémoire vive défectueuse, CPU en surchauffe, coupure électrique), les données stockées ne sont plus accessibles.

Impact : Perte de disponibilité des données.

Contre-mesures [119] :

- **Réseau entièrement redondé :** il est primordial de redonder le cœur du réseau et les switchs de distribution ainsi que l'ensemble des composants intermédiaires (pare-feu, VPN et répartition de charge).
- **Failover :** les logiciels de stockage et les bases de données NoSQL possèdent des systèmes automatiques de basculement en cas de détection de perte de « noeuds » de stockage ou de calcul. Les activer permet de réduire drastiquement les problèmes de disponibilité car les multiples répliques de la donnée prennent alors le relais.

IV.1.3. Risques de perte de confidentialité :

➤ **Dus au vol/diffusion/fuite d'informations confidentielles :**

Origine : Attaque sur le système menant au vol d'information.

Scénario : Un attaquant s'introduit dans le système, et vole des informations, tel que des N° de Carte bleue.

Impacts : Fuite ou perte d'information, impact sur les clients.

Contre-mesures [120] :

- **Sécurisation de l'infrastructure :** Il est impératif de réaliser des audits réguliers de l'infrastructure et de vérifier, notamment, les permissions d'accès aux données et la configuration des systèmes critiques. Le monitoring constant peut permettre, en cas d'attaque, de détecter celle-ci rapidement et d'empêcher la fuite d'information.

➤ **Dus à un accès à des données confidentielles, par une personne illégitime :**

Origine : Base de stockage mal sécurisée à cause d'une mauvaise gestion des droits.

Scénario : Un employé mécontent s'empare des informations stratégiques d'une entreprise, par un accès à une base de données, depuis l'intérieur, puis les revend à des entreprises concurrentes.

Impact : Fuite d'information.

Contre-mesures [121] :

- Audit régulier de la base des permissions pour détecter les profils laissés actifs par oubli et les permissions trop larges.
- Limitation des privilèges : Gestion fine des droits d'accès. Ne permettre l'accès à la base de données qu'aux comptes administrateurs et aux applications.

IV.1.4. Risques opérationnels :

➤ Dus à la collecte d'informations non pertinentes ou inutilisables

La collecte doit être limitée aux seules données utiles et l'organisation ne doit pas se mettre à « **tout enregistrer** » en vue d'un possible enrichissement des données par corrélation ultérieure. Cela entraîne une perte de place considérable, de temps de collecte et peut saturer inutilement les infrastructures de stockage.

➤ Dus à une dépendance des autres applications à une solution technologique unique

Origine : En l'absence d'un langage de requête intermédiaire, comparable au SQL pour les bases de données, l'hétérogénéité des systèmes Big Data entraîne une dépendance du SI envers une unique solution [122].

Scénario : Du fait de la forte singularité des systèmes Big Data (interfaces non compatibles), en cas de changement de technologie, il faut modifier toutes les applications en dépendant.

Impacts : Augmentation importante du coût du changement de technologie. On se retrouve « prisonnier » d'une seule solution.

Contre-mesures :

- Utiliser plusieurs technologies différentes en parallèle (pour différentes données et traitements).

IV.1.5. Risques liés aux modes d'intégration :

L'intégration d'une plateforme Big Data peut se faire de différentes manières :

- **Interne** : Toute la plateforme est contrôlée par l'entreprise.
- **Externe / Cloud** : L'entreprise utilise une plateforme mutualisée ou un service de type « Cloud ».

Le cas de **l'intégration interne** n'apporte pas de risques supplémentaires si ce n'est la mauvaise gestion de la solution.

En revanche, **l'utilisation d'une plateforme hors du contrôle de l'entreprise** peut poser des problèmes juridiques ou des risques de perte de confidentialité. Ces risques ne sont pas exacerbés par l'utilisation du Big Data, mais sont inhérent aux infrastructures mutualisées sous le contrôle d'une entreprise tierce. Les risques engendrés sont [123] :

➤ **Confidentialité :**

- **Fuite d'information** : La sécurité des données dépend de la sécurité de la solution proposée. Le contrôle sur celle-ci est généralement inexistant et les attaques sur les plateformes de grande envergure sont plus « intéressantes » pour les pirates qui ne manqueront pas de trouver et exploiter les failles

➤ **Légalité :**

- **Localisation des données** : La perte de capacité à localiser l'emplacement géographique de ses données peut poser des problèmes légaux de transfert de données. Les données localisées dans un autre pays tombent sous les lois dudit pays, posant aussi des problèmes de confidentialité.
- **Juridiction et gestionnaire de la plateforme** : bien que mes données soient localisées sur des serveurs de mon pays, si l'entreprise qui gère le service est localisée dans un autre pays, les injonctions émanant de l'état de cet autre pays peuvent-elles lui permettre d'accéder à mes données ?
- **Appartenance des données** : certains services possèdent des clauses de propriété sur les données manipulées sur leurs infrastructures, il y a alors un risque de perte de propriété intellectuelle autant que juridique car on ne contrôle plus les données manipulées.

➤ **Dépendance au Cloud Provider :**

- **Sécurité :** la sécurité des données est laissée au gestionnaire de la plateforme et il est difficile, sinon impossible, de pouvoir auditer la sécurité de celui-ci.
- **Dépendance à une technologie :** le risque de dépendance forte à un acteur unique et à sa technologie sous-jacente est exacerbé car celui-ci met à disposition non seulement le matériel, mais aussi le logiciel (qui varie suivant le service).

Tous ces risques doivent alors être pris en considération lors d'une intégration dans le Cloud ou sur une plateforme gérée par un organisme externe. La **contre-mesure** consiste à internaliser les traitements et à posséder sa propre infrastructure, mais cela demande des capacités techniques de pointe et peut demander des investissements conséquents [124].

Conclusion Générale & Perspectives

Tout au long de ce travail, nous avons découvert une nouvelle facette du monde qui nous entoure, celle de la Data. Qui eût cru que notre monde générerait autant de données au quotidien ? Qui eût cru que visiter des sites ou effectuer des recherches sur internet pouvait servir à nous identifier, à comprendre notre façon de penser, décrire notre personnalité, connaître nos passions Etc. Et de ce fait devenir des clients potentiels pour des entreprises.

Mais les Big Data ce n'est pas que sur internet, nous avons vu que l'information peut venir de n'importe où que ce soit des téléphones mobiles, des tablettes, des capteurs ... etc. Aujourd'hui avec l'évolution de la technologie on peut même ajouter : Les lunettes connectées, les montres, les téléviseurs, etc. Désormais on équipe même les ballons de football de capteurs. Pour ainsi dire que tout ce qui nous entoure dans la vie est source potentielle d'information pour peu qu'on sache la capter et l'analyser.

Concernant les domaines d'application, ils sont nombreux et tous les domaines qui gèrent et génèrent des données sont concernés car comme on l'a déjà vu, l'apport de l'analyse des Big Data est énorme d'une valeur inestimable, on se rend pas compte à quel point les données que nous manipulons peuvent être d'une grande utilité.

On peut citer quelques domaines qui peuvent tirer profit des Big Data :

En recherche médicale, la recherche sur l'ADN génère 1 To de données par expérimentation et nécessite un fort besoin de traitement analytique. Big Data peut répondre à ces problématiques.

En géologie, un séquenceur pour calcul sismique génère 2 To par an et des millions de fichiers associés qu'il faut prendre en compte, traiter et interpréter. Seul Big Data a la capacité à assurer ces traitements en un temps raisonnable.

Les responsables marketing qui veulent pouvoir faire un mailing bien ciblé à partir des données sur les réseaux sociaux comme Facebook peuvent utiliser Big Data pour acquérir, qualifier, traiter et analyser les données.

Les compagnies pétrolières doivent assurer l'exploration et la surveillance avec un pilotage temps réel et une détection rapide des problèmes. Big Data est capable de traiter les millions d'informations collectées par l'ensemble des capteurs.

Dans le domaine des télécommunications, c'est des milliards d'évènements qui sont gérés par chaque opérateur téléphonique, rien qu'un abonné individuel peut générer à lui tout seul des centaines voire des milliers d'informations par jour. Tout opérateur doit être en mesure d'analyser ses logs à tout instant.

Partant de ces constats, nous déduisons que l'implantation du Big Data dans les entreprises doit venir d'un besoin fort du métier car la complexité de la mise en place d'une telle solution ne doit pas être sous-estimée. Une solution Big Data entraîne un coût et des compétences à acquérir qui ne sont pas forcément rentables si un besoin spécifique n'est pas clairement établi.

Par ailleurs, il est toutefois intéressant de se projeter en émettant l'hypothèse que toutes les données peuvent être utilisées un jour dans une solution Big Data, et que la mise en corrélation de celle-ci avec d'autres données peuvent entraîner de nouvelles opportunités. Reste à savoir si l'on dispose des moyens et compétences pour se lancer et surtout savoir bien choisir le SGBD et le modèle de base de données NoSQL qui convient le mieux pour nos besoins.

A la fin, nous avons vu que les Big Data n'ont pas que des avantages, ils peuvent également poser problème si on ne sait pas les utiliser ou que les données n'ont pas été vérifiées au préalable. Parmi les problèmes qu'on peut rencontrer, il y a ceux liés aux données personnelles. Il ne faut pas confondre données publiques et données privées. Pour éviter des problèmes avec la justice, les données à caractère personnel doivent être déclarées à l'autorité concernée.

Nous pouvons citer par exemple le cas de Google et facebook qui sont condamnés plusieurs fois par la CNIL pour violation de vie privée, et à ce jour, personne ne sait où vont les données collectées par ces deux géants du web et surtout quel usage il en est réellement fait.

Perspectives

L'avenir des Big Data est très prometteur et ses possibilités d'utilisation sont nombreuses et ne font que grandir avec l'évolution de la technologie. Aujourd'hui plusieurs constructeurs automobiles investissent dans les Big Data pour créer des SmartCars (voitures intelligentes sans chauffeur), qui peuvent parcourir des centaines de kilomètres et plusieurs villes sans aucune assistance humaine. Ces voitures ne serviront pas que pour le déplacement, elles serviront également pour la récolte d'informations.

Un nouveau domaine commence à voir le jour, le « **Car Analytics** ». A travers ce dernier, les constructeurs de véhicules visent à combiner les données techniques remontées des véhicules avec des informations de positionnement géographique.

L'objectif étant de saisir le comportement du conducteur dans un contexte particulier. On ne conduit pas de la même manière sur une route de montagne qu'en ville ou encore moins que sur une autoroute. De là, l'objectif est bien d'affiner la conception du produit en cohérence avec l'analyse marketing qui pourra être réalisée sur cette base. Nous pouvons alors imaginer de nouveaux services numériques qui pourraient être bâtis à partir des données générées par les voitures, et surtout d'analyser les besoins du client dans ce domaine.

De nombreux travaux et expérimentations sont menés dans ce domaine, et **Mercedes-Benz** a annoncé l'année dernière, la commercialisation du premier système de communication inter-véhicule. Baptisée **Car-to-X**, cette technologie va permettre l'échange d'informations entre des modèles de la marque ainsi qu'avec des infrastructures de gestion du trafic. Elle est disponible via l'option **Drive Kit Plus**. Mercedes a également mis à disposition de ces clients, une application mobile gratuite **Digital DriveStyle** qui transforme le véhicule en **émetteur-récepteur**. La technologie Car-to-X repose sur une connexion cellulaire en reliant le smartphone au véhicule grâce à une station d'accueil.

Concrètement, le conducteur connecte son smartphone au système multimédia de sa voiture sur lequel il recevra à l'avance des alertes pour des dangers potentiels se trouvant sur son trajet : véhicule en panne, accident, ralentissement, obstacle, croisement dangereux, etc.

Dans le domaine de la médecine, Le volume des données constituant le dossier médical a largement augmenté, notamment avec l'intégration de l'imagerie médicale. Le début de la généralisation des études portant sur la génomique (discipline de la biologie moderne qui étudie le fonctionnement des organes) va également produire un volume considérable de données. Par conséquent, il faudra conserver les données brutes pour pouvoir réétudier les bases de données en fonction de l'avancement de la recherche médicale, et ce, sur une longue période.

Grâce aux Big Data, les médecins seront en mesure de suivre leurs patients à distance : Lire les données de leurs pacemakers, mesurer leurs taux de cholestérol, leur tension artérielle, ... etc.

Google travaille sur des lentilles de contact connectées pour aider les diabétiques à gérer leur taux de sucre. Ces lentilles intelligentes sont capables de mesurer le taux de glucose dans les larmes d'une personne.

Dans le domaine de l'électroménager, les appareils de votre maison seront tous connectés et interactifs, ils s'occuperont des tâches ménagères de manière autonome, votre réfrigérateur commandera du lait et du yaourt tandis que votre téléviseur se chargera de vous trouver les meilleurs films et séries du moment, sans aucune intervention humaine.

Glossaire

IDC : (International Data Corporation)

Une entreprise américaine spécialisée dans la réalisation d'études de marché dans les domaines des technologies de l'information et de la communication et de l'électronique grand public. Fondée en 1964, son siège est basé au **Massachusetts, USA**.

Data Mining : Le Data Mining, ou exploration de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

Gartner : c'est une entreprise américaine de conseil et de recherche dans le domaine des techniques avancées. Elle mène des recherches, fournit des services de consultation, tient à jour différentes statistiques et maintient un service de nouvelles spécialisées. Son siège social est situé à **Stamford, Connecticut**.

Médias Sociaux : L'expression « **médias sociaux** » recouvre les différentes activités qui intègrent la technologie, l'interaction sociale, et la création de contenu. Les médias sociaux utilisent l'intelligence collective dans un esprit de collaboration en ligne. Par le biais de ces moyens de communication sociale, des individus ou des groupes d'individus qui collaborent créent ensemble du contenu Web, organisent le contenu, l'indexent, le modifient ou font des commentaires, le combinent avec des créations personnelles

Crowdsourcing : Un des domaines émergents de la gestion des connaissances, il consiste en l'utilisation de la créativité, de l'intelligence et du savoir-faire d'un grand nombre de personnes, pour réaliser certaines tâches. Pour les éditeurs de sites, il consiste à utiliser les internautes pour créer des contenus, répondre aux questions d'autres visiteurs, voire participer à la conception du site. En mutualisant les ressources et compétences de leurs visiteurs, les sites peuvent alors proposer des produits et services à des coûts très bas.

Google Analytics : c'est un service d'analyse d'audience d'un site Web gratuit proposé par Google. Avec environ 85 % du marché mondial, Google Analytics est le service d'analyse de visites de sites internet le plus utilisé au monde.

Google propose des outils pour les développeurs afin d'accéder aux données de Google Analytics : un service Web est disponible et permet à chacun de récupérer des informations sur les comptes dont il est propriétaire.

Business Drivers : c'est une ressource, processus ou condition qui est vitale pour la continuité du succès et du développement de l'entreprise. Ce terme est largement devenu populaire notamment dans le domaine du management, il fait référence à toute partie importante d'une entreprise. Pour une société spécialisée en informatique, il représente les différents produits de gamme, les innovations technologiques et une assistance client en continue.

Captain Dash : c'est une société française créée fin 2009 et proposant de créer des tableaux de bord épurés principalement à destination des fonctions marketing de l'entreprise. Surfant sur la mode « big data », Captain Dash rapproche les données des systèmes d'information de l'entreprise avec des données externes à l'entreprise.

McKinsey & Company : un cabinet de conseil auprès des directions générales, créé en **1926**, son siège social se situe à **New York, USA**. En 2007, pour la cinquième année consécutive, McKinsey a été classé à la première position du classement Vault des cinquante plus grands cabinets mondiaux. Il compte parmi ses clients 93 des 100 premières entreprises mondiales, ainsi que plus de 50 gouvernements.

Newsletter : (ou **lettre d'information**) est un document d'information envoyé de manière périodique par courrier électronique à une liste de diffusion regroupant l'ensemble des personnes qui y sont inscrites.

La lettre d'information peut également être téléchargée, à la suite d'un abonnement, depuis un site internet.

Business model : Le modèle d'entreprise ou « modèle d'affaires », est une représentation synthétique censée décrire les principaux aspects de l'activité d'une organisation, tant au niveau de ses finalités (but, offres, stratégies) que des ressources et moyens déployés (infrastructure, organisations, pratiques de diffusion ou distribution, processus et règles de fonctionnement). Il s'agit de garantir l'existence et le partage d'une valeur ajoutée entre parties prenantes sur une période et pour un domaine d'activité clairement identifiés.

Churn : (ou **taux d'attrition**) : est, au cours d'une période donnée, la proportion de clients perdus ou ayant changé de produit et service de la même entreprise. Ce terme est principalement utilisé dans les secteurs des télécommunications et bancaire, notamment autour de la fidélisation aux offres, mesurée par le taux de fidélité.

Cloud : Le **Cloud Computing** est un concept qui consiste à déporter sur des serveurs distants des stockages et des traitements informatiques traditionnellement localisés sur des serveurs locaux ou sur le poste de l'utilisateur.

Patron d'architecture : c'est un patron de conception, c'est-à-dire un modèle de référence qui sert de source d'inspiration lors de la conception de l'architecture d'un système ou d'un logiciel informatique.

JDBC : Java DataBase Connectivity, est une interface de programmation créée par Sun Microsystems depuis rachetée par Oracle Corporation, pour les programmes utilisant la plateforme Java. Elle permet aux applications Java d'accéder par le biais d'une interface commune à des sources de données pour lesquelles il existe des pilotes JDBC. Normalement, il s'agit d'une base de données relationnelle, et des pilotes JDBC sont disponibles pour tous les systèmes connus de bases de données relationnelles.

Sérialisation : En informatique, la sérialisation (de l'anglais américain serialization) est un processus visant à coder l'état d'une information qui est en mémoire sous la forme d'une suite d'informations plus petites (dites atomiques) le plus souvent des octets voire des bits. Cette suite pourra par exemple être utilisée pour la sauvegarde (persistance) ou le transport sur le réseau (proxy, RPC...). L'activité symétrique, visant à décoder cette suite pour créer une copie conforme de l'information d'origine, s'appelle la désérialisation (ou unmarshalling).

JSON : (JavaScript Object Notation – Notation Objet issue de JavaScript) est un format léger d'échange de données. Il est aisément analysable ou générable par des machines. Il est basé sur un sous-ensemble du langage de programmation JavaScript. JSON est un format texte complètement indépendant de tout langage, mais les conventions qu'il utilise ressemblent aux langages descendant du C, comme par exemple : C lui-même, C++, C#, Java, JavaScript, Perl, Python et bien d'autres. Ces propriétés font de JSON un langage d'échange de données idéal.

Scalabilité : La scalabilité est la capacité qu'a l'architecture pour évoluer en cas de montée en charge si nécessaire. On distingue :

- **La Scalabilité horizontale :** Possibilité d'ajouter des serveurs d'un type donné. Par exemple : ajout possible de serveurs d'application web avec répartition de charge.
- **La Scalabilité verticale :** Possibilité d'upgrader un serveur (ajout de processeurs, RAM, disques...).

Solucom : c'est un cabinet de conseil en management et en système d'information auprès de clients grands comptes, créée en 1990. Son siège social est situé à **Paris**.

CNIL : La Commission nationale de l'informatique et des libertés (CNIL) est une autorité administrative indépendante française, chargée de veiller à ce que l'informatique soit au service du citoyen et qu'elle ne porte atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.

Failover : Le failover, ou basculement, est la capacité d'un équipement ou d'une infrastructure à basculer automatiquement sur un/une autre en cas de panne partielle ou totale d'une de ses parties.

Bibliographie & Webographie

- [1] http://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es
- [2] <http://www.commentcamarche.net/contents/104-bases-de-donnees-introduction#utilite-d-une-base-de-donnees>
- [3] <http://www.commentcamarche.net/contents/104-bases-de-donnees-introduction#la-gestion-des-bases-de-donnees>
- [4] <http://www.commentcamarche.net/contents/104-bases-de-donnees-introduction#les-principaux-sgbd>
- [5] http://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es#Typologie
- [6] http://benhur.telug.uqam.ca/SPIP/inf6104/article.php3?id_article=17&id_rubrique=4&sem=2
- [7] <http://www.synomia.fr/fr/notre-vision/synomia-menu-la-data-non-structuree>
- [8] S. Abiteboul and P. Buneman. Data on the Web : From Relations to Semistructured Data and XML. Morgan Kaufmann, 1999
- [9] http://www.kernix.com/blog/les-possibilites-offertes-par-les-donnees-non-structurees_p3
- [10] <http://www.it-expertise.com/les-big-data-un-nouveau-paradigme/>
- [11] <http://www.irt-systemx.fr/simplexite-des-big-data-la-simplicité-au-service-de-la-complexité/>
- [12] http://www.liberation.fr/economie/2012/12/03/donnees-le-vertige_864585
- [13] <http://www.synomia.fr/fr/notre-vision/big-data-et-entreprises>
- [14] http://fr.wikipedia.org/wiki/Big_data
- [15] **Rapport d'étude sur le big data - par : Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – Solucom – EPITA - SRS Day 2012.**
- [16] <http://www.redsen-consulting.com/2013/06/big-data/>
- [17] <http://www.statistique-2013.fr/11-statistique-big-data.shtml>
- [18] <http://davidfayon.fr/2013/01/big-data-mot-numerique-2012/>
- [19] <http://www.awt.be/web/res/index.aspx?page=res,fr,foc,100,208>
- [20] <http://www.astrosurf.com/luxorion/big-data-mining.htm>
- [21] <http://www.reuters.com/article/2012/01/23/us-google-youtube-idUSTRE80M0TS20120123>
- [22] <http://www.numerama.com/magazine/25453-youtube-6-milliards-d-heures-de-videos-sont-vues-chaque-mois-maj.html>

- [23] <http://www.zdnet.fr/actualites/facebook-500-to-de-donnees-enregistrees-quotidiennement-39775333.htm>
- [24] http://fr.wikipedia.org/wiki/Big_data#Volume
- [25] http://fr.wikipedia.org/wiki/Big_data#Vari.C3.A9t.C3.A9
- [26] <http://www.soft-concept.com/surveymag/big-data-fin-des-etudes-marketing-2.htm>
- [27] <http://ibmgefrance.com/2013/03/20/big-data-les-3-v-sont-morts-vive-les-5-v/>
- [28] <http://www.data-business.fr/big-data-definition-explication-des-3v/>
- [29] <http://lecercle.lesechos.fr/entreprises-marches/high-tech-medias/informatiques/221164510/manque-t-v-big-data>
- [30] <http://www-03.ibm.com/software/products/fr/category/SWP10>
- [31] Livre Blanc – **Big Data : Fin ou renouveau du Marketing** - APROGED |ADETEM – Bernard Normier | Jean Marc Goachet.
- [32] <http://www.virtualscale.fr/index.php/big-data/big-data-qu-est-ce-que-c-est>
- [33] <http://ibmgefrance.com/2012/11/15/captaindash-designee-startup-europeenne-de-lannee-lors-de-ibm-smartcamp-berlin-2012/>
- [34] <http://en.wikipedia.org/wiki/CartoDB>
- [35] <http://www.soft-concept.com/surveymag/big-data-fin-des-etudes-marketing-3.htm>
- [36] <http://lecubevert.fr/pourquoi-faire-le-choix-du-big-data-en-entreprise/>
- [37] <http://www.altic.org/decouvrir-notre-offre/business-intelligence/big-analytics/118-hadoop/350-big-data-pour-vos-projets-decisionnels>
- [38] <http://www.soft-concept.com/surveymag/big-data-fin-des-etudes-marketing-4.htm>
- [39] **Les enjeux du Big Data pour l'Entreprise** - These professionnelle -Touria Engohan- MBA ISC PARIS - **Page 34.**
- [40] **Les enjeux du Big Data pour l'Entreprise** - These professionnelle -Touria Engohan- MBA ISC PARIS - **Page 35.**
- [41] **Guide d'achat sur l'intégration du Big Data** – CITO Research | Pentaho – Septembre 2013.
- [42] « **Big Data: The next frontier for innovation, competition, and productivity** » - McKinsey & Company - Mai 2012.
- [43] **Les enjeux du Big Data pour l'Entreprise** - These professionnelle -Touria Engohan- MBA ISC PARIS - **Page 36.**

- [44] **Les enjeux du Big Data pour l'Entreprise** - These professionnelle -Touria Engohan- MBA ISC PARIS - **Page 38.**
- [45] **Les enjeux du Big Data pour l'Entreprise** - These professionnelle -Touria Engohan- MBA ISC PARIS - **Page 40.**
- [46] <http://www.journaldunet.com/solutions/emploi-rh/emplois-et-salaire-informatique-en-2013-2014/data-scientist.shtml>
- [47] <http://www.journaldunet.com/solutions/analytics/metier-big-data-data-scientist.shtml>
- [48] <http://www.lesechos.fr/idees-debats/cercle/cercle-82865-metier-data-scientist-une-hybridation-pas-une-revolution-1001028.php>
- [49] <http://www.journaldunet.com/ebusiness/expert/55180/big-data---enjeux---risques-juridiques.shtml>
- [50] **Analyse des big data Quels usages, quels défis ?** – Commissariat général à la stratégie et à la prospective – **Note D'analyse N°8** – Marie-Pierre Hamel | David Marguerit - Novembre 2013, **Page 4**
- [51] TechAmerica Foundation (2012), **Demystifying Big Data: A Practical Guide To Transforming The Business of Government**, p. 12 ; McKinsey Global Institute (2011), **Big Data. The Next Frontier for Innovation, Competition, and Productivity**, 156 p.
- [52] Siegel E. (2013), **Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die**, John Wiley & Sons.
- [53] **Analyse des big data Quels usages, quels défis ?** – Commissariat général à la stratégie et à la prospective – **Note D'analyse N°8** – Marie-Pierre Hamel | David Marguerit - Novembre 2013, **Page 6**
- [54] Guide de Planification : **Démarrer avec le Big Data** – Intel IT Center – Février 2013.
- [55] Appel à projets : Cœur de filière numérique | **Cloud Computing et Big Data** – investissements d'avenir – Février 2014, **Page 4.**
- [56] Appel à projets : Cœur de filière numérique | **Cloud Computing et Big Data** – investissements d'avenir – Février 2014, **Page 5.**
- [57] **Big Data, tirer parti de l'explosion des données** - François Stephan | Renaud Bonnet – CRIP - Numéro 2 – Janvier 2012.
- [58] Guide de Planification : **Démarrer avec le Big Data** – Intel IT Center – Février 2013, **Page 6.**
- [59] <http://pro.01net.com/editorial/558530/au-coeur-du-big-data-decryptage-du-phenomene-hadoop/>
- [60] <http://fr.wikipedia.org/wiki/Hadoop>
- [61] <http://www.egilia.com/fr/articles/54849-au-c-ur-big-data-decryptage-phenomene-hadoop.html>

- [62] Propos de **Charles Zedlewski**, vice-président produit de Cloudera, l'une des start up pionnières dans le big data
- [63] <http://hadoop.apache.org/>
- [64] http://fr.wikipedia.org/wiki/Hadoop#Hadoop_Distributed_File_System
- [65] <http://blog.khaledtannir.net/2013/10/hadoop-et-mapreduce-hdfs/>
- [66] <http://www-igm.univ-mlv.fr/~dr/XPOSE2012/Le%20framework%20Apache%20Hadoop/architecture.html>
- [67] <http://fr.wikipedia.org/wiki/MapReduce>
- [68] «**MapReduce, Algorithmes de Parallélisation des traitements**» - Khaled TANNIR (Doctorant CIFRE LARIS/ESTI) – "Université de Cergy-Pontoise" - Mars 2011, **Page 8**.
- [69] http://fr.wikipedia.org/wiki/MapReduce#Un_mod.C3.A8le_de_programmation
- [70] «**MapReduce, Algorithmes de Parallélisation des traitements**» - Khaled TANNIR (Doctorant CIFRE LARIS/ESTI) – "Université de Cergy-Pontoise" - Mars 2011, **Page 13**.
- [71] «**MapReduce, Algorithmes de Parallélisation des traitements**» - Khaled TANNIR (Doctorant CIFRE LARIS/ESTI) – "Université de Cergy-Pontoise" - Mars 2011, **Page 17**.
- [72] «**MapReduce, Algorithmes de Parallélisation des traitements**» - Khaled TANNIR (Doctorant CIFRE LARIS/ESTI) – "Université de Cergy-Pontoise" - Mars 2011, **Page 20**.
- [73] «**MapReduce, Algorithmes de Parallélisation des traitements**» - Khaled TANNIR (Doctorant CIFRE LARIS/ESTI) – "Université de Cergy-Pontoise" - Mars 2011, **Page 21**.
- [74] «**MapReduce, Algorithmes de Parallélisation des traitements**» - Khaled TANNIR (Doctorant CIFRE LARIS/ESTI) – "Université de Cergy-Pontoise" - Mars 2011, **Pages 22-23**.
- [75] Guide de Planification : **Démarrer avec le Big Data** – Intel IT Center – Février 2013, **Page 7**.
- [76] <http://averain.blogspot.com/2011/05/v-behaviorurldefaultvml-o.html>
- [77] <http://www.linusnova.com/2013/09/hadoop-pig-traitement-des-flux-de-donnees/>
- [78] <http://www.linusnova.com/2013/09/hive-le-data-warehouse-de-hadoop/>
- [79] <http://fr.wikipedia.org/wiki/HBase>
- [80] <http://www.linusnova.com/2013/09/hbase-la-base-nosql-de-hadoop/>
- [81] <http://www.open-source-guide.com/Solutions/Developpement-et-couches-intermediaires/Big-data/Sqoop>
- [82] <http://www.lemondeinformatique.fr/actualites/lire-apache-fait-passer-sqoop-pour-hadoop-au-rang-de-projet-prioritaire-48428.html>

- [83] <http://blog.octo.com/introduction-a-flume-ng/>
- [84] http://en.wikipedia.org/wiki/Apache_ZooKeeper
- [85] <http://en.wikipedia.org/wiki/Oozie>
- [86] http://en.wikipedia.org/wiki/Apache_Avro
- [87] <http://www.infoq.com/fr/articles/BigDataPlatform>
- [88] <http://blog.ippon.fr/2013/05/14/big-data-la-jungle-des-differentes-distributions-open-source-hadoop/>
- [89] Propos de **Peter Cooper-Ellis** - vice-président Engineering chez **Cloudera**.
- [90] <http://www.iguanesolutions.com/expert/articles/technique/nosql>
- [91] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 8**.
- [92] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 10**.
- [93] http://fr.wikipedia.org/wiki/Propri%C3%A9t%C3%A9s_ACID
- [94] <http://fr.wikipedia.org/wiki/NoSQL#Th.C3.A9orie>
- [95] <http://www.infoq.com/fr/articles/graph-nosql-neo4j>
- [96] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 11**.
- [97] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 13**.
- [98] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 14**.
- [99] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 15**.
- [100] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 18**.
- [101] <http://fr.wikipedia.org/wiki/BigTable>
- [102] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 19**.
- [103] http://librairie.immateriel.fr/fr/read_book/9782212135602/ch010

[104] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 21**.

[105] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 22**.

[106] « **Big Data & NoSQL** » - Ahmed AYADI |Jean-Sébastien CALVIER |Adrian GUARRERA - POLYTECH NICE-SOPHIA, **Page 23**.

[107] http://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es_orient%C3%A9e_graphe

[108] "**Au-delà de la sensibilisation : l'urgence croissante de la gestion des données sur le marché européen des PME**". L'enquête menée par **Iron Montain**, spécialiste des services de conservation et de gestion de l'information, en association avec le cabinet PwC – 2013.

[109] **Le rapport Security Brief de RSA (division sécurité d'EMC)** analyse l'impact des Big data sur les méthodes et technologies sécuritaires - Janvier 2013.

[110] <http://magazine.qualys.fr/conformite-organisation/big-data-risques/>
Chadi Hantouche et Gêrôme Billois – Consultants chez **Solucom** – **GS Days 2014** - Mars 2014.

[111]
<http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006417958&cidTexte=LEGI TEXT000006070719>

[112] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 26**.

[113]
<http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006417977&cidTexte=LEGI TEXT000006070719>

[114] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 27**.

[115] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 28**.

[116] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 29**.

[117] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 30**.

[118] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 31**.

[119] **Rapport d'étude sur le big data - par :** Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom** – **EPITA** - SRS Day 2012, **Page 32**.

[120] Rapport d'étude sur le big data - par : Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom – EPITA - SRS Day 2012, Page 33.**

[121] Rapport d'étude sur le big data - par : Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom – EPITA - SRS Day 2012, Page 34.**

[122] Rapport d'étude sur le big data - par : Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom – EPITA - SRS Day 2012, Page 37.**

[123] Rapport d'étude sur le big data - par : Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom – EPITA - SRS Day 2012, Page 38.**

[124] Rapport d'étude sur le big data - par : Mickaël CORINUS, Thomas DEREY, Jérémie MARGUERIE, William TÉCHER et Nicolas VIC – **Solucom – EPITA - SRS Day 2012, Page 39.**