GRENOBLE ECOLE DE MANAGEMENT
Management Technology & Innovation

Quels facteurs sont à prendre en compte pour passer d'une Business Intelligence d'Entreprise au Big Data

QUELS ARCHITECTURES ET IMPACTS SUR L'ENTREPRISE

Thibaud Ryden | Mastère spécialisé en Business Intelligence | 2012-2013

TABLE DES MATIERES

REMERCIEMENTS

Avant de présenter mon travail réalisé pour ma thèse professionnelle, il est important de citer les personnes qui, en m'aidant, ont contribué à l'élaboration de ce document. Qu'elles soient au sein de mon entreprise ou de mes écoles l'ESIEA et GEM, je souhaite exprimer mes remerciements à toutes les personnes de la liste suivante :

- Tout d'abord, mon tuteur en entreprise, pour m'avoir donné l'opportunité de travailler dans ses équipes.
- L'ensemble des chefs de projets qui m'ont accompagné durant toute mon alternance.
- Les étudiants de l'ESIEA avec lesquels j'ai partagé mes cinq années de formation d'ingénieur.
- Les étudiants de GEM avec lesquels j'ai partagé ma dernière année d'études.
- Mes tuteurs école GEM et ESIEA qui m'ont accompagné dans la rédaction de chacun de mes documents.
- L'ensemble du corps professoral que ce soit de l'ESIEA ou GEM sans qui je n'aurais pas pu acquérir les compétences et connaissances que je possède à ce jour.
- Toutes les personnes interviewées qui m'ont aidé à développer ma réflexion dans ce document et qui m'ont beaucoup appris sur le Big Data.

Je remercie également toutes les personnes que je côtoie tous les jours et qui m'ont soutenu tout au long de mes études.

INTRODUCTION

Un système décisionnel est « la capacité de présenter les interrelations entre des faits de telle sorte que cela permette de guider les actions pour atteindre le but espéré» (Luhn, 1958). Cette première définition de la Business Intelligence BI a bien évolué ; de nos jours, la BI correspond à un ensemble de données organisées de façon spécifique, facilement accessibles et appropriées à la prise de décision (Goglin, 2001).

Les systèmes décisionnels d'Entreprise se fondent sur les données dont elles disposent. Le management de ces données a été une grosse problématique du 20ème siècle. Les systèmes permettant de les stocker et de les exploiter n'ont cessé de se développer et d'évoluer en fonction des nouveautés technologiques. Le stockage de ces données se fait sur des bases de données BDD. Initialement, les BDD étaient relationnelles avec un but opérationnel et surtout transactionnel (Inmon, 2002).C'est-à-dire, que les données étaient organisées en tables reliées entre elles par des clés, les liens étant logiques. On parle alors de relation entre les tables. Ce modèle se base sur les propriétés ACID (Atomicité, Cohérence, Isolation et Durabilité). L'Atomicité assure qu'une transaction s'est bien ou mal déroulée. La Cohérence implique que l'état du système est valide avant et le restera après la transaction. L'Isolation permet l'exécution simultanée de transactions apportant le même résultat que l'exécution en série des transactions. La Durabilité assure qu'une transaction terminée restera toujours enregistrée dans le système même suite à une panne.

Puis, petit à petit, la partie analytique et informationnelle a fait son apparition ; il n'était plus uniquement nécessaire d'effectuer que des transactions. Le data Warehouse DWH a donc vu le jour, il s'agit d'une collection de données orientées sujet, intégrées, non volatiles et historisées, organisée pour le support d'un processus d'aide à la décision (Inmon, 2002). En d'autres termes, le DWH est une importante BDD contenant des informations fortement structurées, sur des sujets précis (employés, Chiffre d'Affaires, clients, produits, etc.), destinées à ne pas être modifiées sur une période de temps très large. Il permet de procéder à des analyses de données avec des temps de réponse rapides sur une plus large profondeur d'historique. On parle alors de BDD décisionnelle. L'évolution des performances et des capacités de ces systèmes de stockage ont eu une influence sur les systèmes décisionnels. La BI est devenue plus accessible.

Depuis les années 2000, le volume de données qui transite sur la toile n'a cessé d'augmenter. Le volume de données généré par jour ne cesse d'augmenter ; tous les deux jours, nous créons autant d'informations que du début de l'humanité à 2003 (Schmidt, 2013). Les réseaux sociaux et les plateformes de partage ont ouvert les portes à des informations de plus en plus variées ne pouvant être exploitées avec les technologies existantes. Jusqu'à 2010, les technologies et leurs prix ne permettaient pas de conserver cette quantité énorme d'informations dans des conditions de performance acceptables. Les entreprises étaient dans l'obligation de sélectionner les informations les plus pertinentes et de les utiliser suivant des contraintes de temps et d'interprétation très précises.

La démocratisation du Cloud Computing a permis de dématérialiser les supports physiques, de virtualiser les applications et de distribuer les ressources machines. À ceci, s'ajoutent l'arrivée des supports de stockage ultra rapides SSD (Solide State Drive) ainsi que des technologies permettant d'utiliser ces nouveaux supports pour améliorer les performances. De plus, de nouveaux concepts sont venus challenger les technologies d'exploitation des

données, les nouvelles bases de données apportent de nouvelles méthodes adaptées à ces problèmes de vitesse, volume et variété.

Le Big Data apparait comme la nouvelle rupture du management des données au même titre que les bases de données relationnelles l'ont été dans les années 1970. Le Big Data se définit comme les concepts d'exploitation de données de types variés, avec des temps de réponses rapides et sur de grands volumes permettant d'améliorer la précision des analyses et la prise de décision (Oxford Dictionnary, 2013). Gartner prédit également que d'ici à 2015, 85% des 500 plus grandes organisations ne seront pas en mesure d'exploiter les avantages compétitifs du Big Data (Gartner, 2013). Le Big Data est donc un sujet d'actualité qui va beaucoup évoluer au cours de la prochaine décennie.

Le principe du Big Data est de pouvoir gérer de gros volumes de données à haute vitesse. Les données peuvent être structurées (données définies par un type, un nom et étant stockées dans des BDD relationnelles) ou non structurées (photo, vidéo, son, fichier texte, etc.). Chaque année, la quantité de données générée se sépare entre 30% de structurées et 70% de non structurées. Pour exploiter tous ces différents types de données en conservant des temps de réponses rapide, il faut passer par de nouveaux types de BDD dite non relationnelles s'affranchissant des contraintes ACID. Différents types ont vu le jour comme le Not Only SQL NoSQL ou encore l'écosystème Hadoop regroupant un ensemble de solutions. Toutes ces nouvelles BDD sont nées de l'initiative des Géants du web comme Amazon, Facebook, Google, qui ne pouvaient pas traiter leurs données avec les BDD relationnelle. Certaines bases sont spécialisées dans un format de données bien particulier; par exemple les BDD spatiales pour exploiter les données de géolocalisation.

Ces nouvelles solutions s'appuient sur des traitements parallélisés. L'intérêt du Big Data et de ses technologies est de pouvoir séparer le traitement d'un gros volume d'informations sur différents serveurs de façon à profiter de la puissance de calcul de chacun de ces serveurs pour traiter de plus petits volumes de données. Généralement, un cluster contient un ensemble de racks, chaque rack contient des serveurs et chaque serveur possède plusieurs nœuds. Le principe est alors de prendre un gros volume de données en entrée, de le séparer en petits volumes et de les répartir sur les nœuds. À l'intérieur d'un cluster, les connexions sont proches et de hautes performances, l'échange d'information entre les nœuds est donc favorable. Le Cloud Computing a un rôle clé dans le Big Data. En effet, le Cloud met à disposition un ensemble de serveurs pour échanger des informations. Il est alors possible d'utiliser ces serveurs pour y faire des traitements.

Le Big Data apporte de nouvelles technologies, nécessite de nouvelles compétences et demande un engagement important de la part des entreprises. Dans notre étude, l'entreprise X est prise comme référence car elle dispose d'une culture de la donnée et possède des architectures BI variées. La BI reste un premier pas pour gérer ses données et prendre des décisions, le Big Data est une nouvelle couche de management de ces données. Toutes les entreprises n'ont pas le même niveau de maturité vis-à-vis de la donnée. La transition au Big Data va dépendre de sa culture d'entreprise et de sa maturité en termes de traitement de l'information.

Cependant, quels facteurs pourraient impacter une entreprise souhaitant ajouter la couche Big Data à celle de la BI ? Le fait est que les solutions Big Data dépendent énormément de l'entreprise. Dans la suite, nous définirons clairement ce qu'est la BI et le Big Data. Puis, nous

aborderons différents scénarios d'implémentation d'une solution Big Data sur un système décisionnel existant et répondant à des besoins métiers précis. Enfin, une étude sera faite sur les facteurs à prendre en compte ainsi que sur les impacts possibles d'une telle implémentation sur la gouvernance, la transformation des processus métiers et des méthodes de management.

LA BUSINESS INTELLIGENCE

La Business Intelligence ou informatique décisionnelle représente l'ensemble de moyens permettant d'améliorer la compétitivité des entreprises en exploitant intelligemment les informations dont elle dispose pour prendre des décisions. Apparus dans les années 1980, les outils de BI n'ont cessé de s'améliorer pour satisfaire au mieux les besoins des utilisateurs au sein des entreprises. Avec l'arrivée d'internet et du développement des systèmes d'information, de nouveaux concepts sont apparus, ils appartiennent à la BI 2.0.

De ces concepts découlent de nouvelles opportunités. Les nouvelles technologies ont permis de briser la barrière du stockage, les contraintes de temps de réponse, etc. Internet a donné naissance à de nouveaux comportements de consommation, sociaux. Ces comportements font que les données disponibles explosent. De plus, les nouvelles technologies disponibles depuis le début 2010 permettent de tirer profit de ces données pour extraire de la valeur et prendre des décisions encore plus précises. Ces nouvelles opportunités sont représentées par le Big Data. Les différents aspects de la BI 2.0 se retrouvent dans le Big Data mais sur des échelles differentes.

1. Concepts

La première définition de la BI a été énoncée par Hans Peter Luhnoffre en 1958 ; de ce premier point de vue, la BI était : « la capacité à appréhender les interrelations entre des faits disponibles de manière à guider l'action vers un but désiré ». La recherche de prise de décision en rapport avec des évènements connus était déjà bien ancrée. Petit à petit, la BI a pris forme pour fournir à partir des années 1980 les premiers offres et outils. Ce n'est qu'à partir des années 1990 que la BI s'est installée dans les entreprises.

Initialement on parlait de systèmes d'information transactionnels pour les systèmes de gestion de bases de données. Ils permettaient d'organiser les informations de références (clients, commande, fournisseur, employés) de manière relationnelle. La transaction d'une commande est alors plus simple; par exemple, si un client achète un produit B à la date XX/YY/ZZZZ pour tant d'euros, alors sa représentation en base de données se devra d'être logique.

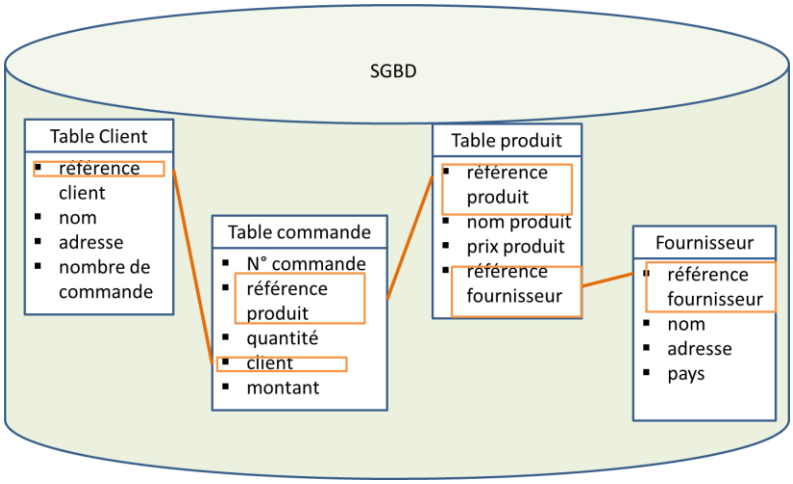

Figure 1: BDD transactionnelle

La BI a amené les systèmes décisionnels qui sont un ensemble de données organisées de façon spécifique, facilement accessibles et appropriées à la prise de décision. En comparaison aux systèmes transactionnels, Les systèmes décisionnels organisent les données pour piloter les activités de l'entreprise. Ils sont optimisés pour faire accéder facilement aux données utiles de façon à prendre des décisions. Les systèmes décisionnels s'appuient sur les systèmes transactionnels et donc respectent les propriétés ACID. Ceci implique que ces systèmes appartiennent bien aux systèmes relationnels. Ils exploitent donc des données hautement structurées exploitables via le langage, de base pour adresser les bases de données qu'elles soient transactionnelles ou décisionnelles, le SQL (Structured Query Language). De plus, le SQL est le langage sur lequel s'exécutent tous les logiciels de BI.

De nos jours, les projets de BI ont toujours autant de succès, l'importance d'avoir un système d'information décisionnel est devenue cruciale que ce soit pour des contraintes légales ou stratégiques. La BI permet aux entreprises de contrôler leurs activités, de mesurer leur performance et surtout de prendre des décisions pour aligner leur stratégie d'une part dans toute l'entreprise et d'autre part, par rapport au marché. La prise de décision se fait via un processus qui se découpe en plusieurs étapes de la chaine de la BI ; on trouve toujours des sources, un processus d'alimentation, de stockage et de restitution.

2. La chaine BI

2.1. Les sources

Les sources sont le point d'entrée de la donnée, ce sont elles qui fournissent un ensemble d'informations aux BDD dédiées à la BI. Au début de la chaine, quand un client passe une commande sur un site d'internet, il donne un certain nombre d'informations personnelles et à celle-ci s'ajoutent les informations des produits qu'il a achetés. Qu'elle soit effectuée en face à face dans une agence ou via un site internet, les informations seront, d'une manière ou d'une autre, saisies dans le SI. À ce moment-là, les données seront envoyées à une ou plusieurs sources ; ces sources alimenteront ensuite des BDD spécifiques. Ces systèmes sont dits opérationnels, car ils récoltent les données brutes venant d'applications diverses. Plus communément, il s'agit des systèmes transactionnels comme les ERP (Entreprise Resource Planning) et CRM (Customer Relationship Management). Un certain nombre d'applications web (Google, Twitter, Facebook) fournissent des données. Enfin des fichiers textes, CSV ou Excel peuvent également être considérés comme sources.

Système transactionnel

Un système transactionnel est un système informatique où un ensemble de transactions sont réalisées. Ils peuvent impliquer une ou plusieurs applications, acteurs et départements d'entreprise. Chaque action engendre une mise à jour sur une BDD. Les applications peuvent communiquer entre elles via les différentes transactions. Généralement, les plus courants en entreprise sont les ERP et CRM.

- Un Entreprise Resource Planning ERP est un outil permettant de centraliser les données et fonctions de gestion de l'entreprise. Il contient beaucoup de données utiles dans la prise de décision ; ils permettent de décloisonner les fonctions de gestion de l'entreprise pour unifier les services et ainsi constituer des sources de données transverses et stratégiques (Fernandez , 2008).

- Le Customer Relationship Management CRM représente les outils de gestion de la relation client. Il regroupe les informations de type vente, marketing et service client pour assurer et optimiser la relation avec le client dans le but de le conserver dans la durée (Fernandez, 2008).

Les ERP et CRM sont des systèmes transactionnels car les données y sont stockées et évoluent constamment au fur et à mesure du processus. Ils fournissent des sources de données transverses constituant une très grande partie des données utilisées dans les systèmes décisionnels.

Applications

Les BDD peuvent être alimentées par différents types de sources. En effet, lors de la saisie dans le SI des données brutes, les données sont par la suite utilisées via des applications qui traitent différents types d'informations. Les applications peuvent être internes à l'entreprise ou externes comme les interfaces web. Par exemple, une personne s'inscrivant sur un site web d'une entreprise (recrutement, demande d'informations, inscription à un forum, etc.) devra saisir un certain nombre d'informations comme nom, prénom, e-mail, téléphone ; ces informations seront par la suite enregistrées dans une BDD pour de futures exploitations (ERP, CRM, BI, etc.).

Fichiers externes

Une source peut également être ni plus ni moins qu'un fichier. Le fichier peut soit alimenter une BDD en étant intégré directement à l'intérieur de celle-ci, soit être lu indépendamment par les outils d'alimentation du système décisionnel. Ces outils sont indispensables pour un système décisionnel.

Peu importe le type de source, il faut passer par un processus d'alimentation le plus utilisé étant l'ETL.

2.2. Le processus d'alimentation

Le processus, Extract, Transform and Load ETL, est la procédure qui consiste à extraire les données provenant des sources, de les transformer pour les rendre compatibles avec les standards de nommages (nom, type, table cible, filtre, etc.) et enfin de les charger dans la BDD.

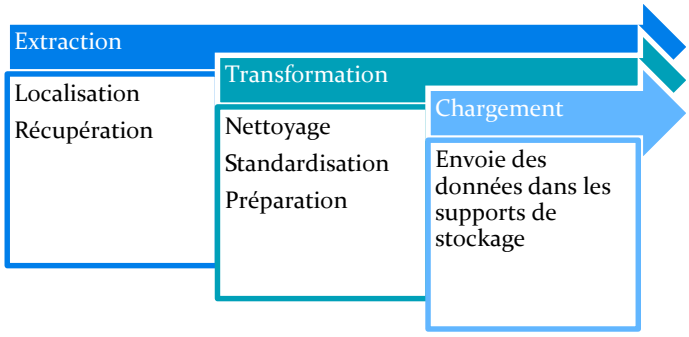

Figure 2: Processus ETL

La procédure doit tenir compte de ce qu'elle a en entrée (données fournies par les sources) et de ce qu'elle délivre en sortie. Typiquement, un flux provenant d'une source contient un ensemble de colonnes avec des données de qualités variées. La sortie de l'ETL est généralement un ensemble de tables de plusieurs BDD ne devant contenir que certaines lignes et colonnes du fichier source, tout dépend de l'architecture de la BDD cible.

Extraction

Lors de l'extraction, l'ETL s'intègre à une source de données pour récupérer les informations dont il a besoin. Ces données sont généralement brutes avec une structure simple. La première étape est de récupérer ces informations et de les stocker dans les entrées de l'ETL en tant que table ou fichier plat de façon à garantir la cohérence de la source si une erreur survenait lors du processus de transformation. De plus, le fichier ou table créé pour accueillir ces données brutes peut être réutilisé simultanément pour définir différentes sorties (Kimball, 2004).

Transform/Transformation

La transformation des données les formate pour leurs futures intégrations dans le système décisionnel. Il s'agit tout d'abord de les nettoyer, puis de les rendre conformes à l'architecture cible via un processus de standardisation et enfin de préparer le chargement.

- Nettoyage

Le nettoyage consiste à faire un travail de qualité de données. La qualité de données est essentielle et pourtant souvent sous-estimée dans les projets BI. Une mauvaise qualité de données a un impact direct sur les décisions prises ; si les données ne sont pas bonnes dans le système décisionnel, les analyses et restitutions seront faussées et donc les décisions erronées.

Le processus de nettoyage vise à rendre le jeu de données brutes utilisable en sortie. Pour cela, l'intervention humaine est nécessaire pour interpréter et choisir le résultat. Par exemple, un fichier Excel contenant des colonnes ou champs sur un client : les informations de nom, prénom, adresse, téléphone et mail sont essentielles. Cependant, le fichier étant originaire d'une saisie manuelle et humaine, toutes les personnes n'ont pas la même manière de remplir ces informations.

- Le nom peut être:
  - écrit en minuscule, majuscule ou première lettre en majuscule.
  - il peut y avoir des noms composés.
  - des champs peuvent être vides.
  - dans une langue différente ne gérant pas les mêmes caractères.
  - etc.
- Le prénom (de même que pour le nom)
- L'adresse :
  - les types de voies peuvent être totalement écrits ou abrégés (avenue=av=AV, etc.).
  - les adresses peuvent ne pas exister.
  - le numéro de voie peut être manquant ou erroné.

        - etc.
    - Le téléphone :
        - écrit selon le format classique (0144557766) ou avec extension +33144557766.
        - contenir des erreurs (trop ou pas assez de chiffres).
        - ne pas exister 0999999999.
        - autre.
    - L'e-mail
        - nom de domaine inexistant.
        - faute de frappe.
        - etc.

Plusieurs moyens existent pour améliorer la qualité de données directement depuis la source en mettant en place des procédures de contrôles mais elles ne corrigent pas tout et le travail de nettoyage reste donc indispensable.

Cette phase permet de supprimer les données non utilisables, retirer les doublons et sélectionner les données qui seront standardisées pour obéir aux contraintes des BDD.

- Standardisation

La standardisation permet de rendre le futur jeu de données conforme à la sortie attendue en travaillant sur des sources multiples. Le format doit coïncider entre les sources (un nombre à 10 chiffres devra rester un nombre à 10 chiffres, les chaines de caractères devront être définies avec la même taille, etc.).

- Préparation

La préparation consiste à créer la sortie de l'ETL, les supports de stockage. Egalement, une étape d'harmonisation est nécessaire pour que des données venant de sources différentes puissent s'intégrer facilement.

Chargement

Le jeu de données d'entrée une fois nettoyage, standardisé et préparé est prêt à être intégré. Le chargement peut s'effectuer en direct ou via des procédures planifiées. Le plus souvent il s'agit de batch (script s'exécutant sur les serveurs hébergeant les BDD).

Le processus d'alimentation est le plus long, il couvre les ¾ du projet et est celui le plus coûteux (80% des coûts). Une fois effectué, il peut être automatisé pour les traitements récurrents. Les batchs tournent pendant la nuit pour mettre à jour les BDD en sortie. Il peut toutefois s'avérer nécessaire de refaire des traitements ou de les modifier en fonction des évolutions ou des besoins sur les restitutions ou l'arrivée de nouvelles sources.

2.3. Stockages

Le processus ETL permet d'alimenter les zones de stockage du système décisionnel après avoir structuré les données pour qu'elles soient de la plus grande qualité possible pour prendre de meilleures décisions. La dernière étape du processus charge les données vers

l'espace de stockage. Cette espace regroupe différents types de stockages répondant à des fonctions et caractéristiques différentes

Les différents types de stockages

- Le data warehouse, DWH est une collection de données orientées sujets, intégrées, non volatiles et historiées ; organisées pour le support d'un processus d'aide à la décision
    - Orientée Sujet : structurée par thèmes ou contextes d'analyse
    - Intégrée : Une donnée a une description et un codage unique et est qualifiée (contrôle et validation)
    - Non Volatile : Les données sont en lecture seule et une requête doit fournir un résultat constant dans le temps
    - Historiée : Une donnée ne doit pas être mise à jour

  Le DWH stocke l'ensemble des données de détail utiles au processus de décision, assure la qualité et la cohérence des données intégrées dans le système décisionnel, constitue le socle d'information et contient l'historique des données d'entreprise.

- Le data Mart, DTM est une collection de données orientées sujets mises à la disposition des utilisateurs dans un contexte décisionnel
    - Il sert de support aux interfaces décisionnelles
    - Il est orienté « utilisateurs » afin de satisfaire l'ensemble de leurs besoins autour du sujet donné
    - Il est constitué à partir des données du Data Warehouse

  Le DTM a plusieurs rôles, il peut être déployé pour optimiser une problématique fonctionnelle ou de reporting, améliorer les performances, confidentialité en séparant physiquement les données entre différentes divisions métiers.

- Operational Data Store, ODS est un espace de stockage tampon entre les données opérationnelles sources et les données décisionnelles du DWH. Parfois, il se limite à une copie des données de production.

Cubes

Un cube est une représentation multidimensionnelle des données disponibles dans le DWH ou les DTM. Les informations dans un cube reprennent certaines données des dimensions que l'on appelle axes du cube. Les informations des axes peuvent être hiérarchisées, renommées pour que ce soient plus ergonomiques vis-à-vis de l'utilisateur final. Une hiérarchie est une suite logique organisée d'une dimension. Par exemple pour la dimension temps contenant les champs Année, semestre, trimestre mois, semaine, jours, date, etc., la répartition peut se faire telle que :

- Année
- Mois
- Jours

OU

- Année
- Trimestre
- Mois
- Semaine
- Jours

Les informations disponibles se mettant à jour en fonction de la manière dont on navigue dans le cube. Un produit peut par exemple se décliner selon son offre/gamme/série/produit.

Un cube possède à minima 3 axes, au-delà de 3, on parlera d'hypercube ; cependant, pour des raisons de performance et compréhension (globale et analytique), il ne devrait pas contenir plus de 6 axes.

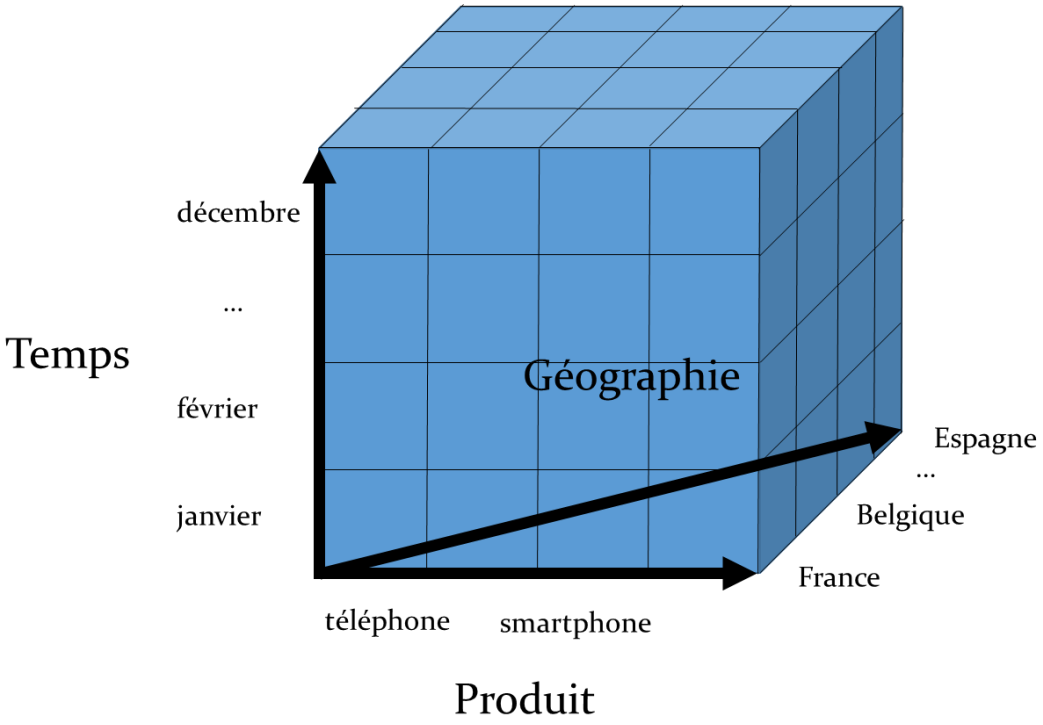

Figure 3: Cube multidimensionnelles

Grâce au cube, il est possible en suivant le schéma ci-dessus, de connaitre le nombre de smartphones vendus en France sur la période janvier à février ou encore le nombre de téléphones écoulés sur le mois d'août en France et en Belgique.

Les cubes se basent sur la technologie OLAP (OnLine Analytical Processing) ; différents types de cubes peuvent être construits : MOLAP, ROLAP, HOLAP.

- OLAP, OnLine Analytical Processing par opposition à OLTP (OnLine Transactional Processing) correspondant à la technologie de calcul des bases transactionnelles des sources. Cette technologie permet d'adresser les DWH et DTM. La technologie OLTP ne peut pas supporter les requêtes des systèmes décisionnels, l'OLAP vient combler cette lacune en offrant des performances adaptées aux requêtes sur de plus gros volumes de données venant de domaines différents. Plusieurs types de construction OLAP permettent de gérer les performances des systèmes : M – R – H.

- MOLAP, Multidimensional OLAP est considérée comme BDD OLAP « pure ». Dans ce cas-là tous les agrégats (indicateurs utilisés avec une fonction d'agrégation comme une somme, produit, comptage, etc.) sont calculés. La complexité du Cube MOLAP est donc plus grande, l'espace disque requis est plus important mais les temps de réponse sont beaucoup plus rapides étant donné que les calculs ont déjà été faits.

- ROLAP, Relational OLAP aucun calcul n'est établi au niveau du cube. Lors d'une navigation dans le cube, les requêtes vont directement adresser les tables du DWH ou DTM sur lesquels le cube est connecté et les calculs des agrégats se font à la volée. Ce genre de cube a des temps de réponse moins bon que les MOLAP et l'espace disque nécessaire est moins important. L'inconvénient est que si plusieurs personnes cherchent à calculer les mêmes agrégats au même moment, les temps de réponse risquent d'augmenter fortement.

- HOLAP, Hybrid OLAP est un mélange entre le M et H. Les agrégats les plus utilisés sont déjà pré-calculés. Cependant, les informations associées aux axes du cube sont requetées sur les BDD.

Tableau 1: Spécifications Cube M-R-H OLAP

Cube	Agrégats calculés	Espace disque	Temps de réponse
MOLAP	Tous	Important exemple 200Mo	Très bon
ROLAP	Aucun	Faible 5Mo	Plus long, risque de crash
HOLAP	Les plus utilisés	Faible 20Mo	Bon

Modélisation

La particularité des supports de stockages décisionnels tient compte de la modélisation. Si les BDD relationnelles opérationnelles sont basées sur une modélisation Entité/Relation, organisées de façon logique ; les systèmes décisionnels ont besoin de plus de flexibilité pour interroger les données avec de meilleures performances et pouvoir faire évoluer le modèle plus facilement.

La modélisation de ces systèmes réside sur une table de fait entourée de tables de dimension. Les systèmes décisionnels visent à prendre des décisions en pilotant la performance. Ce pilotage est possible grâce à des indicateurs ou mesures qui sont déterminés grâce aux

données du DWH. Pour calculer ces indicateurs, il est nécessaire d'établir des faits qui seront les témoins de ce que l'on cherche à mesurer. La modélisation va donc consister à définir les faits à étudier et selon quels critères. Cela se traduit par les faits et dimensions du modèle. Les dimensions sont des tables organisées par thème tandis que la table de fait est centrale ; elle regroupe l'ensemble des liens vers les dimensions et permet de relier les dimensions entre elles tout en calculant les faits.

Par exemple si un décideur souhaite suivre l'évolution des ventes des produits de sa gamme sur le premier semestre 2013 en France, cela se traduira par la table de fait et dimensions suivantes :

Dimensions :

- Temps : la profondeur d'historique. Ici on se limite au S1 2013, la table aura donc les champs mois et années
- Produit : produit de la gamme du décideur
- Géographie : on se limite aux ventes en France
- Vente : information de chiffre d'affaires.

Faits :

- Clés : vers les tables de dimensions (généralement ID_produit, ID_pays, etc.)
- Faits :
    - Nombre de produits vendus
    - Chiffres d'Affaires total
    - Etc.

Plusieurs types de modélisation telle qu'Étoile, Flocon ou encore Constellation sont possibles pour le DWH. Le data mart quant à lui est le plus généralement en étoile.

- Etoile

Pour permettre une utilisation facile des données par les utilisateurs, le schéma en étoile est le plus simple.

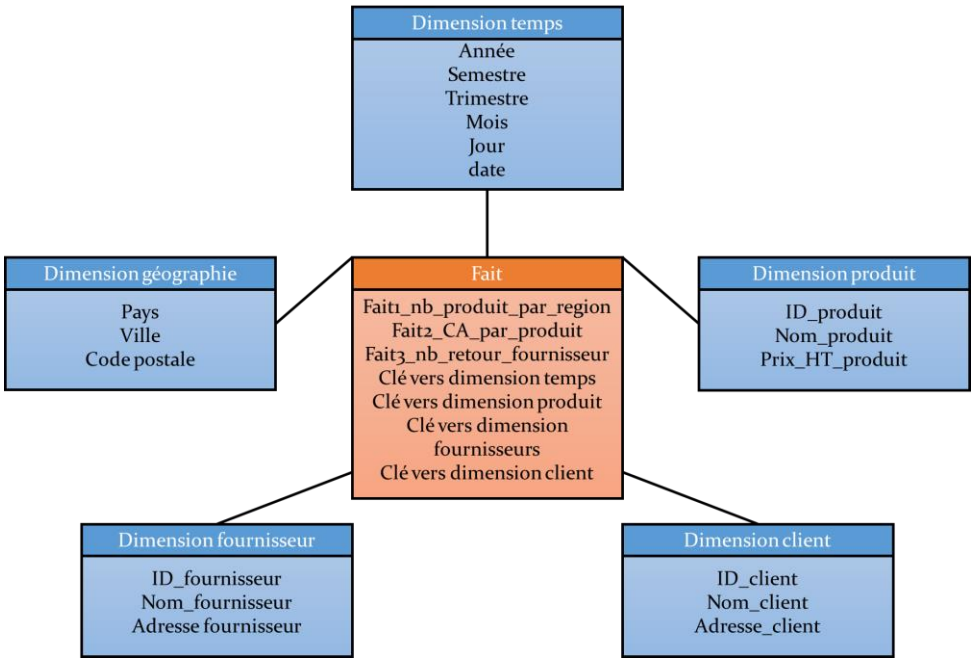

Figure 4: Schéma en Etoile

- Flocon

Le flocon intervient quand le schéma en étoile n'est pas possible. Dans ce cas, le modèle nécessite d'implémenter un schéma relationnel au niveau des dimensions. Celles-ci se déclinent en « sous dimensions » en y mettant des niveaux de détails.

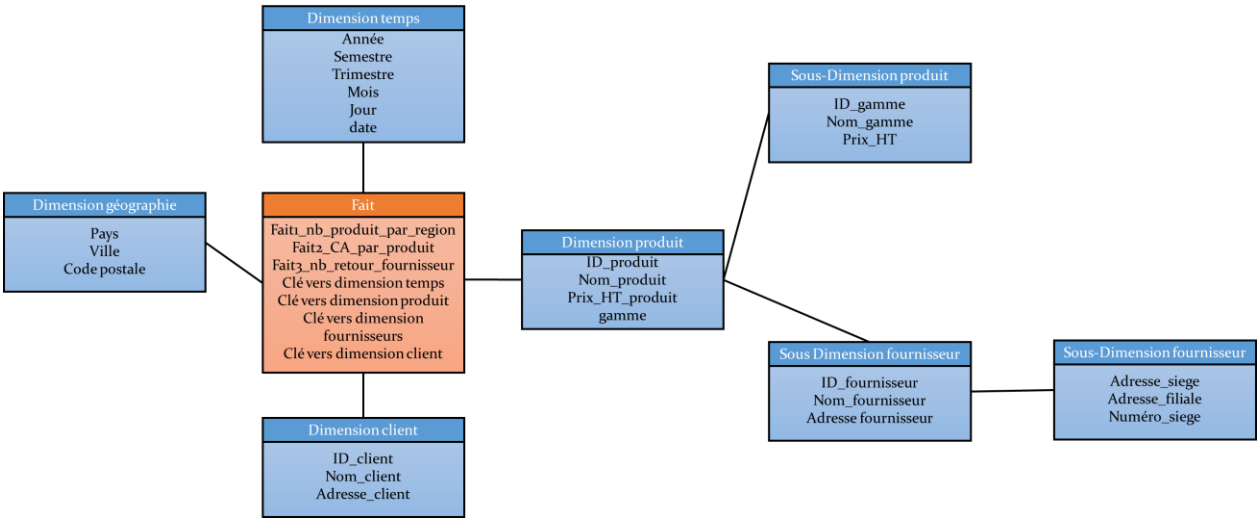

Figure 5 : Schéma en Flocon

- Constellation

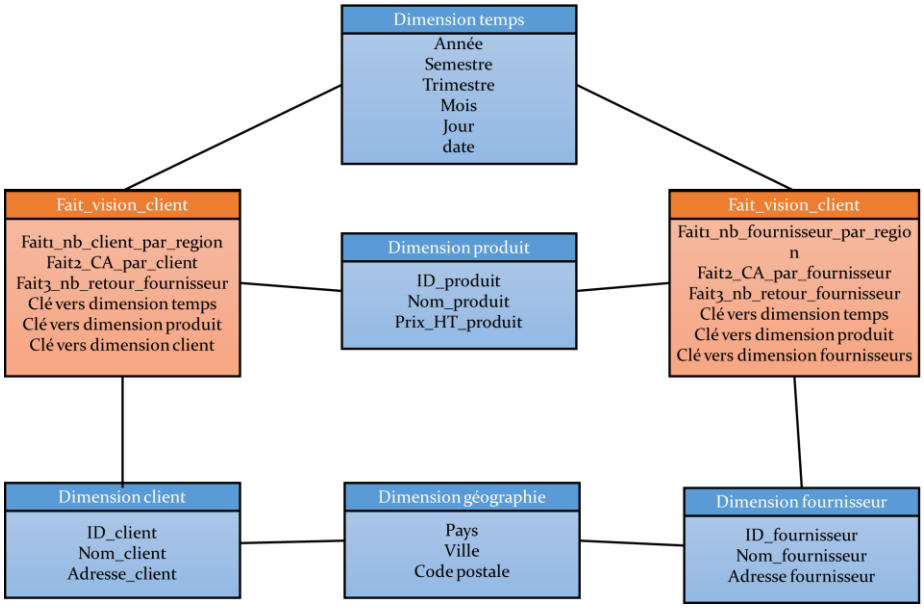

Figure 6: Schéma en Constellation

En suivant la comparaison astronomique, une constellation est un « dessin » d'étoiles. Dans notre DWH, il s'agit donc de schémas en étoile répondant chacun à des problématiques spécifiques avec la particularité d'être reliés entre eux ou d'avoir des dimensions communes.

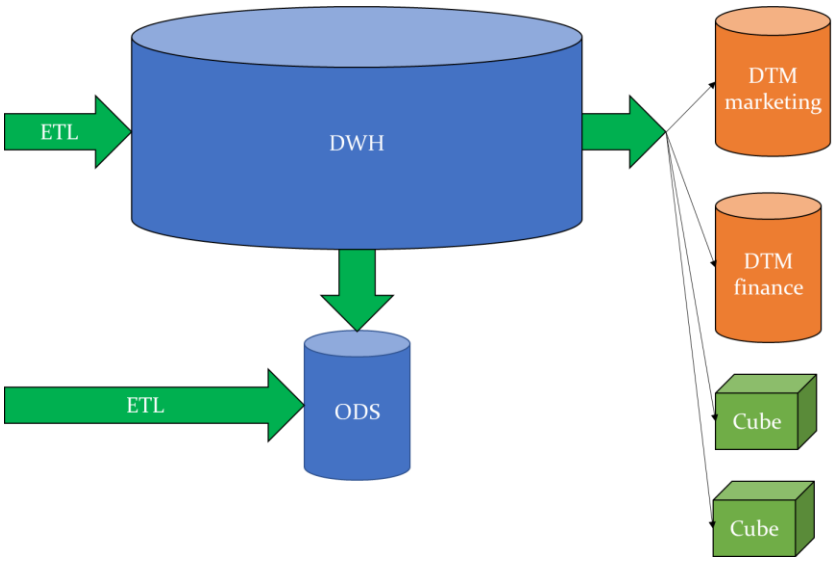

Figure 7: Les supports de stockages en BI

La suite du processus consiste à analyser et exploiter les données précédemment travaillées et stockées pour piloter l'activité. L'étape suivante est une étape d'analyse et de restitution.

2.4. Analyses et restitutions

Pour restituer les informations, plusieurs méthodes sont possibles. Un utilisateur peut adresser un DWH ou DTM directement via un outil de requête SQL, utiliser un cube ou encore passer simplement par des outils BI de Reporting se connectant directement aux BDD et cubes.

Requêtes SQL

Les systèmes décisionnels utilisent des BDD dont la modélisation est spécifique pour améliorer les accès aux données. En d'autres termes, il s'agit de BDD et comme toutes BDD, le langage permettant d'y accéder est le SQL. Avec un outil de requête SQL, le plus souvent celui fourni par le revendeur de la BDD (Oracle, Microsoft, Teradata, etc.), un utilisateur sachant coder en SQL et ayant une bonne connaissance des tables des BDD est en mesure d'effectuer des requêtes plus ou moins complexes.

Une extraction de données des bases est alors possible et celles-ci peuvent être par la suite retravaillées sous un logiciel de traitement de données comme Excel ou encore des outils d'analyses mathématiques plus poussés tels que R, Mathematica, Mathlab, etc.

Reporting

Le reporting est la technique qui par l'utilisation d'outils spécifiques permet de restituer les données sous forme de rapport. Un rapport peut contenir des tableaux, tableaux croisés dynamiques et graphes de tout type. Dans un rapport il est possible d'insérer des filtres pour que l'utilisateur final n'ait accès qu'à certains types d'informations. Les rapports sont connectés directement aux DTM ou cubes. Les outils de reporting peuvent servir à manipuler les données, faire des tableaux de bord ou encore créer des rapports ad/hoc.

La manipulation sert à faire des analyses de données, à déterminer de nouvelles informations et à vérifier l'état de l'entreprise. Par exemple, Excel peut se connecter très simplement à une source de données pour créer des tableaux ou graphes.

Les tableaux de bord sont des rapports mettant en évidence les indicateurs clés du décideur. En un coup d'œil il doit être en mesure de voir l'état de son périmètre.

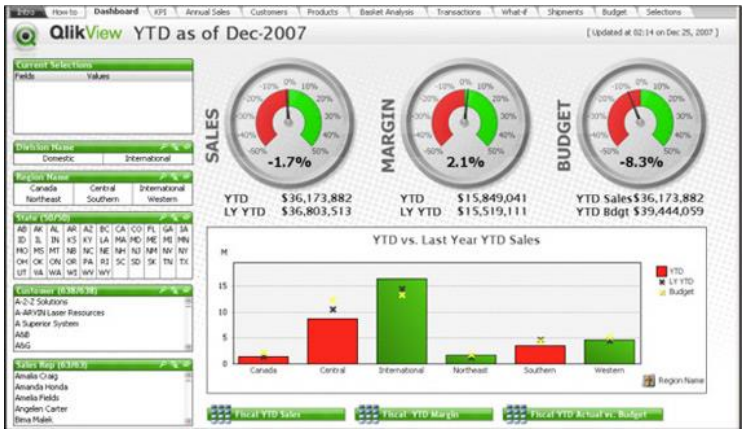

Figure 8: Exemple de reporting sous QlikView (etechnoforte, 2009)

Un rapport ad/hoc est un rapport créé par un utilisateur final ; ce dernier a accès à un certain périmètre de données et peut modifier ou visualiser ce qu'il souhaite analyser.

Data Mining

Pour aller plus loin dans l'analyse de données, au niveau de la BI, les techniques de Data Mining ou fouille de données permet à l'aide d'algorithmes statistiques et méthodes spécifiques de faire parler les données.

Deux familles distinctes sont présentes :

- Descriptives pour décrire la situation actuelle.
- Prédictive pour simuler l'avenir en se basant sur des évènements passés. Les techniques suivantes sont utilisées :

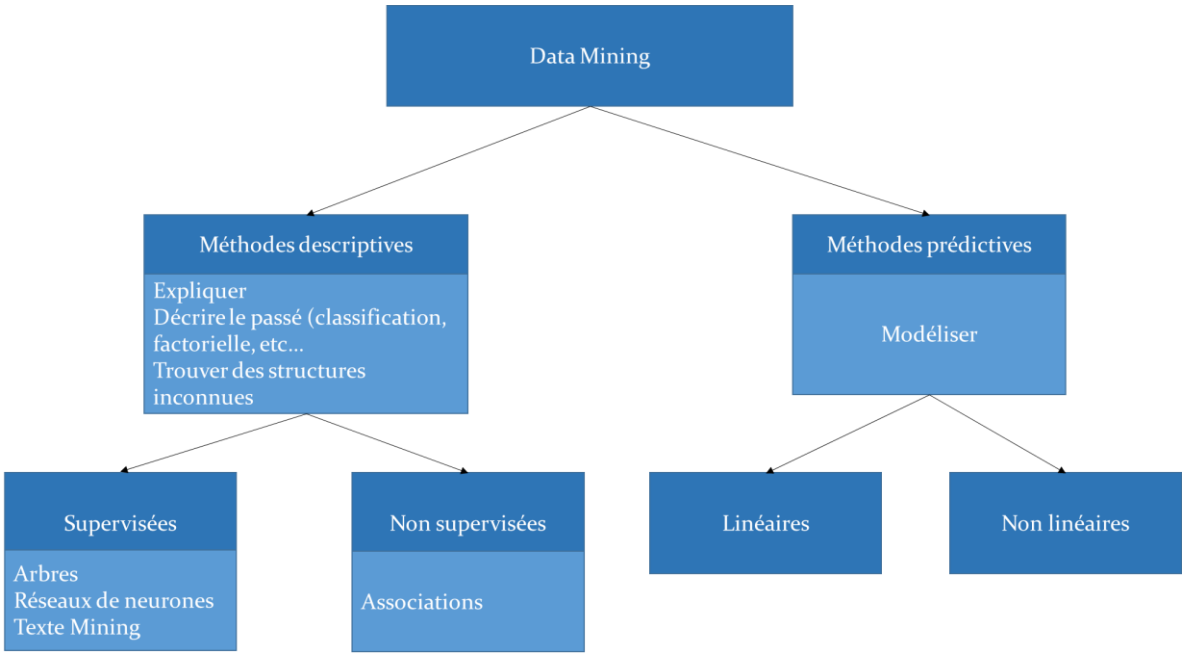

Figure 9: Arbre de décision du Data Mining

3. Data Management en BI

La manière de gérer les données varie en fonction du degré de maturité d'un système décisionnel de l'entreprise. Les plus en retard ne procèdent que par batch quotidien pour mettre à jour les données du DWH et/ou des sources, tandis que les plus en avance peuvent être proches du temps réel avec des mises à jour dirigées par évènement ou *event driven*.

3.1. Batch quotidien pour le DWH

Dans ce cas, les données mises à jour sont les données du DWH et de l'ODS (s'il y en a). Les batchs tournent une fois via l'ETL pour le rafraîchissement. Les outils de reporting peuvent effectuer des requêtes volumineuses avec des temps de réponse faibles.

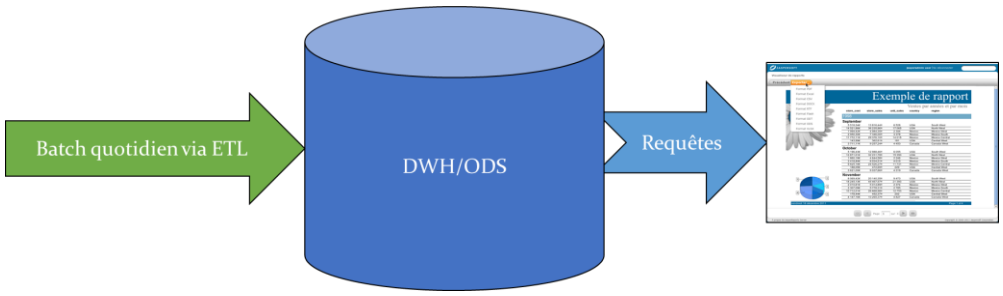

Figure 10: Data management maturité 0

3.2. Batch quotidien pour les sources et le DWH

Pour ce type de management de la donnée, les données proviennent du DWH. De plus, certaines sources peuvent directement s'ajouter au reporting. Les requêtes sont moins importantes mais les temps de réponse sont moyens.

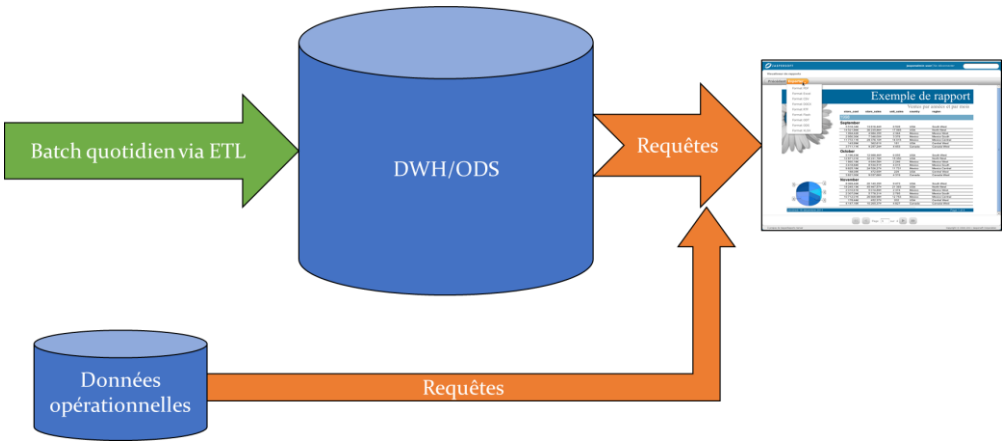

Figure 11: Data management maturité 1

3.3. Batch quotidien en continu pour les sources et DWH

Pour les batchs en continu, on rajoute aux batchs habituels des batchs pour rafraîchir uniquement certaines informations critiques pour les décisions. Les requêtes peuvent être importantes avec des temps de réponse moyens.

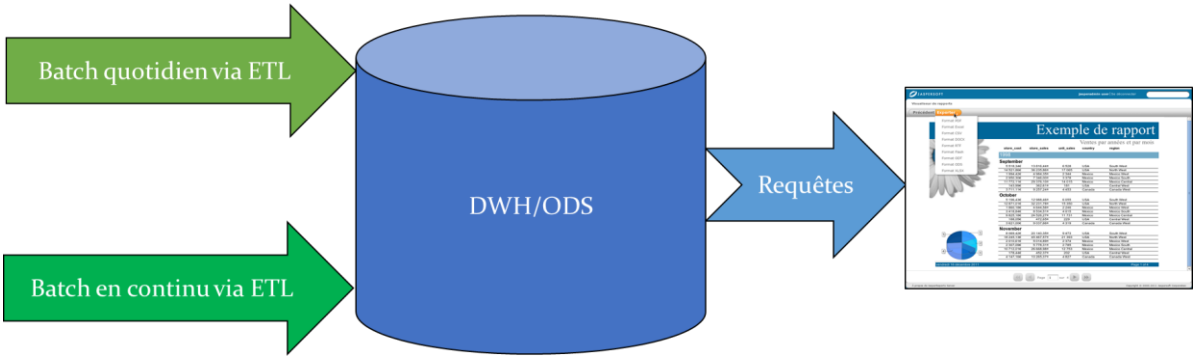

Figure 12: Data management maturité 2

3.4. Batchs commandés par évènements (event driven)

Enfin, les architecture event driven sont proches du temps réel. En plus des batchs classiques, un bus enregistre les évènements des capteurs intégrés dans l'entreprise et en fonction de ces capteurs, les données se mettent à jour. L'avantage est de pouvoir contrôler et réagir en quasi temps réel. Ce type de management de la donnée appartient à ce qu'on appelle BI 2.0 décrite ci-après.

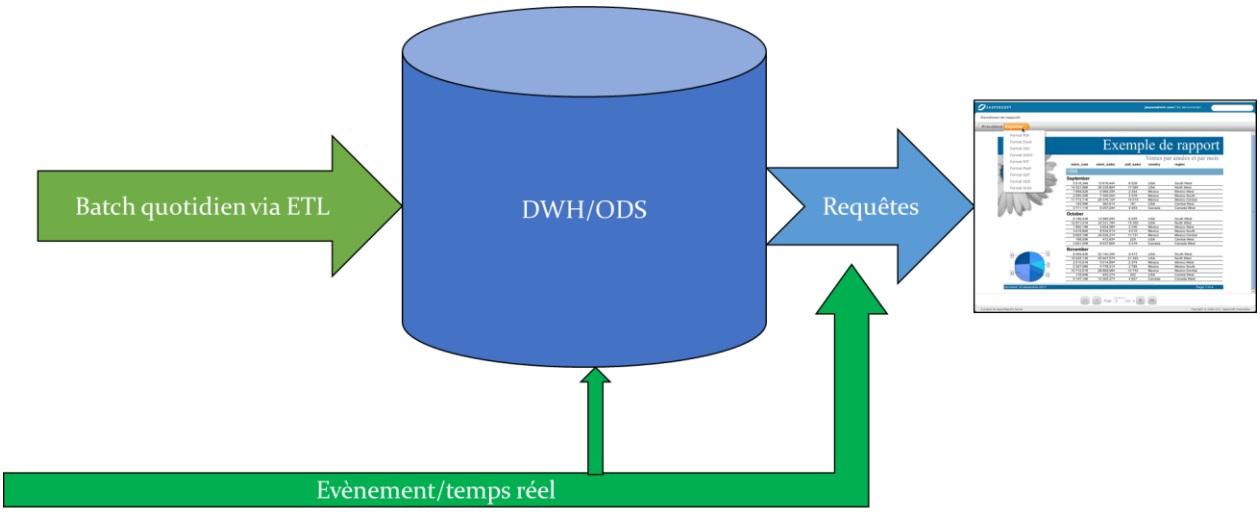

Figure 13: Data management maturité 1

Tableau 2: représentation des niveaux de maturité

Types	Fréquence de mise à jour	Types de requêtes (volume)	Temps de réponse
ETL - Batch sur DWH	Quotidien	Large	Long
ETL – Batch sur DWH et Sources	Quotidien	Moyenne	Moyen
ETL – Batch sur DWH et Sources avec batchs supplémentaires sur certaines sources	Quotidien	large	Moyen
ETL batch et Event driven	Quotidien et temps réel	Large	Rapide

4. La BI 2.0 ou BI de demain

La nouvelle version de la BI appelée aussi BI 2.0 se positionne sur des thèmes d'actualité en Technologie de l'information. Que ce soit les méthodes de conception/management, les technologies, les transformations des habitudes d'utilisation ; la BI 2.0 va venir s'installer petit à petit dans les entreprises. Les bases In-memory révolutionnent le reporting, les méthodes agiles viennent cadrer les réalisations par itération pour délivrer plus rapidement une solution tout en s'approchant au plus près des attentes business. La BI Mobile s'adapte aux habitudes des décideurs qui veulent pouvoir consulter leurs tableaux de bord directement sur leurs smartphones et/ou tablette. Le Cloud Computing qui change les moyens de consommation de la BI dans les entreprises et le Big Data qui offre des possibilités d'analyses sur des volumes de données incomparables avec ceux de la BI actuelle, sous des temps de réponse ultra rapides ; le tout en traitant des données non structurées inexploitées par la BI traditionnelle.

4.1. Bases In-memory

Les BDD traditionnelles ont pour particularité de s'exécuter sur des serveurs dont les supports de stockage physiques sont des disques durs. Les disques durs sont les supports de stockage classiques pour les ordinateurs et serveurs. Les machines informatiques comme les PC sont tous composés de la même manière. Le processeur est le cerveau de la machine, la mémoire vive ou RAM est un stockage volatil ultra rapide sur lequel, la plupart des applications s'exécutent ; de plus, aucune donnée n'y est stockée. Enfin, le disque dur permet de stocker les données dont on a besoin sur le long terme (programmes installés, fichiers, etc.).

Les stockages flash

Depuis quelques années, les supports de stockage Flash ont fait leurs apparitions, il s'agit de mémoires similaires à la mémoire vive mais avec de plus grandes capacités de stockage (sur un PC ordinaire on trouve 4 à 8 Go de RAM). Les disques dit SSD (Solide State Drive ou Disque à Etat Solide) représentent ces nouveaux supports de stockage flash. Ils sont aussi rapide en lecture et écriture que la RAM et peuvent stocker jusqu'à 1 To de données (le plus souvent 250Go).

Puissance du flash

Le prix de ces nouveaux disques a permis de booster les capacités des serveurs et PC. De plus, la particularité des bases In-memory est qu'elles n'utilisent pas seulement ces SSD mais elles disposent également de beaucoup plus de mémoire vive. Le principe est donc de disposer d'une BDD In-memory en sortie du DWH, de charger les données que l'on souhaite analyser ou visualiser directement dans la RAM. De ce fait, les analyses sont instantanées, la visualisation est dynamique et agréable. Le temps de latence noté lors des phases de reporting dû au temps d'acquisition des données dans le cube ou le DTM est inexistant.

Exemple de revendeur

À titre d'exemple, les plus grands fournisseurs de BDD ont tous leur propre base In-memory : SAP HANA, Oracle Exalytics ou IBM BLU Accelerator. De plus, QlikView de Qliktech propose ce service sur de plus petites infrastructures.

4.2. Approche agile

Limites des approches classiques

La plupart des projets BI se font selon une méthode de gestion de projet classique en cascade.

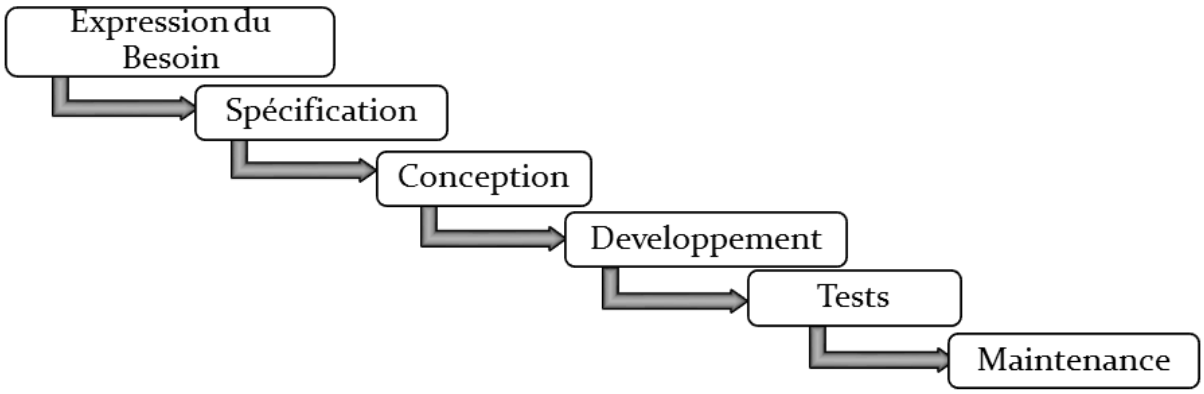

Figure 14: Cycle projet en cascade

Les projets BI peuvent rencontrer des problèmes aux niveaux des sources, de la juridiction sur le croisement des informations. De plus, les besoins métiers peuvent évoluer. Le modèle de données doit donc être évolutif. Le schéma en étoile se doit de pouvoir être modifié sans impacter le reste de l'architecture. Le problème du cycle en cascade est l'effet tunnel ; c'est-à-dire qu'entre le moment où le cahier des charges est établi et celui de la livraison, les besoins peuvent changer, les restitutions peuvent ne pas donner le résultat attendu. Autant de raisons qui engendreront une modification du modèle de données et donc risque de provoquer un allongement du projet associé à des coûts supplémentaires.

Valeurs des méthodes agiles

Une méthode agile se définit en étant itérative et incrémentale, menée dans un esprit collaboratif en mettant l'accent sur la livraison d'un produit de haute qualité en tenant compte de l'évolution des besoins client.

Elle se base sur quatre valeurs :

- Les individus et leurs interactions plus que les processus et les outils
- Des logiciels opérationnels plus qu'une documentation exhaustive
- La collaboration avec les clients plus que la négociation contractuelle
- L'adaptation au changement plus que le suivi d'un plan

Il est ainsi possible de développer le système décisionnel en délivrant rapidement des solutions et le faisant évoluer en fonction des contraintes rencontrées ou besoins du client. Les méthodes les plus connues sont Scrum, XP, et vis-à-vis de la BI, la méthode agile intelligence.

Scrum

Scrum propose une approche par itérations. Une itération est un sprint, il s'agit d'une période fixe pendant laquelle l'équipe devra concevoir un produit répondant aux spécifications définies pour le sprint. Un sprint ne dépasse généralement pas quatre semaines et est encadré par des réunions de début de sprint, quotidiennes, fin de sprint et de retour d'expérience.

L'organisation du projet est divisée en trois rôles. Le Product Owner est responsable du produit, il définit et priorise les tâches du Product Backlog. Le Scrum Master est responsable de l'application de la méthode au sein du projet. Le reste de l'équipe est responsable du développement des tâches du Product et Sprint backlog.

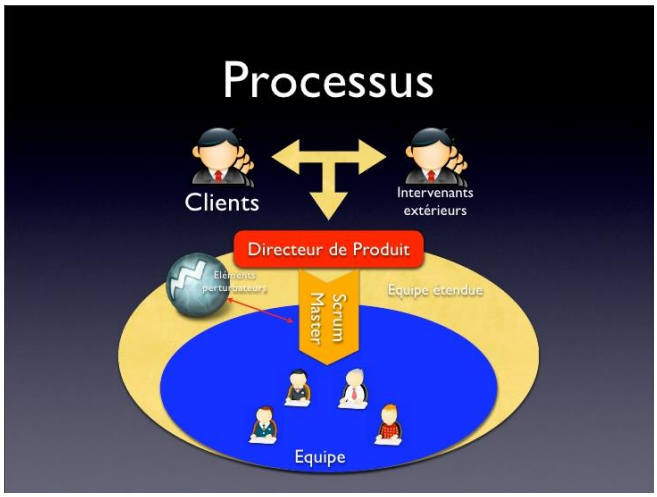

Figure 15: Méthode Scrum (Perriault, 2008)

Un dossier dit Product backlog contient l'ensemble des spécifications à réaliser pour le projet avec une priorisation des tâches. Des éléments peuvent être retirés ou ajoutés à ce dossier. À celui-ci s'ajoute le sprint backlog qui ne contient que certaines tâches spécifiques au sprint en cours. Ces tâches sont détaillées et évaluées en termes de charge. Si elles ne sont pas toutes réalisées, elles pourront être incluses dans le sprint suivant. À la fin de chaque itération, une version du produit est livrée et le client à la possibilité de modifier ses spécifications pour ajuster la version finale.

XP

L'objectif principal de la méthode XP ou EXtreme Programming est de réduire les coûts du changement. Pour cela elle se décline selon plusieurs principes :

- Une revue du code permanent
- Développement piloté par les tests
- Conception au fur et à mesure du projet (refactoring)
- Faire au plus simple
- Privilégier les métaphores
- Intégrer les modifications quotidiennement
- Cycle de développement court

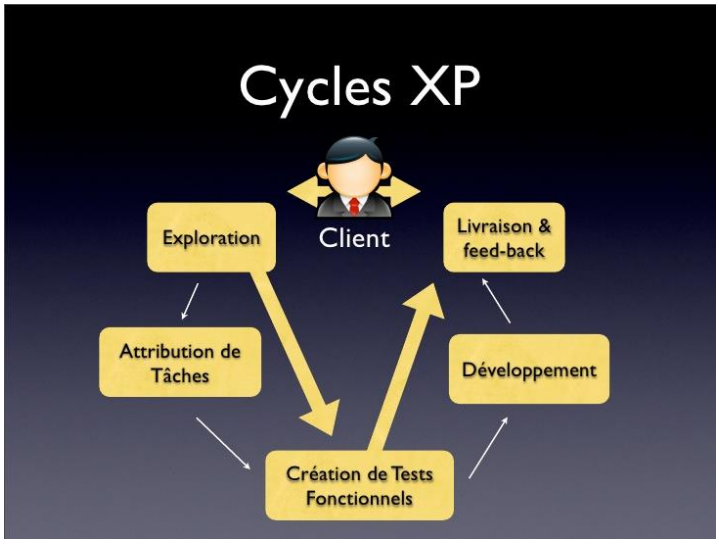

Figure 16: Méthode Agile XP (Perriault, 2008)

Agile intelligence

Agile Intelligence reprend les concepts des méthodes comme Scrum et XP. Les bases des méthodes se fondent sur une approche itérative et incrémentale, avec des petits cycles de feedback et une organisation pratique. Les principes de la méthode sont de créer un pilotage risque/valeur, miser beaucoup sur les retours d'expérience entre les itérations et d'avoir une vision large du projet (penser à tous les départements où la BI s'applique) tout en appliquant les développements sur le département actuel (HARTANI, 2012).

En BI, l'application des méthodes agiles permet de livrer des rapports ou tableaux de bord assez rapidement en concevant un modèle de données évolutif. Une nouvelle méthode est apparue depuis quelque année : le Lean Startup. Cette méthode dite de l'innovation continue peut s'apparenter à un guide de développement de produits/services mêlant le Lean management et les méthodes agiles. La méthode consiste à produire un service en fonction des besoins utilisateurs. Chaque itération vise à développer un produit/service, le faire tester par des utilisateurs potentiellement acheteurs et de le faire évoluer en fonction de leurs retours. L'approche peut facilement s'adapter à la BI.

4.3. Mobile

La mobilité est réellement un changement dans la manière de piloter l'activité. Les smartphones et tablettes sont devenus les outils indispensables du quotidien des décideurs comme des opérationnels. L'intérêt de la mobilité dans la BI est que peu importe le niveau de décision, les personnes responsables n'ont pas nécessairement besoin d'être à leur poste, devant leur PC pour prendre des décisions sur le pilotage. La mobilité ou BI mobile apporte les rapports et tableaux de bord directement sur les téléphones et tablettes.

Les rapports et tableaux de bord des décideurs sont généralement conçus pour qu'ils aient un état des lieux de l'activité en un coup d'œil ; ils ne passent que très peu de temps dessus chaque jour sauf dans le cas où il y a une chute visible de l'activité.

La BI mobile permet donc de répondre aux besoins des décideurs en proposant plusieurs outils d'aide à la décision adaptés aux supports mobiles en fonction de l'appareil (mobile, tablette), plateforme (iOs d'Apple, Android de Google ou Windows de Microsoft), taille d'écran (3.7" à 10").

Les principaux revendeurs ont leurs outils de reporting mobile : SAP, SAS, IBM. Néanmoins, des éditeurs spécialisés comme Roambi ou QlikView mobile proposent des services de BI mobile.

4.4. Big Data et Cloud Computing

Derniers grands sujets : le Cloud Computing et le Big Data. Le Cloud Computing décentralise la BI en entreprise. Le Big Data permet de traiter et analyser de plus gros volumes de données à grande vitesse sur des formats de données non pris en charge par la BI actuelle. Ces deux sujets sont traités en détail dans la partie qui suit.

LE BIG DATA

La gestion d'un volume de données de plus en plus important est un challenge mis en avant dès le début du 21ème siècle (Laney, Doug; Gartner, 2001). Les transformations des business modèles, la mondialisation, la venue des services personnalisés pour les clients ainsi que de nouvelles sources de données ont provoqué cette croissance exponentielle du volume de données générées à chaque instant.

Les entreprises ont changé de business au cours de la dernière décennie. Auparavant, la réputation et le succès des compagnies étaient évalués sur un produit vendu de qualité ayant une utilité. Désormais, les entreprises se doivent de fournir un service plus qu'un produit; les clients souhaitent pouvoir participer et faire évoluer le produit en fonction de leurs besoins. De ce fait, les compagnies sont passées d'une vision orientée produit à une vision orientée service. Ce changement a été accéléré par la mondialisation et la connectivité des clients. Les clients peuvent s'exprimer à tout moment depuis différentes plateformes. Récupérer les informations qu'un client dépose sur un produit, un service ou une entreprise permet de réagir plus rapidement, mais le volume de données engendrées explose. En effet, récupérer tout ce qui se dit sur différents canaux par des millions d'utilisateurs impliquerait de stocker des volumes énormes de données.

La mondialisation joue également un rôle dans cette croissance. Avec Internet, les téléphones mobiles, PC, tablettes et d'autres technologies, le monde entier est connecté et échange des informations sur le web. De plus, le commerce mondial a évolué, les échanges et transactions se font virtuellement. Les données de types variés viennent s'accumuler à échelle mondiale.

Un client satisfait est un client qui promouvra plus facilement un produit ou service d'une entreprise. La satisfaction client devient un enjeu crucial dans la fidélisation et le gain de nouveaux clients. Le client "2.0" souhaite se sentir comme unique; recevoir des offres adaptées et personnalisées lui permet d'avoir ce sentiment d'exclusivité. Cependant, pour pouvoir offrir un tel niveau de service, il est nécessaire de très bien connaitre son client. La connaissance client devient donc une nécessité qui se traduit par l'acquisition de données sur ses goûts, préférences, envies, etc. La récupération de ces informations est disponible via les commandes, mais surtout par la navigation sur le web. Chaque site web peut enregistrer les déplacements de tous les clients sur leurs sites. Ces fichiers, les emails, les demandes dans les call-centers représentent une grande quantité d'information permettant d'améliorer l'expérience client. De la même manière, pouvoir analyser les comportements de navigation permet d'adapter l'ergonomie et le design d'un site pour qu'un client se sente le plus à l'aise possible.

Les données pouvant être récupérées sont nombreuses, disponibles sous différents formats sont actuellement non exploitées par les systèmes décisionnels. Le fait est que ces données existent depuis le début du web, l'arrivée des réseaux sociaux et du web 2.0 ont accéléré les échanges d'informations. Grâce au Big Data, il est désormais possible d'acquérir, d'analyser et de gagner de la valeur à partir de ces données. Le fait est qu'auparavant, les technologies et les prix associés ne permettaient d'obtenir un ROI valant la peine de s'y engager. Depuis la dernière décennie, les supports de stockage ont évolué, le prix associé à la capacité de stockage, c'est-à-dire la quantité d'informations disponible par disque, a grandement baissé. Les supports de stockage sont devenus plus performants tant sur la capacité que sur la

vitesse; par exemple, les disques durs peuvent supporter plusieurs Téraoctets de données et les nouvelles technologies flash (Solide State Drive) offrent des vitesses de traitement cent fois plus rapides. À ceux-ci s'ajoutent de nouveaux types de bases de données dis non relationnelles par opposition aux systèmes décisionnels actuels utilisant des BDD relationnelles et traitant des données hautement structurées. Ces nouvelles BDD ont des performances nettement meilleures sur de gros volumes de données ayant des formats variés. Ces technologies couplées au support de stockage très performant font que le Big Data est accessible pour les entreprises. Le Big Data se base sur des concepts, technologies et une nouvelle chaine de valeur ; de plus, il peut être appliqué dans tous les secteurs.

1. Concepts

Le Big Data offre des capacités d'analyse et de traitement des flux de données en temps réel. Il repose sur les 3V dont l'enjeu est de créer de la valeur.

1.1. Les 3V

Le Big Data a la capacité d'accéder à de gros volumes de données, avec des temps de traitement très rapides le tout en analysant des données de différents formats. Les caractéristiques du Big Data se trouvent donc dans les 3V : Volume, Vitesse et Variétés. La combinaison de ces caractéristiques permettant d'obtenir des prévisions se définit par le 4e V du Big Data : la Valeur.

Volume

La définition de Volume se traduit par la quantité de données générées de façon continue suivant différents formats. Ces volumes de données peuvent être générés par:

- Les fichiers de log sur les sites web qui correspondent à l'historique des actions effectuées sur un site par tous les utilisateurs. Plus il y a d'actions plus les fichiers sont importants en nombre et taille.
- Les données de nos appareils industriels et personnels. Les données disponibles sont des données sur le fonctionnement, l'utilisation (globale et comportementale) ainsi que des fichiers de log.
- Emails.
- Données externes achetées à des tiers ;
- Contrats tout document concernant les employés, les offres, produits, etc.
- Les données de géolocalisation venant des GPS, téléphones, antennes, etc.
- Médias sociaux le nombre de données générées par individu par minute, sur un produit d'une entreprise par minute, etc.

La quantité de données générées chaque jour est en croissance perpétuelle, les systèmes relationnels classiques ne peuvent supporter un stockage de toutes ces données en conservant des coûts et temps de réponse raisonnables.

Vitesse

Dans les systèmes relationnels traditionnels, la notion de vitesse se fait ressentir sur le temps de traitement des analyses, de la fréquence, de la durée des mises à jour des informations dans les BDD ainsi que sur le temps de restitution. De manière générale, les actions se font

de nuit via des batchs (scripts planifiés s'exécutant sur les serveurs), les temps de traitement pouvant aller de la minute au mois selon les cas. Le V de Vitesse du Big Data approche la notion de flux de données en temps réel ou quasi temps réel, avec des flux pouvant être continus. Les données sont acquises et traitées dans la foulé, offrant des temps de réponse de l'ordre de la seconde ou minute. Par exemple, les données peuvent provenir :

- De capteurs envoyant des informations en continu. Les différents systèmes et équipements possèdent des capteurs envoyant plusieurs milliers d'informations chaque minute. Typiquement, un avion est équipé de plusieurs milliers de capteurs permettant de gérer sa position. En utilisant les données de ses capteurs en temps réel, il est possible de réduire la consommation de carburant, réduire les turbulences, etc.).
- Les réseaux mobiles comme les antennes relais que ce soit pour de la voix ou de la donnée pure (3G/4G) voient circuler un flux énorme d'informations. L'analyse du nombre de personnes connectées à une antenne à un instant t permet d'évaluer la charge que cette antenne supporte, voir si elle n'est pas saturée, obtenir des informations sur les personnes les utilisant (position, utilisation, etc.). Ceci permet d'éviter d'avoir des pertes de signal, gérer le matériel et surtout offrir une qualité de service attendue par les utilisateurs.
- Les médias sociaux comme Twitter reçoivent des milliards d'informations par seconde, pouvoir les traiter instantanément est un défi du Big Data.

En plus des flux en temps réels, les temps de traitement de cette masse de données sont incomparables avec ceux d'un système relationnel classique.

Variétés

Les systèmes relationnels traditionnels n'exploitent que des données structurées. Le Big Data offre la possibilité de traiter deux autres types de données : les données semi-structurées et non structurées.

- Les données structurées sont celles utilisées dans les systèmes décisionnels actuels. Il s'agit des données qualifiées, nettoyées, formatées destinées à être stockées dans les bases de données pour de futures analyses.
- Les données semi-structurées sont des données venant de certaines sources, dont les informations obéissent à une logique. Par exemple, une URL est une information non stockée sous forme de colonne ; cependant, elle peut être utilisée en ne récupérant qu'une partie de l'URL. Par exemple l'URL www.Ecole-de-Grenoble.com/student/dupont se segmente telle que l'élève Dupont appartient au réseau de Grenoble. L'information structurée qui en découle est : Nom = Dupont, activité = étudiant, école = Ecole de Grenoble.
- Les données non structurées se définissent comme non exploitables via les BDD relationnelles classiques. Elles n'ont aucun lien entre elles que ce soit sur le format ou leurs structures. Les principaux formats sont les fichiers texte, PDF, logs de sites web, les images, les vidéos, commentaires de réseaux sociaux, des forums, etc. La structure dépend du format, de sa provenance, de la taille, etc. Pour les données non structurées, il n'y a pas de schéma logique permettant de les traiter avec les BDD

traditionnelles qui elles utilisent uniquement des données hautement structurées leur permettant de requêter des BDD optimisées.

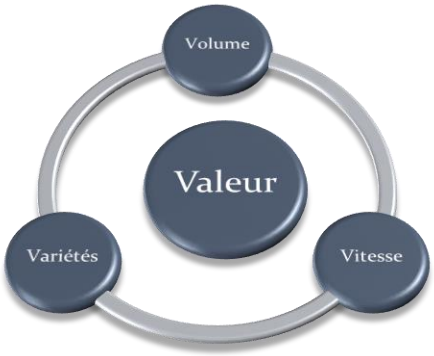

Figure 17: 3V du Big Data

Le dernier V est le résultat des 3V, il peut avoir plusieurs noms en fonction de la situation, le plus fréquent étant Valeur.

1.2. Analytiques

Le plus grand intérêt du Big Data est de pouvoir analyser des évènements ayant eu lieu à l'intérieur ou l'extérieur de l'entreprise. Les plateformes Big Data permettent d'analyser de grandes quantités de données, aux formats variés le tout avec des temps de réponse rapides. Les techniques d'analyse utilisées ne sont pas nouvelles, les analystes utilisent déjà les techniques de Data mining ou fouille de données pour extraire de la valeur des données de l'entreprise. Le Big Data offre la possibilité de traiter plus que ces données d'entreprise qualifiées en allant chercher ces informations soit à l'extérieur (réseaux sociaux, forums, internet) soit à l'intérieur dans des données inexploitables par les BDD relationnelles classiques. En effet, les plateformes peuvent se connecter au Web, aux réseaux sociaux et à tout type de fichier fourni ; ainsi, les analystes ont un plus large champ d'action pour extraire de la valeur ou comprendre les comportements des consommateurs, des architectures techniques, etc. Pour extraire cette valeur, il est nécessaire de passer par un processus de découverte d'information en fouillant dans les données puis de visualisation des jeux de données et enfin s'il est utile la mise en production ou automatisation du processus.

Découverte de l'information

Le processus de découverte d'information s'exécute en quatre étapes :

Figure 18: Processus de découverte de l'information

- Tout d'abord, il faut acquérir les données dont on aura besoin en entrée ; ces données peuvent provenir de différents types de sources (fichiers, BDD, etc.). A cette étape d'acquisition, les données sont rassemblées et préparées pour construire un jeu de données utilisable par les systèmes d'analyses.

- Ensuite les données doivent être taguées ; c'est-à-dire que chacune des données est identifiable.

- Une fois taguée, la donnée est classifiée grâce à un système spécifique par exemple comme le système clé/valeur. En effet, des sous-ensembles du jeu de données principal sont créés et sont associés à une clé.

- Enfin, un modèle est créé, il s'agit d'un jeu de données qualifiant une question qu'une entreprise pourrait se poser pour augmenter sa compétitivité.

Une fois que le jeu de données final est modélisé, la visualisation facilite l'interprétation.

1.2.1. Visualisation

Tout comme pour le reporting en BI, la visualisation de gros volumes de données en Big Data permet d'interpréter les résultats et donc de créer de la valeur. Des logiciels dédiés facilitent cette visualisation. Il s'agit de l'étape la plus importante pour que les utilisateurs adoptent les solutions Big Data. Grâce à la visualisation, les utilisateurs vont pouvoir s'amuser avec les données. La différence par rapport au reporting en BI est que la manipulation des données est beaucoup plus interactive, que le volume de donnée utilisé est plus conséquent et qu'il est possible de créer de la valeur.

Le reporting est surtout utilisé pour présenter des chiffres, faire un état des lieux des données et piloter l'activité. La visualisation va plus loin en permettant de détecter quelles données seraient les plus utiles, lesquelles pourraient être à l'origine d'un comportement, d'une fraude ou d'un besoin.

1.3. Flux de données en temps réel

Le Big Data permet de gérer les flux de données en temps réel. Une machine, par exemple, une antenne réseau, reçoit et génère un volume de données important à chaque seconde. En effet, quand un capteur détecte ou reçoit une information, un évènement est créé et une donnée est générée. Cette donnée peut contenir des informations variées sur cet évènement (date, heure, identifiant du capteur, type de déclenchement, etc.). Sachant qu'à chaque seconde, le nombre d'évènements enregistré est grand, s'il est multiplié par le nombre de capteurs présents, le stockage et l'exploitation de ces informations devient problématique.

Sans le Big Data, la plupart de ces informations ne sont pas utilisées. Grâce aux technologies du Big Data, il est possible de traiter ces informations sans avoir à les stocker. A partir du moment où le modèle analytique est défini, les données peuvent transiter dans le système Big Data et ainsi corriger ou réagir à la situation en quasi-temps réel. La phase de visualisation est également possible en temps réel, cependant, l'intervention humaine est nécessaire.

À titre d'exemple, Air France, sous une architecture IBM a pu optimiser sa consommation d'essence, limiter les turbulences et améliorer l'expérience des voyageurs lors d'un vol long-courrier en installant des serveurs contenant des solutions Big Data. En effet, en temps réel, tous les capteurs présents sur l'appareil ont pu analyser les données de vol et donc optimiser la position et la vitesse de l'avion (Messatfa & IBM, 2013).

1.4. Enjeux

L'intérêt du Big Data réside donc essentiellement sur la possibilité de traiter des données semi ou non structurées sans se préoccuper d'un processus de gestion de données au cas par cas du format ou de la structure des données en entrée. Ces données peuvent être traitées sur de très gros volumes avec des temps de réponse incomparables avec les BDD relationnelles. De plus, disposer d'une architecture Big Data intégrée ou en parallèle des systèmes décisionnels existants offre la possibilité de corréler les informations.

Les technologies du Big Data ont un intérêt quand il est possible d'automatiser le processus de prise de décision. Que ce soit pour des études comportementales, recherches d'informations où l'intervention humaine est nécessaire ou encore analyse de capteurs en temps réel.

2. Technologies impliquées

Les technologies du Big Data doivent permettre de répondre aux 3V et donc de traiter de gros volume de données à grande vitesse le tout en maniant des données de structures différentes. Ces contraintes impliquent que l'architecture devant supporter toutes ces charges soit robuste et puissante.

Le Big Data offre un panel de solutions qui ont vu le jour grâce à des initiatives des géants du Web. En effet, Google, Facebook, Yahoo et d'autres ont lancé des projets open source ; ce sont des projets sous licences gratuites où la communauté peut collaborer dans le but d'améliorer les rendus. L'écosystème Hadoop venant de Apache ou encore les bases NoSQL ont donc répondu aux besoins de ces géants qui se devaient de traiter des péta octets de données pour améliorer l'expérience client. De plus, le Cloud Computing s'étant démocratisé au cours des cinq dernières années, ces technologies sont devenues beaucoup plus abordables.

2.1. Les architectures

La gestion des données se fait selon deux types d'architectures.

Architecture centralisée

La première architecture est dite centralisée. Dans ce cas-là, les données sont collectées puis stockées en un seul et même point (typiquement, il s'agit d'un serveur). Pour ce type d'architecture, la gestion des utilisateurs, de la charge et de l'accès aux données est facilitée. La machine est généralement coûteuse, car elle embarque des capacités de calcul et de stockage importants.

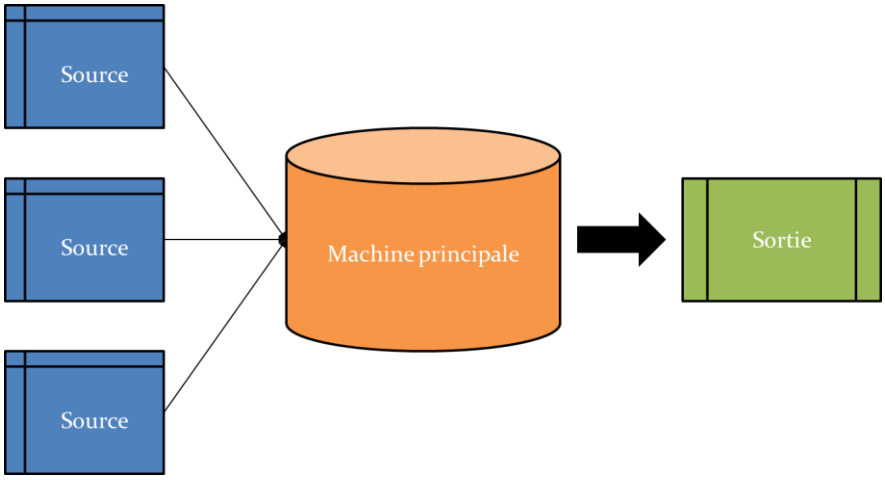

Figure 19: Architecture centralisée

Ce genre d'architecture dispose de capacité d'évolution verticale ; c'est-à-dire qu'étant donné qu'il ne s'agit généralement que d'une machine, cette unique machine sera améliorée sur le plan technique (nombre de processeurs, quantité de RAM, espace de stockage). Ce type d'évolution est limité, car les supports ont une quantité prédéfinie de ports, slots, espaces pour ajouter de nouveaux composants, mais aussi les systèmes ne peuvent pas gérer toutes les technologies.

Architecture distribuée

Le deuxième type d'architecture est représenté par les architectures distribuées, elles permettent de gérer les données en séparant les traitements sur différentes machines. L'avantage ici est de profiter de la puissance de calcul et de la disponibilité de nombreuses machines pour calculer plus rapidement et plus efficacement les traitements.

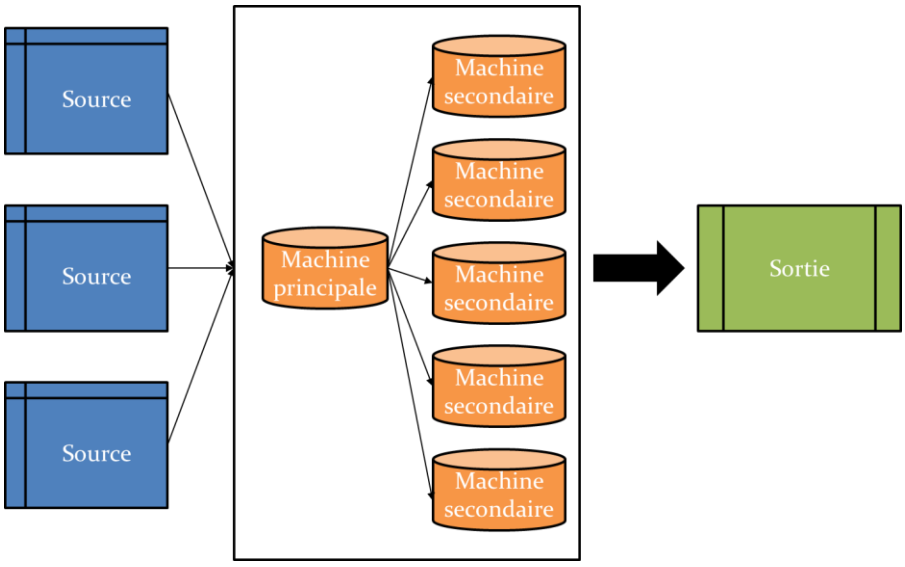

Figure 20: Architecture distribuée

Ce genre d'architecture a la particularité de pouvoir être évoluée horizontalement. En d'autres termes, plutôt que d'améliorer une machine, il suffit de rajouter une nouvelle machine secondaire.

2.2. Le Cloud Computing

Le Cloud Computing est : « *l'accès via un réseau de télécommunications, à la demande et en libre-service, à des ressources informatiques partagées configurables* » (NIST, 2010). En d'autres termes, il s'agit d'un service de location d'infrastructure, de plateforme ou de logiciel informatique que l'on paie en fonction de son exploitation, le tout en passant par internet (Rajkumar Buyya; James Broberg; Andrzej M. Goscinski, 2010).

Quatre points sont essentiels pour définir le Cloud ; un Cloud est un espace virtuel dans lequel les informations sont fragmentées, dupliquées et distribués sur des supports physiques uniques ou multiples possédant une interface permettant de restituer les résultats (CIGREF, 2012).

Un système informatique se décline en couches ; pour faire fonctionner un service informatique, il faut une couche matérielle, logicielle et applicative. En fonction des composants de ces couches, le Cloud offre plusieurs types de services.

De plus, quatre typologies du Cloud permettent de différentier le niveau de confidentialité avec lequel une entreprise peut travailler.

Les types de Cloud

Le service Cloud dépend des couches qu'il fournit. Dans un schéma classique, les couches suivantes sont toutes les couches dont l'entreprise doit s'occuper pour faire fonctionner son activité informatique :

Figure 21: Couches informatiques

Le Cloud permet de fournir des services en fonction de ces couches, le niveau de service s'adapte en fonction du besoin du client. En effet, un client peut n'avoir besoin que d'une architecture technique (serveurs), mais devra se charger de la partie réseau et logiciel ou au contraire vouloir un logiciel complet avec les parties architectures, réseau et logiciel.

Les trois principaux services Cloud sont l'IAAS, PAAS et SAAS ; d'autres types de services existent, notamment le DAAS.

- IAAS, Infrastructure As A Service, est un Cloud d'infrastructure correspondant à l'ensemble des couches matérielles et logicielles. Ces infrastructures virtuelles fournissent un ensemble de solutions de niveau bas par exemple des serveurs, des réseaux, etc. Ils permettent d'aider les entreprises à disposer de machines performantes sans avoir à investir de lourdes sommes. Les entreprises recourant à l'IAAS peuvent installer n'importe quelles logiciels et applications sur leur infrastructure virtuelle.

- PAAS, Platform As A Service, consiste à fournir la couche d'infrastructure d'un IAAS en proposant en plus une couche moyenne permettant aux développeurs d'applications de disposer de solutions pour exploiter les langages dont ils ont besoin. Par exemple pour un développeur Java, un PAAS lui mettra virtuellement à disposition les machines faisant tourner un environnement de développement tel que Eclipse pour produire des applications utilisables en interne.

- Le SAAS, Software As A Service, reprend l'ensemble des couches PAAS plus une couche service. Un catalogue de service met à disposition des utilisateurs des solutions clés en main prêtes à être utilisées (CIGREF, 2012).

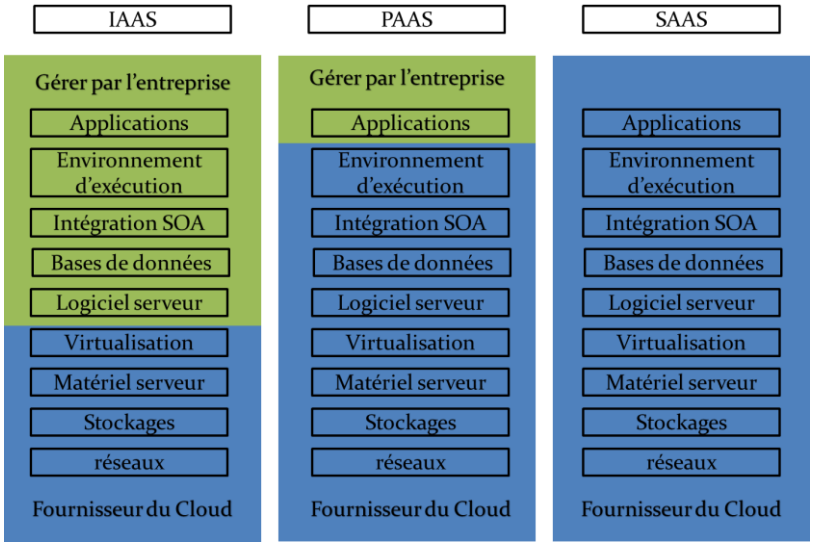

Figure 22: Les types de Cloud

Enfin, DAAS, Data As A Service, est une stratégie Cloud rendant disponible aux métiers des données critiques avec des temps de réponses rapides en assurant la disponibilité et la confidentialité. Ces données critiques sont ainsi stockées en un seul et même endroit ; cela évite la redondance et la multiplication des données au sein de l'entreprise.

Les typologies du Cloud

Quatre typologies de Cloud permettent de définir si le Cloud utilisé est utilisé uniquement pour l'entreprise ou s'il est ouvert ; mais également, s'il est géré en interne ou via une société externe.

- INTERNE/PRIVE L'entreprise gère un Cloud pour son usage interne. Généralement, le Cloud est entièrement géré par la DSI. Grâce à ce Cloud, les employés ont accès à des services fournis par l'entreprise. Il est possible d'adapter le service en fonction de l'utilisateur en paramétrant simplement la couche demandée. De plus, le Cloud interne/privé offre une plus grande maitrise de la qualité de service et des informations qui circulent à l'intérieur ; la sécurité y est plus développée.

- EXTERNE/PRIVE Le Cloud reste à usage interne ; cependant, la gestion est faite par une entreprise externe. Cette solution permet à la DSI de ne pas investir dans des architectures coûteuses. Le cas de l'externe privé est intéressant pour des solutions de location de services à court terme. L'aspect sécurité des informations doit être négocié avec l'entreprise externe. En effet, les données sont la propriété de l'entreprise louant le service, mais elles sont hébergées à l'extérieur. La question des SLA (Service Level Agreement) est très importante, car elle permettra de définir la qualité de service entre le fournisseur et l'entreprise et donc de la satisfaction des utilisateurs en interne.

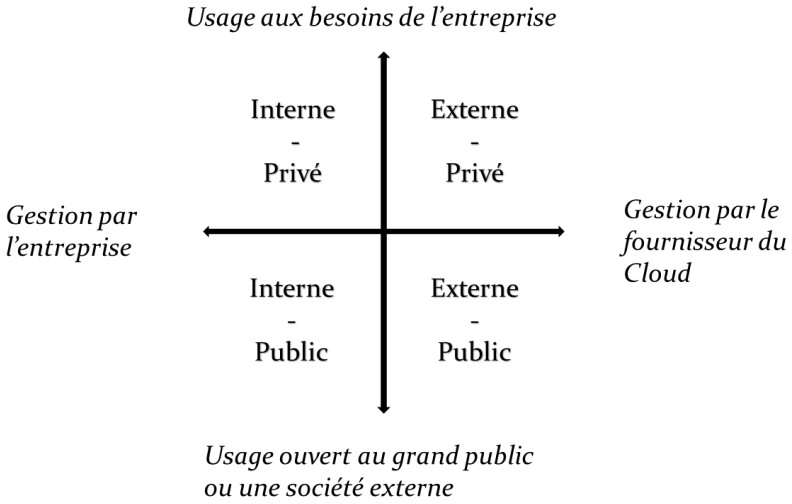

Figure 23: Les typologies du Cloud (CIGREF, 2012)

- INTERNE/PUBLIC cette typologie concerne les opérateurs de Cloud. En effet, le Cloud est géré par l'entreprise, mais les informations circulent vers l'intérieur et l'extérieur.

- EXTERNE/PUBLIC les directions métiers peuvent avoir à passer directement par un fournisseur de Cloud ; c'est le cas de cette typologie. Le Cloud est géré par un fournisseur qui offre un service à des clients.

2.2.1. Intérêt du Cloud Computing pour le Big Data

Le Cloud joue un rôle clé dans le Big Data d'une part parce que le Cloud fonctionne grâce à des data centers (entrepôts de serveurs contenant les données et applications louées aux clients) et d'autre part parce qu'il donne accès à des solutions de Big Data sous forme de service à la demande.

Le Big Data utilise des BDD distribuées dans lesquelles les traitements sont parallélisés ; or les data centers peuvent être utilisés pour paralléliser les traitements. Un data center est par définition un ensemble de serveurs connectés entre eux ; il est ainsi facile d'utiliser toutes ces infrastructures pour exploiter les principes des technologies Big Data.

Une infrastructure, une plateforme ou un service Big Data est coûteux pour une entreprise. Une entreprise a intérêt à investir dans sa propre solution Big Data si elle sait qu'elle a un réel gain pour amortir l'investissement. Cependant, les gains liés au Big Data ne sont pas les mêmes pour toutes les entreprises et il est dur d'avoir de la visibilité. Obtenir un service Big Data via le Cloud permet à une entreprise voulant tester la valeur du Big Data de le faire à moindre cout et en pouvant exploiter les data center d'une autre entreprise (ou de la sienne pour un interne privé).

2.3. Les systèmes de fichiers distribués

Les systèmes de fichiers distribués (File System) sont la base de certaines technologies Big Data. Pour bien les interpréter, il est nécessaire de définir les termes et concepts qui gravitent autour, puis de retracer l'évolution de ces technologies.

Définitions

Les architectures reposent sur un modèle Clients/serveurs. Dans sa configuration la plus basique, un client envoie des demandes à un serveur et le serveur lui répond. Un grand nombre d'applications et notamment les applications web utilisent cette configuration. Le client est généralement une machine simple classique comme un poste utilisateur et le serveur est une machine très puissante capable de fournir des services gourmands en ressources (CPU, RAM, etc.).

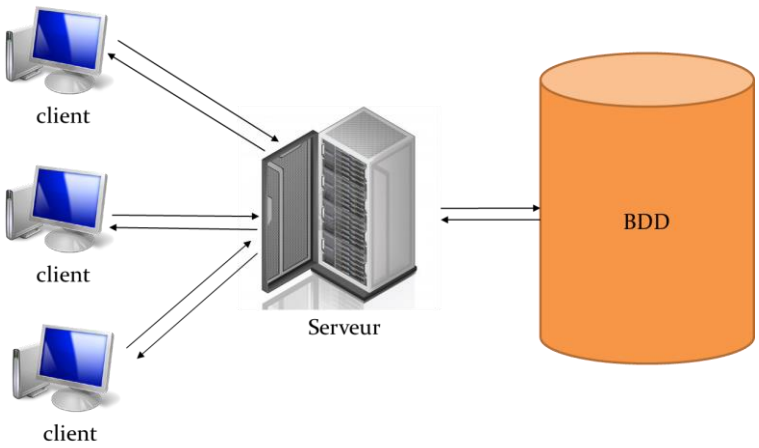

Figure 24: Mode Clients/Serveur

Pour les architectures distribuées, un ou plusieurs clients peuvent envoyer en simultané, des demandes à plusieurs serveurs. Les clients ont tous le même rôle, mais peuvent avoir des interfaces différentes (par exemple, un client java pour demander d'exécuter du code java sur le serveur avec en parallèle un client web qui souhaite récupérer une page web).

Pour gérer les demandes, les serveurs se répartissent selon schéma un maitre (master)/esclave (slave). Le nœud maitre connait le réseau liant les esclaves et sait qui fait quoi et pendant combien de temps. Le serveur esclave est celui qui s'occupe d'effectuer les tâches demandées par les clients, les clients demandent directement aux esclaves et le maitre s'occupe de réguler les tâches.

Pour les architectures Big Data :

- Les serveurs comportent des nœuds, d'où le nœud maitre et le nœud esclave
- Un rack est un ensemble de nœuds
- Un Cluster est un ensemble de racks

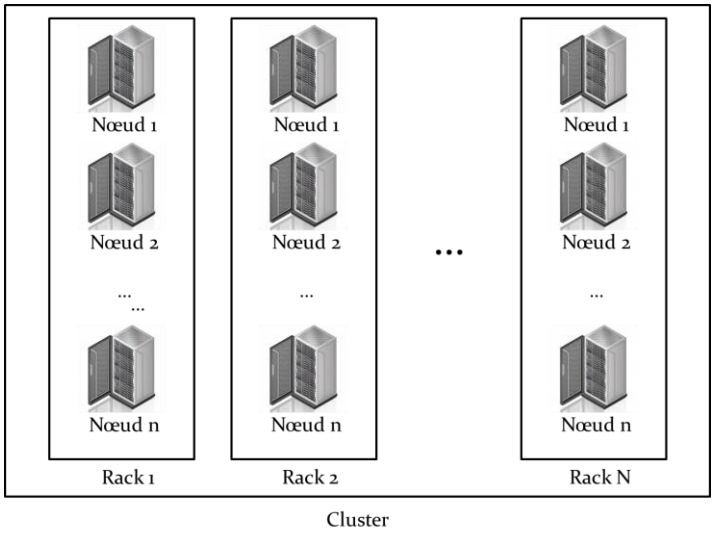

Figure 25: Architecture distribuée classique

L'application du Big Data se fait grâce à ces clusters ainsi que d'algorithmes de traitement de l'information.

Le Google File System

Dans la fin des années 1990, Google cherchait à gérer des volumes de données massives au sein de son moteur de recherche. Cependant, les BDD traditionnelles ne pouvaient simplement pas suivre ses besoins. En 2001, Google a donc trouvé la solution et publié un livre blanc « Google File System GFS » (Google; Sanjay Ghemawat; Howard Gobioff; Shun-Tak Leung, 2001).

Le GFS contient un ou plusieurs clusters ayant chacun un nœud maitre et de multiples nœuds esclaves. Il peut être adressé par des machines externes appelées clients. Le jeu de données principal est divisé en morceaux de tailles fixes répartis dans les nœuds esclaves pour être traités indépendamment (phase de splitting). Le client peut « discuter » avec les esclaves sans à avoir à passer par le nœud maitre.

Le nœud maitre définit où sont placés les morceaux de données, il connait donc tout la carte des répartitions des données dans les clusters. Il est alors dit responsable du mapping. De plus, il connait « l'état de santé » de chacun des serveurs grâce à des systèmes de contrôle réguliers (handshake et checksum). Chaque jeu de données est répliqué sur plusieurs autres nœuds pour récupérer les informations en cas de panne ou d'interruption d'un des nœuds.

Dans cette architecture, le nœud maitre est le SPOF (Single Point Of Failure), il s'agit du point critique d'un système où une panne pourrait entrainer une défaillance totale du système. Pour éviter d'atteindre un tel niveau de criticité, le nœud maitre enregistre dans sa mémoire ROM (mémoire vive non volatil) l'état de la dernière configuration du système (Map, état de santé, etc.) Ainsi, lors d'une panne, il est capable de redémarrer « instantanément » avec les dernières configurations pour minimiser l'impact.

Les technologies

GFS est la référence qui a permis de passer aux plateformes Big Data. En complément de GFS, Google a également développé un algorithme de traitement de fichier spécialisé pour une architecture de fichiers distribué, le paradigme Map/Reduce.

Actuellement, deux types de plateformes sont disponibles offrant chacune ses avantages et inconvénients : Hadoop et les bases NOSQL.

2.4. HADOOP

Hadoop est né d'un projet open source nommé Nutch lancé en 2002 par Mike Cafaella et Doug Cutting dont le but était de créer un moteur de recherche. Nutch s'est heurté aux mêmes problèmes que Google sur la problématique des évolutions de l'architecture pour supporter les volumes de données. Au même moment, Google a partagé avec la communauté sa solution GFS ainsi que son paradigme Map/Reduce. Grâce à cela, Nutch a évolué vers le projet Nutch Distributed File Système NDFS. En 2005, NDFS avait résolu les problèmes d'évolutivités, de stockages et la fonction Map/Reduce a été réadaptée pour satisfaire les besoins du projet. En 2006 le sous-projet Hadoop est apparu, la première version a été utilisée et supportée par Yahoo pour l'améliorer (Apache, 2005).

La plateforme est composée de deux composants indispensables HDFS pour Hadoop Distributed File System qui reprend NDFS et Map/Reduce. L'écosystème Hadoop comprend quant à lui en plus de la plateforme de base, des Frameworks permettant d'adresser des DWH (HIVE), gérer des données massives via un langage particulier (PIG), des transpositions de SQL pour Hadoop (SQOOP), des Frameworks d'administration (ZOOKEEPER) ainsi que la seconde version de Hadoop (YARN).

Lors d'un traitement, l'HDFS partitionne un gros volume de données en blocs de petites tailles égales, chacun de ces blocs est ensuite envoyé sur un nœud du cluster (également répliqué à deux autres endroits). Grâce aux jobs Map/Reduce, les actions de traitement sont effectuées sur chacun des blocs présents sur les nœuds puis le résultat est consolidé.

HDFS

HDFS reprend les mêmes concepts que GFS avec des correctifs ; il s'agit donc d'un ensemble de clusters, chaque cluster possède des serveurs contenant des racks et les racks possèdent plusieurs nœuds. Des systèmes de réplications assurent de ne pas perdre de données pendant

le traitement et des systèmes assurent que lors d'une panne, l'analyse pourra reprendre très rapidement sans impacter les résultats.

HDFS est composé de plusieurs parties :

- Le client HDFS est l'interface Hadoop avec laquelle un utilisateur peut envoyer les jobs Map/Reduce qui permettent de traiter les données au sein du cluster. Il peut interagir avec le NameNode ainsi qu'avec les DataNodes en exécutant les jobs.

- Le NameNode NN est l'équivalent du nœud maitre. Chaque cluster possède un seul NameNode, il est en charge de gérer les blocs au sein du cluster et régule les actions du client HDFS vers les DataNodes.

- Les DataNodes DN correspondent aux nœuds esclaves. Ils s'occupent de l'espace à allouer par bloc ainsi que les données qui vont y transiter. Pour cela, en collaboration avec le NameNode, les DataNodes contrôlent les traitements demandés par le client.

- Le JobTracker JT est unique au sein d'un cluster, il s'occupe de la localisation des informations et de la répartition des traitements dans le cluster.

- Le TaskTracker TT accepte d'exécuter des jobs venant du JobTracker. Il gère le nombre de Jobs qu'il peut exécuter en parallèle.

Lors d'un traitement, un client envoie une demande de job, celle-ci est traitée par le JobTracker qui cherche la localisation des données avec le NN. Une fois la donnée identifiée, le JT cherche un TT disponible pour effectuer le traitement, le TT exécutera alors la tâche sur le DN.

De plus, HDFS possède plusieurs propriétés :

- L'image contient les métadonnées concernant les blocs présents dans le cluster. Elle est stockée en mémoire de façon à pouvoir traiter en simultané les demandes de plusieurs clients différents.

- Le journal est le fichier de log de l'image, il historise tous les traitements effectués entre un client et un DataNode. En cas d'erreur ou de redémarrage, le NameNode se base sur le journal pour savoir où reprendre.

- Le Checkpoint est un fichier de configuration utilisé à chaque redémarrage. Lorsque la machine est démarrée, un fichier de checkpoint est créé et n'est jamais modifié. Il sera remplacé lors du redémarrage suivant.

La communication est essentielle entre le NN et DN pour assurer qu'aucune donnée n'est perdue ou oubliée, qu'un DN fonctionne toujours, que le traitement ne rate pas, etc. Pour assurer la consistance des données, les blocs sont répliqués au sein du cluster. Chaque bloc est copié à deux endroits sur un rack voisin.

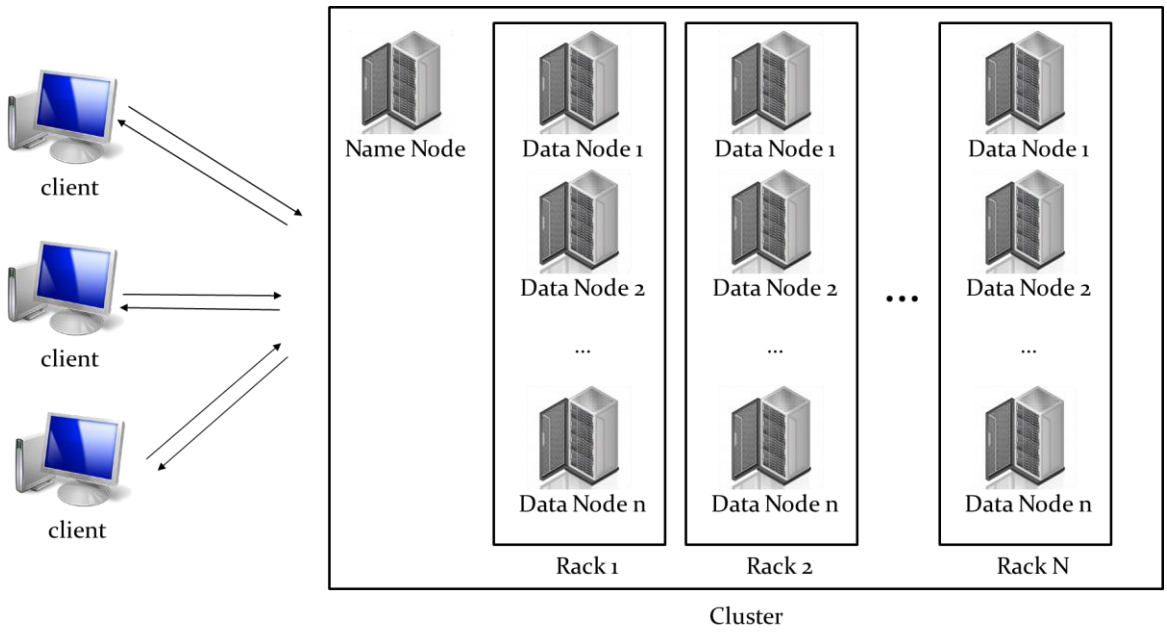

Figure 26: HDFS

Pour s'assurer de l'état de santé du cluster, chaque DN émet toutes les trois secondes, un « battement de cœur ». Il s'agit d'un signal permettant d'avertir le NN que le DN est toujours en marche et ne rencontre aucun problème. Si un DN n'émet plus pendant plus de dix minutes, il est automatiquement retiré du mapping du NN et les blocs traités et répliqués dans ce DN sont envoyés vers d'autres DN. Ce battement de cœur est également effectué entre le TT et le JT.

A un niveau plus haut, des photos (snapshots) sont effectuées. Il s'agit d'une sauvegarde à un instant donné du système HDFS. En cas de mise à jour de la plateforme ou de configuration, des problèmes peuvent survenir et la photo permet de restaurer l'HDFS à son état précédent.

Map/Reduce

Map/Reduce provient d'un modèle de programmation permettant de traiter un gros volume de données, il a été développé par IBM en 1968. L'utilisation de ce modèle est un grand pas en arrière sur l'histoire du traitement de l'information (DeWitt, 2008)Cependant, en 2000, Google en partant de ce modèle a sorti Map/Reduce.

Il est composé de deux fonctions Map et Reduce écrites par l'utilisateur, qui sont exécutées sur des architectures distribuées. La fonction Map projette le code sur l'ensemble des nœuds des clusters, tandis que la fonction Reduce s'exécute pour récupérer ces informations et ainsi produire un résultat (Google; Sanjay Ghemawat; Howard Gobioff; Shun-Tak Leung, 2001).

Pour un jeu de données en entrée, une opération de splitting est effectuée, c'est-à-dire que les données vont être séparées en jeux de données plus petits pour être parallélisées. La fonction Map projette les jeux de données et associe les éléments à une clé. Le shuffle vient rapprocher les éléments communs à des splits différents. Enfin la fonction Reduce consolide les résultats.

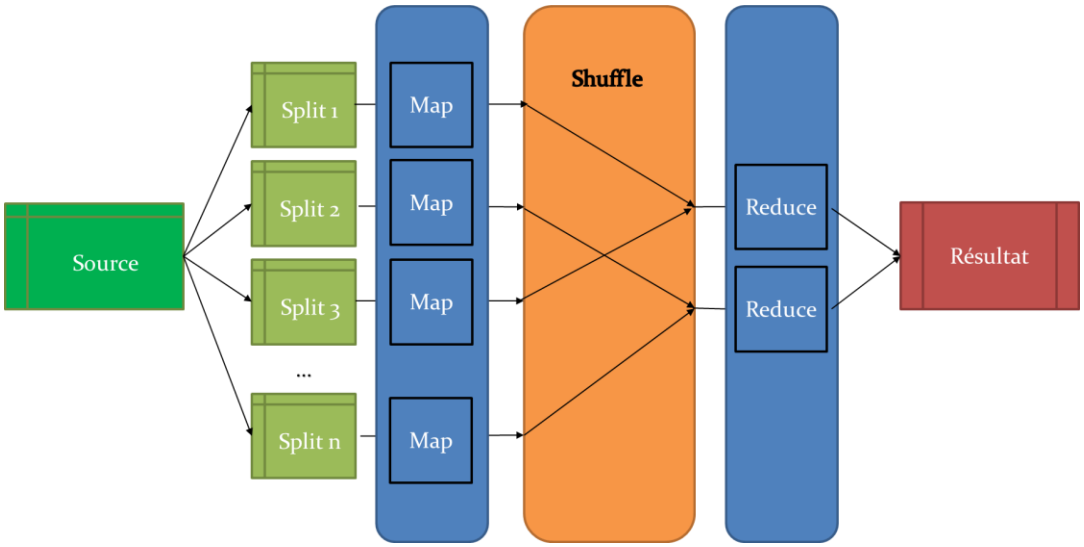

Figure 27: Map/Reduce

Prenons l'exemple ci-dessous qui compte le nombre de mots identique. Notre entrée contient les lignes :

- Pierre fume une cigarette
- Alice joue aux cartes
- Jean conduit sa voiture
- Pierre joue aux cartes
- Alice fume une cigarette

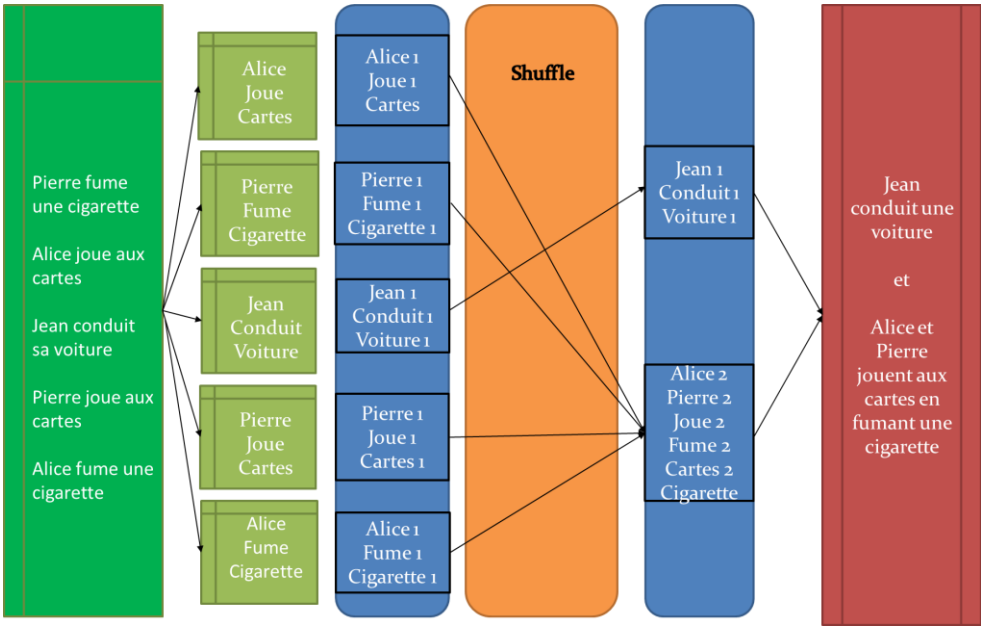

Figure 28: Exemple de job Map/Reduce

Ici, on cherche à déterminer les interactions entre les personnes. La fonction Map va donc extraire le premier mot, le second mot et le quatrième mot. La fonction Reduce va compter l'occurrence de chacun des mots. Au final, il est possible de supposer que Pierre et Alice jouent aux cartes ensemble tout en fumant et que Pierre est seul dans sa voiture.

Map/Reduce s'intègre parfaitement à HDFS :

- le HDFS permet de paralléliser les traitements sur le DN et NN.
- Map/Reduce s'exécute sur les JT et TT.

Deux couches sont ainsi visibles pour effectuer des traitements. La partie HDFS et la partie Map/Reduce.

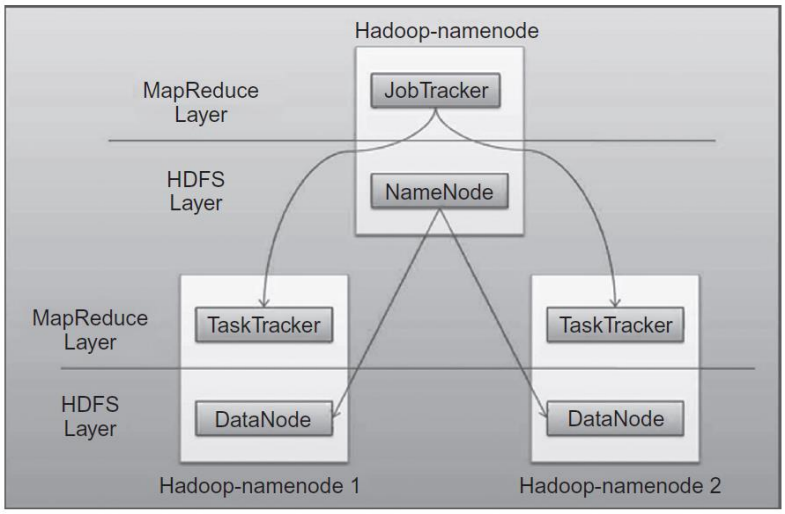

Figure 29: Comportement HDFS-Map/Reduce (Krishnan, 2013)

Écosystème

Un écosystème gravite autour de Hadoop. La nouvelle version YARN apporte de nouvelles fonctionnalités, des Frameworks permettent d'utiliser le SQL, des DWH ou encore de gérer les accès.

- YARN ou map/Reduce V2 se base sur une meilleure gestion des ressources et planification. Initialement effectuées par le JobTracker, la gestion des ressources et la planification sont attribuées à deux nouveaux modules :
  - Le Ressource Manager, RM, s'occupe de l'allocation des ressources en fonction de la planification des tâches.
  - NodeManager qui gère l'application et le bon déroulement du job au sein du nœud.

  En résumé, chaque nœud est découpé en blocs mémoires prêts à recevoir un bloc de fichier. Dans la V1, les blocs ne se différencient pas pour les Jobs de type Map ou Reduce. Au milieu d'un traitement, des Maps peuvent s'exécuter en même temps que des Reduces, ce qui peut provoquer des « embouteillages » ralentissant les traitements.

- HIVE est une solution de DWH pour Hadoop. Il utilise HiveQL pour générer du SQL sous Hadoop et ainsi interagir avec le DWH (Apache, 2013).

- PIG est une plateforme permettant d'analyser un large volume de données ne pouvant être traitées via Map/Reduce. Il génère des séquences Map/Reduce pour diminuer la charge provoquée par un Map/Reduce trop gourmand. Il utilise le langage PIG Latin pour interagir avec la plateforme (Apache, 2013).

- SQOOP permet à Hadoop d'interagir avec un DWH d'entreprise. SQOOP : SQL pour HADOOP (Apache, 2013).

- ZOOKEEPER est une base NoSQL in-memory permettant de coordonner les applications distribuées. YARN se base sur les concepts de ZOOKEEPER pour assurer la gestion des applications. Le rôle principal de ZOOKEEPER est de permettre à l'ensemble des nœuds de connaitre l'état de santé du cluster (Apache, 2013)

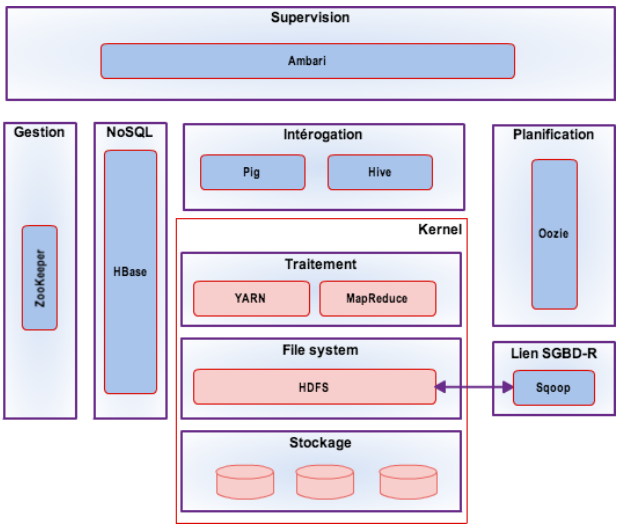

Figure 30: Écosystème Hadoop (Parageaud, 2013)

2.5. Bases NoSQL

Les problèmes des systèmes transactionnels a été de ne pas pouvoir évoluer en fonction des besoins en termes de traitement de larges volumes de données. Ceci est essentiellement dû aux caractéristiques ACID des systèmes relationnels. Ces propriétés font office de résistance quand il s'agit de faire circuler des flux de données en temps réel générées par des capteurs ou des applications web. Un mouvement est donc né à la fin du 20ème siècle, celui du Not Only SQL (NoSQL). Le mouvement a créé une nouvelle classe de base de données qui par opposition à bases relationnelles qui sont structurées, utilise majoritairement le SQL pour extraire des lignes. Le NoSQL est alors apparu comme la réponse aux contraintes de structure et d'évolutivité des BDD relationnelles ; on parle de bases non relationnelles. Ces BDD fonctionnent selon certains concepts, obéissent à un théorème et ont des applications variées.

Concepts

Les BDD NoSQL ont vu le jour tout comme pour l'écosystème Hadoop par la communauté Open Source alimenté par les géants du Web. Plus particulièrement suite à des recherches effectuées par Amazon avec Dynamo une BDD Clé/valeur ainsi que Google Big Table le système de stockage multi dimensionnel de Google utilisant un mapping de la BDD via la position. Par la suite, de nouvelles BDD NoSQL sont sorties des laboratoires de recherche de ces géants du web et de la communauté Open Source.

Ces BDD se classent en quatre grandes familles (Aurélien Foucret; SMILE, 2011) :

- Les bases Clé/valeur dans lesquelles une paire composée d'une clé pointe vers une donnée spécifique que l'on appelle valeur. Ces paires sont indexées dans une table contenant la liste des valeurs associées aux clés appelée table de hashage. La table permet de retourner très rapidement une donnée structurée ou non dans l'ensemble de la BDD.

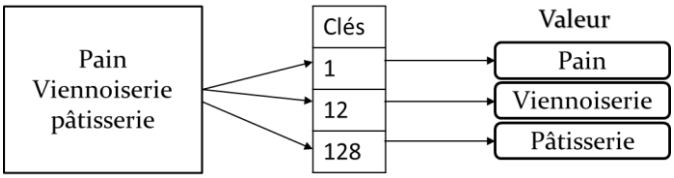

Figure 31: BDD NoSQL clé/Valeur

- Les bases orientées colonnes/familles qui sont une extension des BDD clé/valeur ont la différence que la clé pointe vers une donnée qui est organisée en colonne de même famille.

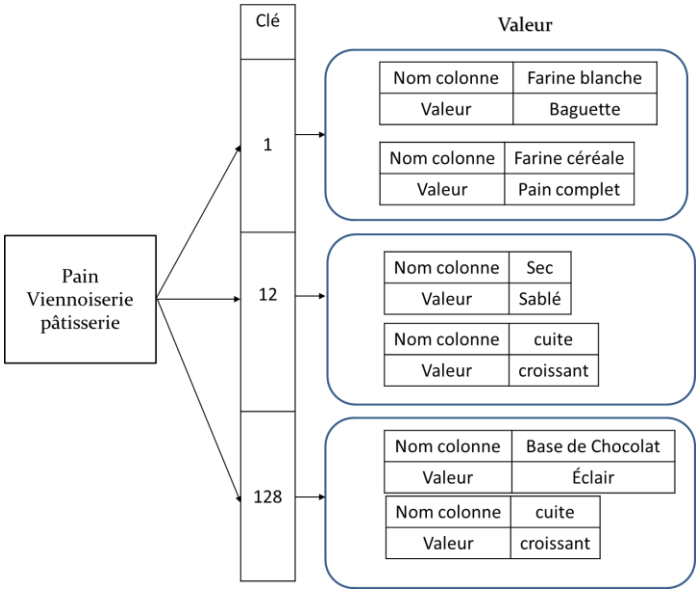

Figure 32: BDD NoSQL orientée colonne/familles

- Les BDD documents ressemblent aux BDD clé/valeur. Cependant, la clé pointe vers un document au format json ou XML. Un document est composé de champs et de valeurs associées. En comparaison aux BDD relationnelles, ces bases sont dites schemaless ; c'est-à-dire, qu'elles n'ont pas besoin d'un modèle de données prédéfini, elles s'adaptent à tout type de contenus et permettent donc d'offrir un compromis performance/agilité. Dans l'exemple ci-dessous, les documents sont des recettes, composées de champs (ingrédients) associé à la valeur (quantité).

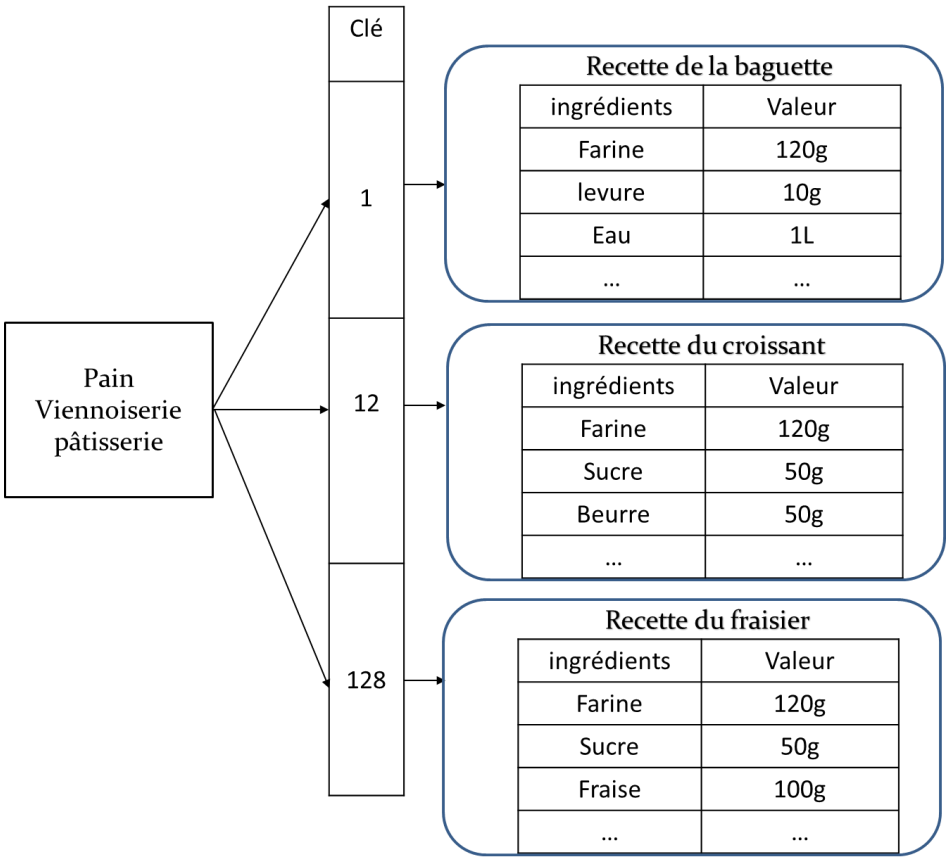

Figure 33: BDD NoSQL Documents

- Les BDD de graphe sont des bases qui supportent la théorie des graphes. Elles sont organisées en clusters tout comme Hadoop et peuvent ainsi traiter des ensembles de données très complexes. Le principe se fonde sur deux parties :
  - d'une part, la partie stockage, elle correspond aux nœuds et se comporte comme une base Documentaire.
  - d'autre part, une partie décrit les arcs entre les nœuds. Les arcs sont orientés et possèdent quelques attributs.

  Dans l'exemple ci-dessous, nous avons Pierre, Alice et Jean sur un réseau professionnel. Pierre était dans la même école qu'Alice et connait Jean depuis l'enfance. Chaque nœud possède les informations de chacune des personnes et les arcs décrivent leurs relations.

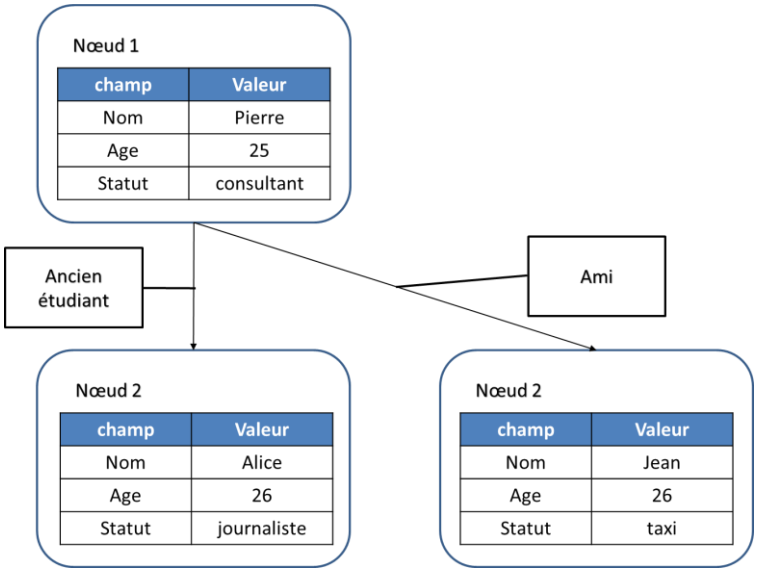

Figure 34: BDD NoSQL Graphe

Peu importe la classe de la BDD, chacune de ces BDD supporte le théorème de CAP.

Théorème de CAP

Le théorème de CAP a été présenté pour la première fois en 2000 lors d'une conférence sur les systèmes distribués. Il stipule que tout environnement distribué doit répondre à au moins deux des trois principes du théorème de CAP ; en d'autres termes, le système doit garantir deux des trois caractéristiques ci-dessous (Strauch, 2011)

- Consistency (consistance) soit que toutes les données disponibles sur chacun des nœuds ne soient pas altérées au cours d'un traitement.
- Availability (disponibilité) chaque requête doit renvoyer une réponse.
- Partition tolerance (tolérance aux partitionnements) le système doit agir sans prendre en considération la disponibilité, la partition ou la perte d'une donnée ou communication.

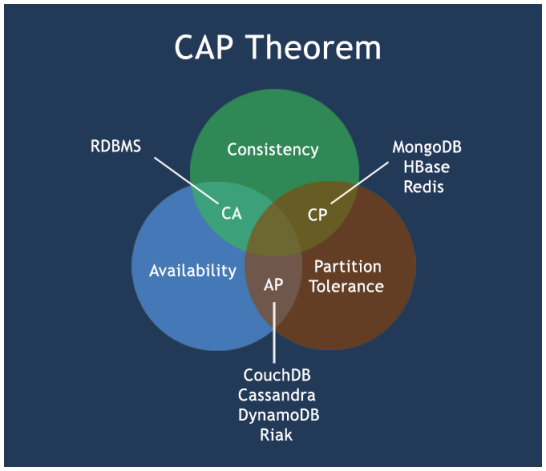

Figure 35: Théorème de CAP (w3resource, 2012)

Les BDD ayant évoluée depuis Google Big Table ou Amazon Dynamo sont typées CP/CA/AP. En fonction des deux fonctions qu'elles remplissent. Pour informations, les BDD relationnelles sont des bases CA.

Exemple de bases NoSQL

Une BDD clé/valeur de type AP simple est celle développée par le réseau social professionnel LinkedIn : Voldemort. LinkedIn avait besoin d'une BDD légère, très évolutive qui pouvait traiter un large volume de données sans avoir à respecter les contraintes ACID ; ils se sont alors inspirés de DynamoDB d'Amazon. Voldemort est une BDD clé/valeur avec la spécificité que la BDD ne travaille pas avec une paire clé/valeur, mais une clé avec une valeur. La différence se fait sur l'organisation des données ; plutôt que d'être indexées dans une table de hashage, la donnée est répliquée sur plusieurs nœuds liés entre eux par un anneau.

Une BDD clé/valeur organisée en colonne/famille est celle de Facebook : Casssandra. Compte tenu du nombre croissant d'utilisateurs Facebook, leur modèle doit reposer sur une BDD sans limites d'évolution orientée sur une disponibilité et une capacité de distribution forte. Le modèle est donc du type AP et combine l'infrastructure de dynamo avec le modèle de donnée de BigTable.

Riak est une BDD document basée sur Cassandra. Ce type d'architecture permet de stocker un grand nombre d'informations en passant par un document. Une clé est associée à ce document permettant ainsi de traiter un nombre important de données non structurées. Par exemple, un document historisant des clics sur page web est stocké sous un fichier XML ou json via la clé X. Si ce fichier doit être modifié (par exemple en rajoutant un champ URL), il suffit de modifier le fichier associé à la clé en y ajoutant le champ désiré. Dans le cas d'une BDD relationnelle, il aurait fallu revoir la structure de la BDD en rajoutant le champ dans la ou les tables nécessaires.

Les bases de données graphe sont les BDD NoSQL les plus complexes et sont nées de l'émergence des réseaux sociaux : Neo4J utilisée par Viadeo ou encore GraphDB. Le modèle pour ces bases est une clé/valeur complexe. Deux types d'objets permettent de comprendre la BDD : les Individus et les liens. Chaque individu est représenté par une donnée et se situe sur un nœud. Il est alors possible de construire un arbre d'appartenance en partant d'un nœud racine et en dépliant les liens.

Les BDD NoSQL permettent de gérer de gros volumes de données en évitant les contraintes ACID des bases relationnelles. Elles sont faites pour évoluer facilement en fonction du besoin de l'entreprise. Ces BDD sont faites pour répondre aux caractéristiques du théorème de CAP et en fonction des besoins de l'entreprise, le choix d'une BDD dépendra de son contexte en sélectionnant des BDD CP ou AP. L'avantage des BDD NoSQL réside dans le fait de pouvoir traiter des données non structurées avec une grande facilité.

3. Chaine de valeur du Big Data

Pour extraire de la valeur du Big Data certaines étapes clés sont nécessaires :

Figure 36: Chaine de Valeur du Big Data (Orange Business Services, 2013)

3.1. Data management du Big Data

Le data Management du Big Data se positionne sur les données et les technologies pour extraire les SMART DATA.

Les données

Pour gérer les données via une architecture Big Data, il faut prendre en compte les contraintes liées à ces données :

Le volume explose, que ce soit les données internes (fichiers Word, PDF, contrats, données clients, call center, etc.) ou externes à l'entreprise (réseaux sociaux, forum, etc.), les volumes à gérer sont très importants.

La variété des données amène les problématiques de structure ; il est compliqué de traiter tous les formats de données en même temps en un même point. Les données structurées venant des systèmes relationnels sont déjà présentes et disponibles. Les données non structurées (Excel, doc, PDF, blog, medias sociaux, vidéos, images, audio, capteurs et données géographiques) représentent le plus gros volume à traiter selon des formats totalement différents.

La complexité de la donnée pose encore des problèmes supplémentaires. Par exemple au sein d'une même application, en fonction des versions utilisées la sortie peut avoir changé. Par exemple, un fichier CSV avec ou sans header, un fichier Word 2003/2007/2010/2013/ODT, données Facebook contre Twitter, etc.

Enfin la vitesse des données ajoute une contrainte supplémentaire selon laquelle ces gros volumes de données contenant des formats aux complexités variées doivent être traités avec des temps de réponse incomparables.

Ce qui mène le management de la donnée, celui-ci est dépendant de deux axes, le premier représente l'incertitude du résultat et le second s'oppose au premier grâce au gain de temps offert par les systèmes non relationnels.

Le schéma classique pour les systèmes relationnels est de Collecter, Standardiser, Découvrir les informations pertinentes, Modéliser le comportement des données et enfin de Structurer le résultat dans un support de stockage pour l'analyse et le reporting.

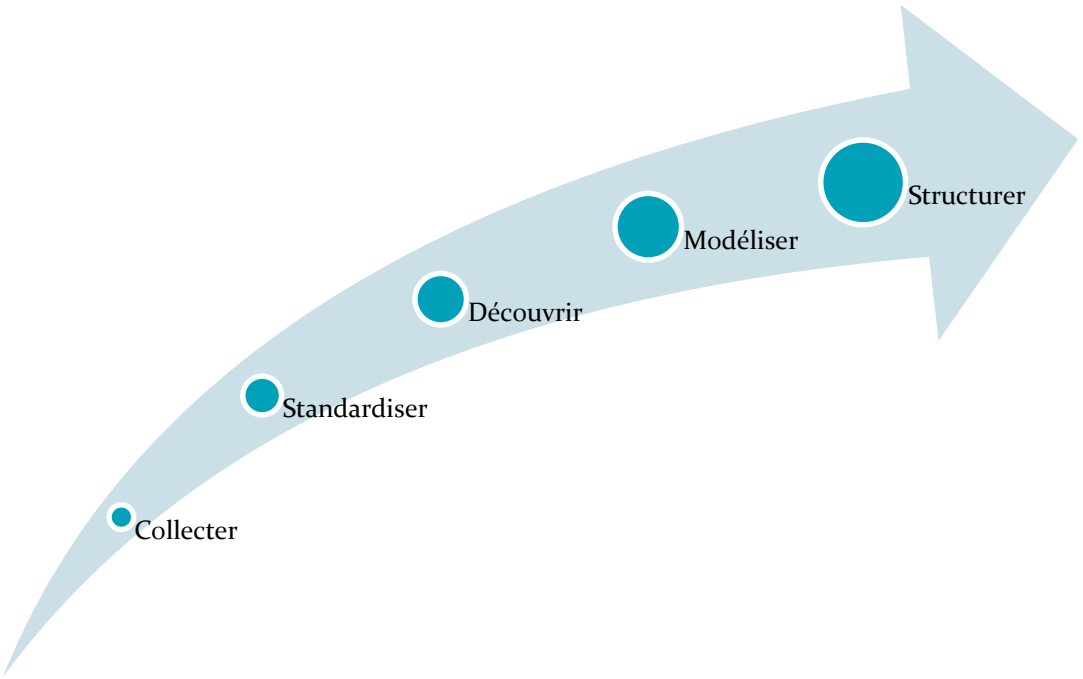

Figure 37: Chaine de traitement des données classique

Pour le Big Data, le processus est différent, il s'agit de Collecter, Structurer, Transformer et Extraire

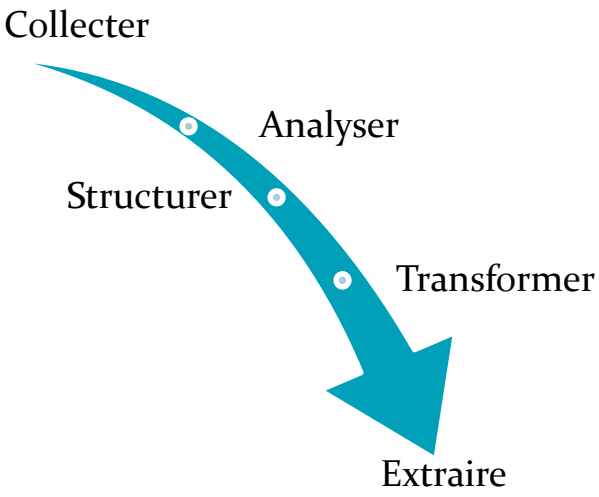

Figure 38: Chaine de traitement des données pour le Big Data

La première étape est donc de récupérer toutes les données dont on dispose peu importe leur format ou leur volume. Il faut que la personne ait le droit d'accéder aux données, mais

également que celles-ci aient un intérêt à être explorées. Ces données multi sources sont ensuite chargées dans un système de fichier de notre architecture de fichiers distribués.

Une fois collecté, il faut analyser en suivant les étapes décrites précédemment (Acquisition, Tag, classification et Modélisation). L'analyse permettra de définir quelle structure apporter au jeu de données.

Ensuite, structure est appliquée aux données via les métadonnées. La métadonnée est une donnée sur la donnée, un libellé associé à une donnée pour connaitre toutes les informations sur cette donnée. Celles-ci va définir la donnée que l'on souhaite traiter et sera l'intermédiaire entre la donnée structurée et la donnée non structurée.

Une transformation est appliquée sur la métadonnée: elle est standardisée et préparée pour de futures analyses.

Enfin les données peuvent être extraites par des systèmes d'analyses ou reporting.

Les technologies

L'interaction avec les technologies comme Hadoop ou les bases NoSQL intervient essentiellement au niveau de la structure et transformation. Lors de la première étape, les données sont collectées puis ordonnées par dossiers de mêmes types qui seront traités indépendamment. Lors de la structure, toutes les métadonnées seront créées en précisant plusieurs informations sur la donnée et surtout en lui assignant une clé. Puis la donnée est transformée. Dans le cas d'une plateforme Hadoop, les clés sont utilisées lors des Job Map/Reduce ; un premier jeu de clés sort du Mapping puis un second lors du Reduce. Dans le cas des bases NoSQL, fonctionnant en Clé/Valeur, les métadonnées sont les valeurs et elles sont reconnues par les clés.

Les SMART DATA

Le réel intérêt du Big Data est de pouvoir extraire de la valeur de tous ces volumes de données. Cependant, seules 20% de ces données sont considérées comme SMART DATA. Les SMART DATA sont ces données qui représentent une valeur pour l'entreprise elles sont découvertes en fin de cycle une fois qu'elles ont été traitées dans la phase de transformation et extraites dans des outils d'analyses ou reporting. Ces données sont considérées comme cruciales et doivent être conservées.

Dans le cas, d'une application aux capteurs, le processus est le même, à la différence que les opérations s'exécutent en continu. Les données sont collectées, transformées, puis traitées et directement utilisées. L'idéal étant de pouvoir croiser les possibilités d'Hadoop et Map/Reduce avec les BDD NoSQL. Ces solutions sont compatibles et il est conseillé de mélanger les approches.

3.2. L'analyse des données du Big Data

Pour le Big Data, l'analyse des données se fait sur des méthodes de Data Mining et de statistiques exploratoires connues. Cependant, les 3V du Big Data augmentent cette complexité. Par rapport à la figure 9 au chapitre sur le Data Mining, le Big Data n'intervient que dans les méthodes descriptives supervisées et non supervisées

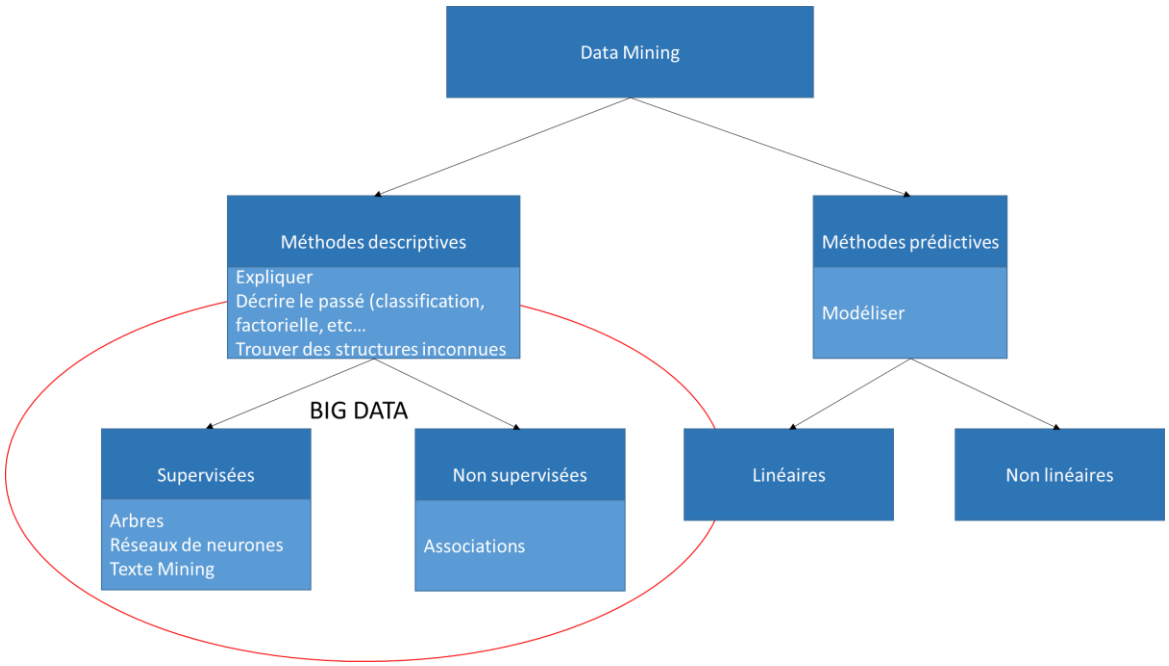

Figure 39: Arbre de décision de l'analyse des données Big Data

En effet, les méthodes prédictives représentent un véritable enjeu pour le Big Data. En Data Mining, les analyses prédictives sont qualifiées par une valeur appelée p-value qui permet d'évaluer de la pertinence d'un modèle et donc de supposer qu'il y a un intérêt stratégique à prendre une décision en se basant sur le modèle. Pour le Big Data, il n'est pas facile de définir cette p-value et donc de prendre une décision sans limiter le risque.

3.3. Automatiser les processus et prises de décisions

En se basant sur les modèles définis lors de l'étape d'analyse, il s'en suit de l'automatisation du processus. Cette automatisation permettra d'avoir des indicateurs mis à jour automatiquement pour piloter encore plus efficacement l'activité et prendre de meilleures décisions.

4. Exemple d'applications

Les applications du Big Data se fondent sur deux catégories et touchent de nombreux secteurs. Ce paragraphe met en évidence des types d'applications que l'on peut avoir avec le Big Data et propose quelques cas. Plusieurs applications existent ; toutes les applications ne sont pas présentées ci-dessous. Deux catégories se distinguent, une première par rapport aux données créées par les personnes et une seconde traitant des données machines.

4.1. Sujet aux personnes

Quand les personnes utilisent un ordinateur, internet ou appellent un helpdesk, elles laissent des traces. Ces traces sont des données non ou semi-structurées. Disponibles dans tous les supports physiques hébergeant des données comme les fichiers de log des sites web ou des enregistrements des incidents pour un helpdesk. Ce type de données peut être utilisé pour

améliorer la gestion de service, augmenter la satisfaction client, anticiper sur des problèmes produits ou encore analyser les comportements.

4.2. Machines

Nos machines génèrent en continu des informations qui peuvent être utilisées. Que ce soit des capteurs, des informations sur une transaction ou un processus interne, ces informations peuvent être utilisées en temps réel pour corriger des traitements, anticiper sur l'état de santé d'une machine ou encore réagir en temps réel sur des problématiques concrètes.

4.3. Exemples d'analyses de données personnelles

Les applications sont variées en fonction de chaque secteur ; cependant, chaque secteur peut rencontrer les deux catégories. Pouvoir traiter ces deux catégories est un gain réel pour tous les secteurs.

Le secteur Marketing utilise le Big Data pour mieux connaitre ses clients afin de proposer du contenu toujours plus personnalisé, mais également établir des stratégies pour définir les meilleurs emplacements pour du contenu (pub, lien, ergonomie, etc.).

L'analyse des réseaux sociaux est indispensable, car ils représentent une mine d'informations. Les gens postent régulièrement et gratuitement des informations sur ce qu'ils aiment ou n'aiment pas, sur leurs envies, leurs goûts. Il est alors possible d'obtenir des tendances sur des produits, évaluer la réputation d'un produit ou d'une entreprise, fidéliser les personnes pour proposer des réductions, des pubs directement aux bonnes personnes.

Internet regorge d'informations, le Big Data permet de connaitre exactement comment une personne se déplace sur les pages web d'un site ce qui permet d'améliorer l'ergonomie d'un site, de déterminer où placer le bon contenu pour qu'il soit le plus visible possible.

Dans l'exemple ci-dessous, nous avons une heatmap, elle représente via des zones « de chaleurs », la position du pointeur d'une souris sur un site web. Plus les zones sont rouges, plus l'utilisateur a passé du temps sur cette zone. Il est ainsi possible d'en déduire les endroits les plus utilisés, où la personne pose le plus longtemps son regard, où elle cherche l'information, etc. Grâce à l'interprétation, l'ergonomie du site peut être améliorée, les publicités mieux placées, etc.

Le Big Data, permet de récupérer ces données de déplacement et extraire des réponses en expliquant le comportement de navigation.

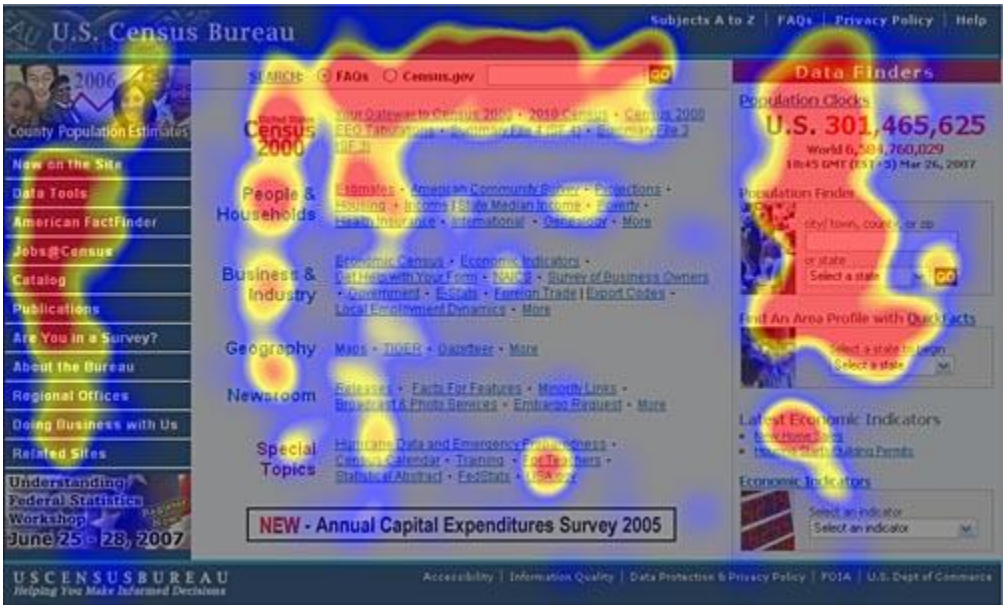

Figure 40: Heat Map pour l'étude des comportements en ligne (Nielsen, 2007)

Pour bien comprendre l'intérêt du Big Data vis-à-vis de la BI, prenons l'exemple de Minolta présenté lors du forum BI Oracle. La société a remarqué avec la BI traditionnelle que les ventes d'un produit étaient en baisse sur le dernier semestre. Ils ont alors utilisé les réseaux sociaux, les informations du web (forum, SAV, etc.) pour trouver une réponse à cette baisse. En croisant les informations, ils ont remarqué que le mot batterie était le plus récurrent. Les analystes ont donc creusé en détail les informations de batteries concernant le produit et il s'est avéré que la batterie de certains modèles de cette gamme était défectueuse. Ils ont alors changé de fournisseurs, modifié la gestion des applications au niveau de la batterie. Ils ont réagi en fonction de l'exploitation de la masse d'informations disponibles en ligne pour améliorer leur produit, ont communiqué sur une nouvelle gamme de produits avec une batterie performante et ont probablement évité une perte significative de clients.

4.4. Exemple d'analyses de données machines

Dans le secteur des télécommunications et plus précisément du mobile, des évènements (action telle qu'allumer/éteindre un terminal, passage d'un appel, etc.) sont enregistrés. Près d'un milliard d'évènements sont enregistrés chaque jour. (Antoine Crochet-Damais; JDN, 2013)Grâce à ces données et en les croisant via des données publiées gratuitement sur internet dites Open Data, il est possible de déterminer les flux de population que ce soit sur les autoroutes, les transports en commun ou à un point précis de la carte.

Dans l'industrie, les données machines seraient celles de toutes les machines sur un site de production. Par exemple, dans une usine de production de vin. La chaine de production part de de la cuve principale et chacune des bouteilles est remplie puis bouchée et enfin empaquetée. A chaque étape de la chaine, plusieurs machines gèrent une fonction. Grace aux technologies du Big Data, il est possible d'optimiser toute la chaine de traitement pour faire des économies.

PROBLEMATIQUE : QUELS FACTEURS SONT A PRENDRE EN COMPTE POUR PASSER D'UNE BI D'ENTREPRISE AU BIG DATA

1. Méthodologies

Pour répondre aux questions de recherche, le cas a été traité avec trois méthodes :

1.1. Interviews

Une enquête de terrain en interviewant :

Tableau 3: Liste des interviews

Qui	Fonction	But	Date	durée
MN	Responsable agence	Définir les besoins métiers	2 septembre 2013	1h
MM	Analyste marketing transverse	Définir les besoins métiers	10 septembre 2013	1h
JPL	Business Developer : Développer l'activité Big Data	Définir les perspectives et actions engagées sur le Big Data	20 septembre 2013	2h
JB	Consultant Valorisation Aider les entreprises à valoriser les données via le Big Data	Définir l'intérêt de la valorisation de données pour une entreprise et l'enjeu du Big Data	22 octobre 2013	1h
JG	Responsable Big Data	Impact sur la gouvernance	26 novembre 2013	1h

1.2. Participation à des forums et conférences

Enquête de terrain lors de conférences et Forum en BI et/ou Big Data.

Tableau 4: Liste des conférences

Nom	Sujet	Lieu	Date
Oracle	Stratégie et Big Data	GEM	13 mai 2013
Big Data chez IBM	Analyses de données Big Data,	ESIEA paris	23 mai 2013

	gouvernances et nouveaux rôles		
HP	Définition du Big Data et cas d'usage grâce à HP Vertica	Maison des centraliens Paris 7em	26 septembre 2013
Forum BI Oracle	Stratégie Oracle sur le Cloud Computing, le Big Data	Siège Oracle France	8 octobre 2013
La mine d'or du Big Data club du débat Polytech/HEC/Dauphine	Débat sur le Big Data, cas d'usage	Dauphine	4 novembre 2013
Talend open show	Big Data, MDM et mobilité chez Talend	Siège de Talend	12 novembre 2013

1.3. Etude documentaire

Une étude documentaire à l'aide de :

- livres sur la BI et le Big Data
- livres blancs
- documentations internes.

Dans la suite, une étude des besoins métiers via des enquêtes de terrain a permis d'établir les besoins actuels et futurs au sein de OBS. Par rapport à ces besoins, de l'existant ainsi que d'une étude documentaire, 3 scénarios d'architecture Big Data sont proposés et pour les deux plus pertinents, les facteurs et impacts sont détaillés.

ANALYSE DES RESULTATS ET PROPOSITION DE SOLUTIONS

Plusieurs scénarios sont possibles pour répondre aux besoins formulés par des intervenants métiers par rapport aux faisabilités du Big Data. Il est important de définir deux approches avant de se lancer dans une architecture: l'approche Make ou Buy. Make, pour « faire » en anglais, implique que la société se lance dans la conception de sa propre architecture Big Data. Buy, pour « acheter », revient à se procurer une solution complète ou appliance.

Concernant le Big Data, il n'y a pas d'architecture type, tout dépend de l'entreprise, de ses besoins en termes de volume, vitesse et variété des données. L'architecture mise en place sera dépendante de facteurs que l'entreprise doit prendre en compte pour appréhender le changement. Il y aura nécessairement des impacts et il est important de les développer. Cependant, les besoins métiers nécessitent-ils une architecture Big Data pour obtenir de la valeur ? Quels sont les scénarios possibles pour passer au Big Data et enfin quels seraient les facteurs et impacts ?

Les réponses à ces questions sont développées dans la suite du document. Une étude des besoins métiers a été effectuée de façon à définir ce que le métier pourraient faire avec du Big Data. Puis, trois scénarios sont proposés. Enfin, les facteurs et impacts sont détaillés.

1. Etudes des besoins métiers

Deux intervenants métiers ont été interviewés. La première personne est un correspondant métier, responsable de la restitution des données du portail des restitutions BI Business Object. Elle fait l'intermédiaire entre les sources de données et les métiers (Agences, Vendeurs, service Marketing et direction des grands comptes). Son rôle est de fournir des données de qualités et des Tableaux De Bords TDB pour le suivi des informations de CA et Parc à ses métiers. (MN, 2013)

La seconde personne est une analyste transverse du secteur Marketing. Elle est responsable des analyses statistiques dont le but est de fournir des données travaillées aux responsables des différents domaines (Voix, internet, mobilité, etc.) (MM, 2013).

1.1. Connaissance du Big Data

Les deux intervenants étaient au courant du Big Data, ils ont eu une présentation via des sociétés externes (IBM, SAS, etc.). Pour eux, le Big Data est un traitement massif de données, permettant d'obtenir une vision plus large avec un niveau de détail plus important dont les temps de réponse se rapprochent du temps réel.

Après explication, deux remarques ont été émises :

- En préalable d'une mise en place Big Data, il est nécessaire de faire un travail sur les référentiels de données.
- Les compétences nécessaires pour faire du Big Data ne peuvent être réunies en une même personne.

1.2. Les sources

Ils utilisent tous les deux une dizaine de sources mais uniquement deux d'entre elles possèdent des données non structurées. L'une dispose d'une base contenant des commentaires sous forme de fichier texte, PDF ou image, sur les remarques clients. L'autre

est également une base contenant des commentaires mais cette fois-ci sur la satisfaction client.

Les intervenants ont fait la remarque que l'entreprise est présente sur tous les canaux et devrait exploiter ces sources de données pour augmenter la satisfaction client, l'amélioration des offres et produits, etc.

1.3. Les besoins actuels

Les besoins actuels pour les deux intervenants sont globalement les mêmes :

- Fournir des données de qualité aux responsables des domaines, agences, etc.
- Pouvoir faire un suivi des données pour alerter les responsables le plus rapidement possible
- Mettre des données et des TDB à jour.

La satisfaction de ces besoins est différente d'une personne à l'autre. De manière générale, les besoins sont satisfaits mais des améliorations sont demandées sur :

- Les temps de réponse : les intervenants souhaiteraient avoir du temps réel et à minima, une réponse dans la minute

- Le niveau de détail : le volume de données devant être extrait dans le cas où la granularité serait plus fine n'est pas possible sur la plateforme Teradata actuelle sans impacter les performances

- Des données plus fraiches : actuellement, elles sont fournies à la fin de chaque mois, à un délai de **mois + 1** par rapport à l'arrivée de la donnée. L'idéal serait d'avoir un flux en temps réel et à minima une disponibilité à **jour + 1**

- Des données de meilleures qualités et disponibles dans une base centralisée

1.4. Les besoins nouveaux dus au Big Data

Les besoins dus au Big Data touchent des problématiques différentes. Pour les deux intervenants, il est nécessaire de pouvoir exploiter les informations des réseaux sociaux sur lesquels l'entreprise est présente, à savoir Facebook, Twitter, LinkedIn et Viadeo.

De plus, les informations de comportement des utilisateurs sur les sites web doivent être analysées. L'entreprise est active sur tous les réseaux sociaux mais à leurs connaissances, l'entreprise n'exploite pas ces informations. Il y a une perte d'information et donc un réel besoin à faire de l'analyse de réseaux sociaux et l'étude des comportements en ligne.

Par exemple :

Savoir si une entreprise cliente est présente sur un réseau social.

Définir son niveau technologique pour adapter une solution ou améliorer sa connaissance client (passer par le MDM pour trouver les bons contacts).

Récupérer les informations échangées avec les utilisateurs (sur les réseaux sociaux, sites web, forums pour améliorer :

- les processus de contact
- la connaissance client
- la satisfaction client

Les besoins ci-dessous résument les besoins fonctionnels :

- Identifier les clients présents dans plusieurs entreprises (influenceurs)
- Mieux identifier les clients et leurs comportements
- Analyses actuelles faites à posteriori
  - Besoins de Temps réel
  - Besoins de Prédiction
- Déterminer les nouveaux clients INSEE afin de les transformer en client
- Fouiller les informations non exploitées dans certaines sources car les analyses sont trop longues à réaliser

Les besoins en termes de Volume, Vitesse et Variétés sont présentés à la figure ci-dessous

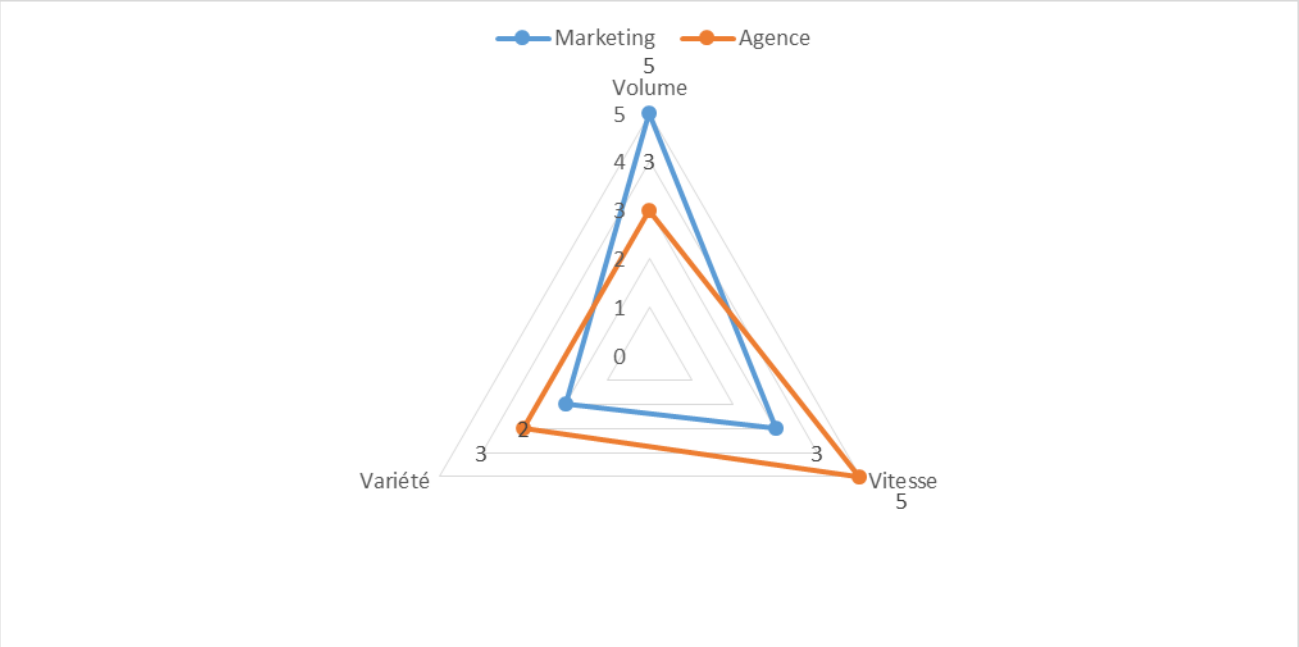

Figure 41: Schéma Radar des 3V selon les intervenants métiers

Le besoin de Big Data est bien présent pour les équipes utilisant les données que ce soit pour des problématiques de Volume, Vitesse ou Variétés. Les analyses de capteurs ont un intérêt dans le secteur des télécommunications mais dans le cas de nos intervenants, ils n'ont pas été discutés. De même pour la partie technologie de traitements de l'information (data management du Big Data). Les intervenants se positionnent surtout sur la partie Big Data Analytics.

2. Scénarios possibles

Pour supporter les contraintes de Volume, Vitesse et Variétés, une architecture BI traditionnelle peut suffire. Cependant, les performances ne risquent pas d'être à la hauteur des attentes. Mettre en place une plateforme décisionnelle gérant des volumes avec des vitesses de traitements rapides est possible avec un fournisseur comme Teradata, la plateforme actuellement mise en place pourrait donc théoriquement supporter les contraintes de Volume et Vitesse. Néanmoins, la variété des données pose un problème. Supporter cette contrainte tout en maintenant un niveau de service suffisant pour les utilisateurs implique de dépenser régulièrement de l'argent dans des améliorations. Il s'agit d'une évolution verticale des bases qui elle-même est limitée (architecture centralisée).

Stratégiquement, il vaut mieux utiliser les nouvelles technologies telles que Hadoop et les BDD NoSQL qui permettent de traiter efficacement ces données non structurées en conservant des temps de réponse rapides sur de gros volumes. La BI doit continuer à servir sa fonction principale qui est de piloter l'activité en utilisant les données structurées, traitées via l'ETL et qualifiées par les métiers. La BI donne une vision sure des données qui sont présentées et il est important de conserver ce premier aspect des données. Les technologies Big Data traitent de gros volumes de données et il est plus difficile de s'assurer que toutes les données sont fiables. La partie analytique du Big Data apporte une qualification sur le management des données fait dans ces technologies. Il est donc utile d'utiliser les différentes technologies Big Data et de les coupler à la BI pour corréler les informations.

Dans la suite trois scénarios sont étudiés. Une approche hybride où les technologies BI, Hadoop et les BDD NoSQL sont connectées entre elles. Une approche Cloud niveau groupe pour fournir différents services en interne et en externe. Enfin une approche Appliance qui en passant par un fournisseur de BDD permettrait d'avoir une plateforme Hybride clé en main. Il est nécessaire de détailler les approches à utiliser : celles du Make et Buy.

2.1. Approche Make or Buy

Les nouvelles technologies du Big Data sont à la base venues de l'Open Source. L'accès à ces technologies est alors gratuit. Certaines entreprises comme Cloudera, Hortonworks ou MapR se sont appropriées ces technologies, les ont retravaillées et proposent des solutions combinant des composants Hadoop, des bases NoSQL et des connecteurs vers les plateformes BI. Des géants des bases de données (comme Oracle, SAP, etc.) propose des appliances ; il s'agit d'une solution unique généralement une machine hébergeant toutes les solutions techniques (Hurwitz, et al., 2013).

Pour se lancer sur une plateforme Big Data, le choix de la concevoir soi-même ou de l'acheter est essentiel (JPL, 2013).

- Concevoir sa propre plateforme est moins coûteux, la solution sera au plus proche du besoin métier mais représente un risque car il n'y a pas de réel support en cas de panne ou de non fonctionnement.
    - Les solutions initiales sont Open Source, le seul support existant est la communauté en ligne participant à l'amélioration du produit et pouvant aider au cas par cas à résoudre les problèmes ;
    - Il y a un temps pour maitriser les outils, monter en compétences et obtenir une solution utilisable ;

- Le gain associé au coût est non négligeable. L'argent est investi dans la nouvelle plateforme et non pas dépensé dans une architecture ou une nouvelle licence.

- Si le choix se porte sur l'achat d'une solution préfabriquée, le support est meilleur, la solution est directement utilisable mais des incompatibilités peuvent survenir, la réponse aux besoins est moins précise et le coût est plus élevé.
  - Logique client / infrastructure de l'éditeur => problème de compatibilité avec des sources ;
  - Des contrats de service régulent le support et la qualité de la prestation ;
  - Support de qualité : les bugs ont moins d'impact ;
  - Plus cher car il y a des coûts des licences, des plateformes, des consultants, etc.

Peu importe la stratégie choisie, il faudra nécessairement que la méthode de management utilisée soit agile, le cahier des charges de la méthode en cascade est trop lourd pour assurer un développement orienté utilisateurs.

2.2. Scénario 1 : approche Hybride

Si la BI met à disposition des données hautement structurées et qualifiées selon des modélisations adaptées à l'analyse de données. Le Big Data lui offre la possibilité de traiter et d'analyser des données non (ou semi) structurées sur des volumes et temps de réponse incomparables avec la BI. Néanmoins, ces deux méthodes de gestion de l'information fonctionnent très bien ensemble ; détenir et utiliser les deux méthodes en parallèle représentent un réel gain.

Les couches de l'approche Hybride (Krishnan, 2013)

En comparant les méthodes de data management, la seule différence existe sur la manière d'intégrer, de traiter et de visualiser les données. Les méthodes de traitement des données de la BI et du Big Data reposent sur trois couches : données, technologies et analytiques.

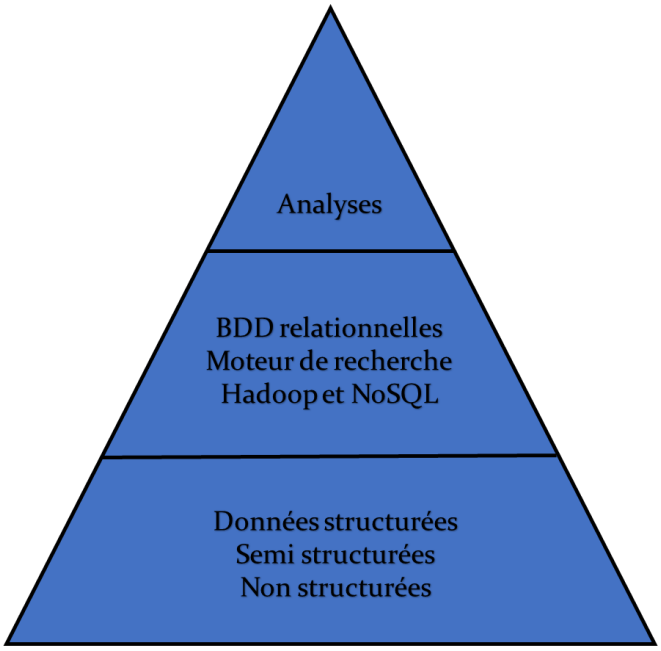

Figure 42: Couches de l'approche Hybride

Une architecture BI voulant traiter des données selon les 3V risque de rencontrer des problèmes de performances. Il est possible de faire du Big Data avec une architecture BI tout comme faire de la BI avec une architecture Big Data. Cependant, chaque architecture a son rôle à jouer et il faut utiliser intelligemment les données.

Les performances

Les temps de réponse et la charge que peuvent supporter une architecture est un problème qui dépend des utilisateurs, des données, du volume requêté et des performances des technologies utilisées. Plus le nombre d'utilisateurs qui requête simultanément sur la base est grand, plus les temps de réponse seront long, de même si les requêtes sont complexes. Cette charge reste vraie pour des données structurées comme non structurées. Il est donc important d'utiliser les bonnes méthodes sur les bonnes données et de séparer les technologies.

L'architecture

Dans cette approche, chacune des méthodes vont suivre leurs processus de management des données. D'un côté, les sources de données structurées seront traitées via l'ETL puis stockées dans le DWH/DTM, puis analysées et enfin visualisées sous forme de rapports, tableaux de bord, etc. De l'autre côté, toutes les données structurées (ou non) internes ou externes sont acquises, puis traitées/stockées dans les bases non relationnelles (Hadoop et NoSQL), ces données sont, par la suite, découvertes et analysées. Pour permettre une communication entre ces deux méthodes de traitement, des connecteurs existent.

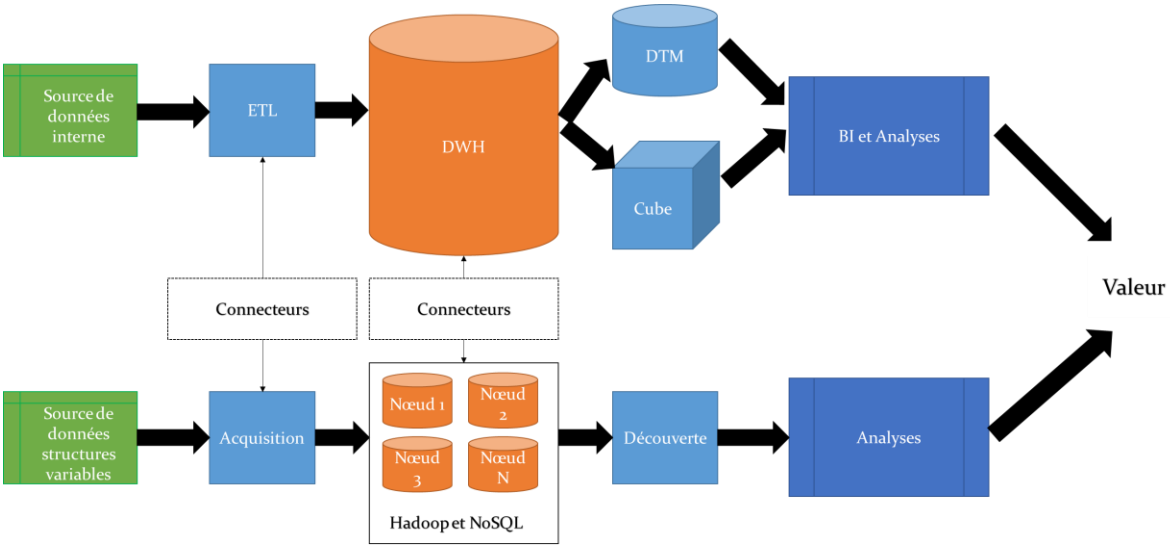

Figure 43: Architecture Hybride

Les connecteurs sont disponibles à différentes étapes :

- Pour la partie stockage, le DWH/DTM et les bases non relationnelles peuvent communiquer entre eux.
- L'ETL a la possibilité de se connecter aux bases non relationnelles pour récupérer des données prétraitées afin de les stocker provisoirement ou définitivement dans le DWH pour de futures analyses (toutes les données ne peuvent pas nécessairement être intégrées).
- Avec la partie analytique les résultats trouvés via les méthodes de datamining sur les données structurées peuvent être approfondis grâce à ceux du Big data. Des connecteurs peuvent alors exister depuis l'outil d'analyse BI vers Big Data ou Big Data vers BI.

Il est important d'utiliser les différentes méthodes de traitements en respectant la fonction de chacune. Elles peuvent communiquer et interagir via des connecteurs. Cette approche se base plus sur la façon de faire, « Make », car elle exploite la BI d'Entreprise existante tout en développant la partie non relationnelle via les technologies Open Source. Il est toutefois possible de passer par des revendeurs comme Cloudera, Hortonworks ou MapR pour connecter le DWH/DTM avec les bases NoSQL.

L'approche Hybride est la plus logique pour passer au Big Data, elle permet de conserver la BI traditionnelle en ajoutant simplement la couche Big Data. Si aucune base non relationnelle n'est utilisée en entreprise, une méthode d'adoption pourrait être d'intégrer la partie Hadoop, de monter en compétence sur le sujet et trouver de la valeur à extraire. Puis de réitérer sur un composant d'Hadoop (SQOOP, HIVE, etc.) ou une base NoSQL. Ceci permettra de construire pas à pas son architecture hybride en conservant les fondations robustes des systèmes décisionnels. La seconde approche repose sur le Cloud Computing pour offrir des services Big Data en fonction des besoins des utilisateurs.

2.3. Scénario 2 : plateforme Cloud niveau groupe (JPL, 2013)

Une stratégie efficace pourrait être de développer une architecture en Cloud IAAS, PAAS et SAAS. Chaque type de Cloud pourrait fournir différents niveaux de service. Cette architecture permettrait de disposer en interne de solutions adaptées aux besoins de chaque division désirant accéder aux opportunités du Big Data.

IAAS

Dans cette infrastructure, un ou plusieurs clusters sont mis à disposition. Les divisions sont responsables des plateformes qu'elles souhaitent installer. Elles doivent configurer leurs composants Hadoop, leurs bases NoSQL et sont responsables du traitement et de l'analyse des données. Les divisions utilisent l'infrastructure technique et gèrent leurs budgets sur la partie logicielle et application.

PAAS

Pour ce type de Cloud, une plateforme dédiée à Hadoop ou à une BDD NoSQL particulière est mise à disposition. Les divisions métiers doivent être capables d'utiliser la plateforme et être responsables du management et de l'exploitation de leurs données. Les utilisateurs ont le choix des outils de visualisation et d'analyse.

SAAS

Sur ce service Cloud, les clusters sur lesquels sont exécutées les solutions Hadoop et NoSQL sont fournis et un catalogue de service contenant les différentes solutions de visualisation et reporting permettent aux utilisateurs d'avoir une solution prête à l'emploi pour traiter les données.

L'architecture Cloud aura la forme suivante :

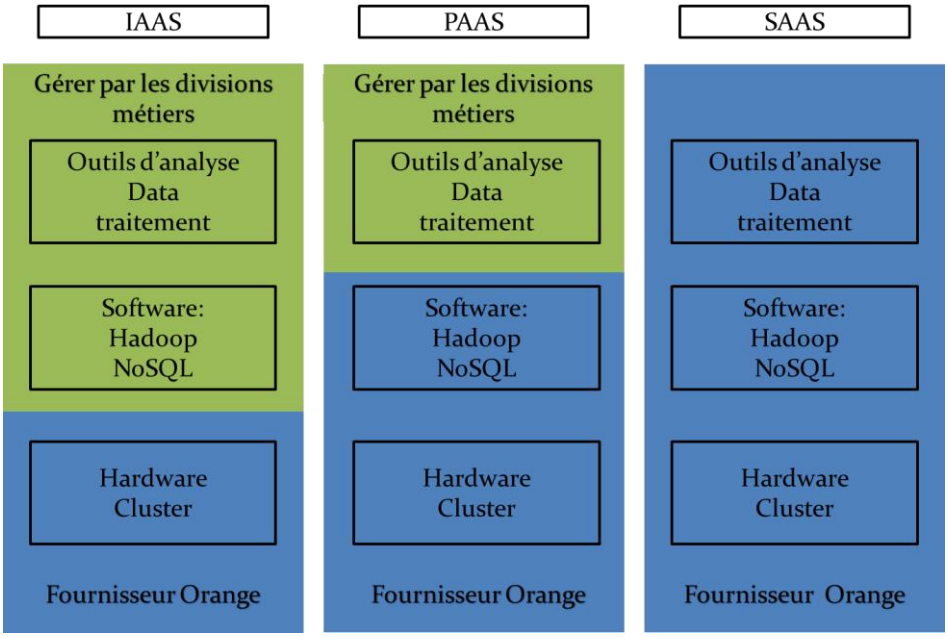

Figure 44: Architecture Cloud niveau Groupe

L'avantage de cette solution est de profiter des data center du Cloud pour utiliser les architectures distribuées et ainsi optimiser les ressources mises à disposition aux utilisateurs. Comme pour le scénario 1, les systèmes décisionnels sont existants et des connecteurs permettent de faire le lien entre les bases relationnelles et non relationnelles.

Cloud privé ou public ?

L'usage principal serait de fournir aux divisions métiers des solutions internes pour exploiter le potentiel du Big Data. La stratégie pourrait également d'être fournisseur de Cloud sur le marché B2B. Néanmoins, il est nécessaire de définir quelle typologie du Cloud à mettre en place. Les données qui vont circuler au sein des solutions peuvent être confidentielles d'une division à l'autre mais aussi depuis l'extérieur. Au vue de la stratégie de ce scénario :

- Un Cloud Interne – Privé est la solution adéquate pour gérer efficacement les services Cloud définis ci-dessus. Il assure la confidentialité et la sécurité des données en interne et permettra de contrôler les aspects technologiques et financiers.

- Un Cloud Interne – Public pourrait être une solution dans le cas où l'entreprise se met en position de fournisseur de Cloud sur les solutions Big Data.

- Un Cloud Externe – Privé, où le Cloud est utilisé en interne mais géré par un tiers, n'est pas une solution adaptée. Il comporte trop de risques en termes de confidentialité, sécurité et qualité de service pour qu'un engagement sur un sujet stratégique comme le Big Data soit intéressant. Néanmoins, ça peut être une alternative pour essayer le Big Data et voir s'il y a un intérêt pour une division métier.

- Un Cloud Externe – Public n'est pas une bonne solution dans notre contexte pour des raisons de confidentialité et sécurité.

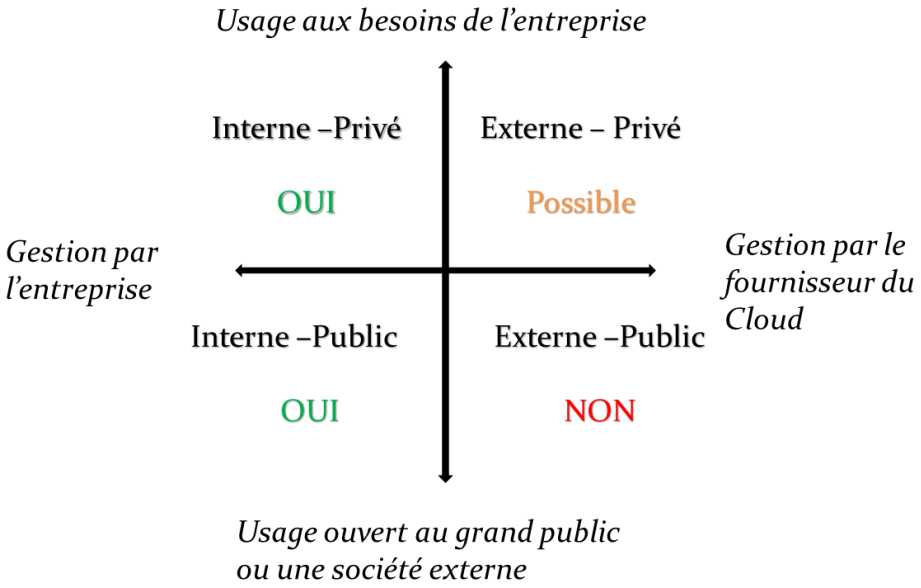

Figure 45: Typologies du Cloud pour l'architecture Cloud niveau Groupe

L'entreprise possède déjà des data centers pouvant supporter la charge d'une architecture Cloud. La société dispose également d'offre Cloud pour le grand public. De plus, elle utilise déjà les différentes solutions Hadoop et NoSQL. Il n'y a donc pas de difficulté technique pour mettre en place cette architecture en Cloud.

Cependant, suite à un échange avec un responsable du Big Data (JG, 2013), le Cloud tel qu'il est, n'est pas adapté par rapport à une stratégie Big Data. Le Cloud actuel est orienté stockage pour particuliers et professionnels. Les architectures techniques mises en place sont donc optimisées avec des machines performantes pour assurer le stockage, l'accès aux données et la consistance des données. Pour le Big Data, une stratégie d'optimisation de coût est nécessaire. Chaque machine utilisée doit avoir une configuration bas de gamme car l'intérêt est de pouvoir utiliser des ressources machines multiples pour de petits traitements. Les machines ne doivent donc pas disposer de stockages performants, de processeurs ultrarapides (les traitements se font sur de petits volumes de données) ou de grandes quantités de RAM. Il faut alors acheter des machines peu coûteuses pour faire évoluer horizontalement le Cloud tout en minimisant les coûts.

Les gains d'une telle architecture seront significatifs, si chaque division métier est en mesure de profiter des avantages du Big Data, que ce soit pour de l'analyse comportementale ou analyse de capteur, les gains seront présents.

Ce scénario met à disposition un ensemble de solutions Big Data, il est ensuite possible de connecter les résultats obtenus avec le Cloud Big Data et ceux de la BI traditionnelle comme proposé au scénario 1. Si l'entreprise souhaite adopter une approche purement Buy, il faudra qu'il passe par de l'appliance.

2.4. Scénario 3 : appliance

Un dernier scénario étudié est l'appliance, il s'agit de solution « sur–étagère » prête à l'emploi et comportant toutes les technologies souhaitées pour faire de la BI et du Big Data sur la même architecture. Les appliances sont fournies par les éditeurs de logiciel et/ou les hébergeurs de BDD. Par exemple, Oracle, initialement fournisseur de BDD mais également éditeur de logiciel BI, est fournisseur de Cloud et de solution Big Data. Oracle propose une solution Big Data qui embarque les plateformes Hadoop, NoSQL, DWH Oracle, BDD in-memory, etc.

Les appliances peuvent être comparées à de grosses boîtes noires dans lesquelles plusieurs couches matérielles et logicielles sont installées pour répondre aux besoins sur une même machine. Les solutions technologies intégrées appartenant à la même société, la communication entre les différentes couches est optimisée (Krishnan, 2013).

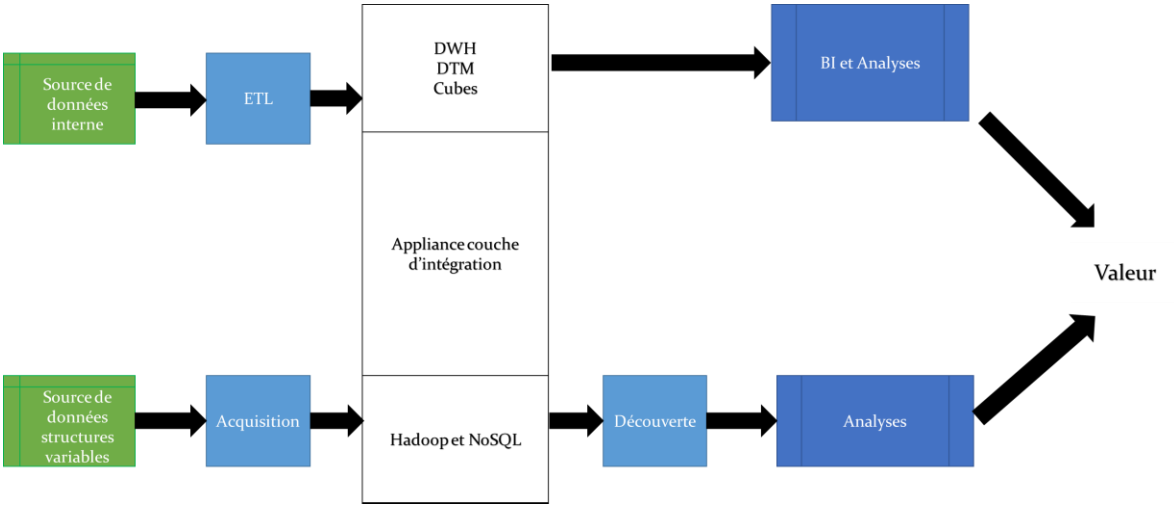

Figure 46: Architecture via appliance

L'avantage de ce scénario est que la solution est immédiatement disponible, les fournisseurs ou des sociétés de conseil peuvent accompagner l'implémentation de la solution et former le client aux technologies ce qui facilite la transformation du business. De plus, les technologies relationnelles et non relationnelles sont totalement intégrées et les logiciels BI/Big Data de traitement des données sont tous disponibles.

L'inconvénient du scénario est simple, choisir de prendre une appliance implique de signer pour plusieurs années sur une solution coûteuse dont les seules modifications pourront être apportées par le fournisseur ou revendeur agréé. De plus, les appliances s'intègrent moins bien avec des solutions (BDD ou logiciels) existantes et provenant de fournisseurs différents.

2.5. Choix du scénario

La stratégie de l'entreprise conduira au choix du scénario. Si la stratégie est de développer sa propre solution en interne les scénarios 1 et 2 seront privilégiés. Si un choix Corporate porte sur un fournisseur/éditeur particulier, une appliance sera probablement un meilleur choix.

Compte tenu que l'entreprise dispose déjà d'appliances, de solutions de nombreux fournisseurs et éditeurs de logiciel ; reprendre une appliance ne serait pas le meilleur choix. Il impliquerait de dépenser de l'argent dans une nième solution avec une démarche projet longue et elle aussi coûteuse.

Le scénario 1 se positionne sur une politique d'investissement sur le long terme pour une division métier précise. Il obéit à une logique de découverte et d'exploitation du Big Data par rapport à une problématique métier donnée. Chaque division métier devra mettre en place sa propre architecture en parallèle de la BI existante. Autrement, la DSI et ses centres de compétences devront proposer des solutions pour ces divisions métiers et gérer les architectures physiques au cas par cas.

Le scénario 2 est une solution d'investissement également sur le long terme mais pour l'ensemble des divisions métiers souhaitant utiliser le Big Data. Il rendra le plus haut niveau de service et semble être la solution la plus adaptée. De plus, il est possible de résumer le

scénario 2 comme une multitude de scénario 1 gérés en interne pour faciliter l'accès aux données.

Le scénario 2 serait l'étape suivant le scénario 1 car il faut déjà maîtriser le scénario 1 pour passer au 2. Dans la suite, les facteurs et impacts portent sur le scénario 1 car il sera possible de les généraliser en évoluant vers le scénario 2.

Peu importe l'architecture, certaines compétences sont nécessaires pour tirer le meilleur profit du Big Data et surtout valoriser un ROI (Retour sur investissement).

3. Facteurs

La BI nécessite des compétences techniques pour requêter sur les bases, utiliser les outils d'alimentation (ETL) ou de reporting. Egalement, l'aspect analyse de données se base sur des compétences statistiques et informatiques maitrisés. Les ROI BI sont bien connus et réputés pour être rapides et élevés. Les méthodes de management en BI sont encore souvent orientées en cycle en V mais tendent de plus en plus vers de l'agilité. Le Big Data apporte de nouvelles technologies, de nouvelles opportunités sur les données à analyser. Cependant, est-il possible de prévoir un ROI sur une architecture Big Data et quelles seraient les méthodes de management à mettre en place ?

3.1. Les compétences techniques

La dimension technique du Big Data et plus exactement la partie Big Data Management, demande des compétences en gestion de bases de données relationnelles et surtout non relationnelles, des compétences en développement logiciel et enfin en architecture informatique (JPL, 2013).

Concepts de gestion de Bases de données

Le Big Data fonctionne grâce à des systèmes de fichiers distribués et des BDD non relationnelles. L'écosystème Hadoop est composé du noyau principal avec HDFS et Map/Reduce et des Frameworks supplémentaires avec HIVE, PIG, SQOOP, etc. L'écosystème étant récent, il est difficile de trouver une seule et même personne capable de tout maitriser. De la même manière, les bases NoSQL se découpent en base orientée colonnes/familles, graphe, document. Chacune de ces BDD sont spécifiques, elles ont des avantages, des inconvénients et sont plus ou moins récentes. Il est donc difficile d'avoir toutes les compétences nécessaires pour utiliser l'ensemble des solutions du Big Data pour en tirer le meilleur profit. Les ingénieurs BI et SI seront capables de répondre à ces tâches.

Développement

Tous les logiciels et architectures techniques sont fondés sur des langages de programmation. Pour simplifier, ils se distinguent en trois catégories :

- Les langages de scripts. Ce sont des fichiers qui lancent des instructions en ligne de commande directement au système d'exploitation (environnement sur lequel toutes les applications tournent comme Windows, Linux, Mac OS). Par exemple, les bash et batch respectivement pour Linux et Windows. Ces scripts peuvent faire toutes les choses qu'un utilisateur exécute en graphique comme par exemple ouvrir, créer, supprimer un dossier, lancer un programme, etc.

- Les langages séquentiels comme le langage C sont des programmes qui exécutent des fonctions les unes après les autres en utilisant les principes d'algorithmie.

- Les langages Orientés Objets comme le Java ou le C++ permettent de définir des Classes d'objets et de créer des programmes basés sur ces classes.

Les technologies comme Hadoop et les BDD NoSQL viennent de l'Open Source, elles sont majoritairement conçues en langage Java et tournent sur des invites de commandes. Il est donc indispensable de connaitre un langage de script car ces scripts pourront ouvrir et exécuter des programmes Orientés Objets comme Java ou séquentiel. Les développeurs pourront effectuer ce rôle, mais il sera probablement nécessaire d'en embaucher plusieurs maitrisant les différents langages.

Architecture de base de données

Pour utiliser les architectures, développer des programmes de traitements et donc extraire de la valeur du Big Data, il faut être capable d'installer et de maintenir les plateformes Big Data. Un Data architecte est donc indispensable, il est responsable du bon fonctionnement des architectures BI et Big data dans l'entreprise (Messatfa & IBM, 2013).

3.2. Les compétences analytiques

Pour la partie Big Data Analytique, les fondements sont les mathématiques et les statistiques. Les méthodes de data et web Mining ont une place importante. Enfin la Data Visualisation pour l'exploration de données est indispensable (JPL, 2013).

Mathématiques et statistiques

Pour appliquer les méthodes descriptives et prédictives, les mathématiques et statistiques sont nécessaires. L'implémentation des modèles prédictifs pour faire parler les données et trouver de valeur dans les données. Un ingénieur de formation en statistiques ou mathématiques appliquées est nécessaire (MN, 2013).

Data/Web mining

Un data et un web miner pourront relier les méthodes statistiques avec les outils informatiques. Le data miner s'occupe des informations du type satisfaction client, lien entre produit et CA, etc. Le web miner exploite les informations des sites web pour exploiter de la valeur.

Data Visualisation

La data visualisation est indispensable au Big Data, les data miners peuvent remplir ce rôle. Ils peuvent utiliser la data visualisation pour représenter les données qu'ils viennent de travailler. Néanmoins, les Data Analyst, que l'on peut voir comme les data miner du Big Data sont les personnes à embaucher pour faire parler les données avec ces outils de Data visualisation.

3.3. Compétences fonctionnelles

Pour répondre efficacement à un problème en faisant parler les données structurées et non structurées, il faut une très bonne connaissance métier. Si cette connaissance est insuffisante, les résultats attendus risquent d'être erronés et donc les décisions prises fausses. En effet, un modèle mathématique qualifié pour prédire un comportement, peut au final, mener à de mauvaises décisions si les enjeux Business n'ont pas été bien compris.

3.4. Le Data Scientist

Pour remplir toutes les compétences ci-dessus, un métier est apparu avec le Big Data : le Data Scientist. Cette personne doit avoir les compétences : informatiques, statistiques et fonctionnelles. Il doit être capable de récupérer les données structurées et non structurées, faire des analyses statistiques poussées pour explorer et modéliser les comportements, savoir visualiser et présenter les résultats. Il doit surtout avoir une forte composante Business pour répondre efficacement aux problématiques métiers.

En résumé, il est au croisement des métiers de statisticien, ingénieur informatique et expert métier.

3.5. ROI

Il est difficile d'évaluer le ROI sur les projets Big Data. Actuellement, très peu d'entreprises françaises osent franchir le pas du Big Data. Généralement, elles ont effectué des Proof Of Concept POC qui consistent à essayer les technologies soit en interne soit avec l'aide de sociétés de service spécialisées et de voir s'il y a un intérêt à franchir le cap Big Data (JB, 2013), (JPL, 2013).

Les entreprises doivent donc aborder une démarche Test & Learn dans laquelle elles essaient les technologies en passant par un Cloud ou en installant les plateformes Open Source. Puis, elles montent en compétences sur le Big Data Management et le Big Data Analytics. Si les services rendus sont satisfaisants alors elles peuvent tenter de se lancer en configurant ces propres plateformes.

Dans tous les cas, si un investissement devait être fait sur une technologie Big Data, il y aurait :

- Des coûts d'infrastructure pour se procurer des clusters où installer les nouvelles technologies

- Des personnes aux compétences bien spécifiques à embaucher ou à former en interne (les profils sont rares et donc peuvent coûter cher) par exemple en France un Data Scientist est embauché entre 45 et 120K€ selon le profil et de l'ordre de 150K$ aux Etats-Unis. Les Data Scientists, Data Analysts, Ingénieurs BI/Big data et experts métier sont nécessaires.

- Des prestataires de service peuvent aider à accompagner le changement (Maitrise d'Ouvrage MOA), aider les équipes à l'implémentation des solutions MOE (Maitrise d'Œuvre, etc.

De plus, la stratégie Make ou Buy aura un impact sur l'investissement. En fonction des technologies et logiciels de traitement de l'information, des coûts de licence peuvent s'y ajouter.

Pour assurer un déploiement efficace, il faut mettre en place des méthodes de management adaptées.

3.6. Méthodes de management (JPL, 2013)

Nous avons vu que l'approche Test & Learn est la plus adaptée pour déterminer les gains possibles du Big Data. Un projet Big Data n'a donc pas de but précis lors de son lancement, le but global est de trouver de la valeur venant de données non structurées mais rien n'assure que de la valeur sera extraite des données que l'on souhaite analyser. En suivant cette logique une méthode de projet en cycle en V n'est pas conseillée. En effet, les besoins risquent fortement d'évoluer au cours du temps, leurs définitions est donc compliquées et l'effet tunnel à de forte de chance de faire échouer le projet.

Pour s'adapter à la méthode de Test & Learn, utiliser les méthodes agiles est un bon compromis. Elles permettent d'adapter le développement et les analyses en fonction des besoins des utilisateurs. Une itération peut représenter une première analyse, si cette analyse amène à un résultat, l'itération suivante sert à l'approfondir.

De plus, il pourrait être bon de s'inspirer des méthodes de management à l'américaine, là où le Big Data a une place importante et stratégique.

- Encourager le succès ;
- Evoluer facilement ;
- Avoir le droit à l'erreur.

Ces trois points facilitent l'adoption du Big data. En effet, le Big Data apporte de nouvelles technologies, n'assure pas une réussite de 100% du travail effectué, se tromper permet d'évoluer, de recommencer en étant plus efficace et donc de perdre de moins d'argent sur le long terme.

4. Impacts

Le passage au Big Data peut être compliqué si l'entreprise n'est pas préparée. Une mauvaise préparation aura des impacts notables sur les restitutions et donc sur la prise de décision. Pour faire communiquer les architectures décisionnelles avec celles du Big data, un travail sur les métadonnées est nécessaire à chaque phase du cycle de vie de la donnée. De même, pour assurer une cohésion sur les données, un projet de Master Data Management MDM doit être mené. Enfin, une gouvernance solide de la donnée doit voir le jour au sein de l'organisation pour définir les règles, assurer la sécurité et la responsabilité des membres de l'entreprise.

4.1. Les métadonnées

Les métadonnées sont des informations sur les données. Elles ont une place centrale dans tous les environnements contenant des données car elles permettent d'intégrer et d'interroger facilement ces données. Dans une BDD, elles fournissent des informations

variées telles que la description du contenu de chacun des champs de toutes les tables. Les utilisateurs peuvent ainsi avoir une cartographie de la BDD.

Les types de métadonnées (Krishnan, 2013)

Les métadonnées peuvent être générées automatiquement par des règles business, créées par les concepteurs des systèmes ou fournies par des sources externes. Au total, près de neuf types de métadonnées circulent dans les systèmes :

- Les métadonnées techniques qui sont associées aux règles de transformation, des structures de stockages ainsi que des couches sémantiques.

- Les métadonnées business qui décrivent les informations des données d'un DWH ou DTM telles que :
  - La structure de la donnée
  - La valeur des attributs
  - Date et heure de la création ou dernière modification
  - Règle de changement.

- Les métadonnées de contexte qui donnent des informations sur la manière de traiter des données non structurées (texte, image, vidéo, etc.).

- Les métadonnées d'alimentation avec les informations de :
  - Table source
  - Table cible
  - Algorithmes
  - Règles de gestion
  - …

- Les métadonnées de programme avec les informations sur le programme et ses actions.

- Les métadonnées d'infrastructure comportant pour les sources et cibles, les informations de plateformes, réseaux et contacts.

- Les métadonnées de business avancé qui fournissent des informations sur la précision, la fréquence de mise à jour et des systèmes utilisant les métadonnées Business.

- Les métadonnées opérationnelles contenant des informations, de fréquences d'usage, durée d'exécution et sécurité.

- Les métadonnées BI apportant des informations sur la manière dont les données sont traitées, filtrées, analysées et affichées.

Les métadonnées sont très importantes pour le Big Data parce que les données ne sont plus structurées et transitent dans des flux volumineux et rapides. Dans une analyse de plusieurs Téraoctet d'information, il n'est pas possible de connaitre toutes les informations sur

chacune des données. Les métadonnées permettent de suivre le traitement de ces gros volumes et d'assurer une traçabilité sur les données analysées. A la variété s'ajoute l'ambiguïté des données ; c'est-à-dire que la donnée d'un type particulier peut être définie avec des versions différentes (exemple document word .doc ou .docx). L'organisation de ces données non structurées est un vrai challenge ; celles-ci ont besoin de métadonnées de contexte. Cependant, sans une contextualisation adaptée, le rattachement à d'autres données ne pourra se faire, ce qui provoquera de faux résultats.

4.2. MDM et référentiels (Iafrate, 2013)

Le Master Data Management est un processus de gestion de l'information qui vise à standardiser les données centrales d'une entreprise. Il permet de centraliser des données d'excellente qualité sur un sujet précis. Ces données sont dites de référence également appelées Master Data. Elles concernent des données de type :

- Client
- Produit
- Fournisseur
- Employé
- Etc.

Sans MDM, ces données centrales peuvent être dupliquées avec des attributs différents dans plusieurs supports de stockage à travers l'entreprise.

Par exemple, le service Marketing web dispose des informations suivantes sur un client :

Nom : Dupont

Adresse : 1234 avenue de la paix Paris

Numéro de téléphone : 06 12 34 56 78

Mail : Dupont@monmail.com

Du côté du service de marketing direct, les employés disposent des informations suivantes :

Nom : Dupont

Adresse : 1345 avenue de la paix Paris

Téléphone : 06 98 76 54 32

Lors d'un traitement, chaque service va effectuer ses requêtes sur sa propre base. Cependant, lors de la consolidation des informations, des erreurs peuvent apparaitre parce que les données utilisées ne sont pas qualifiées.

Figure 47: Gestion des référentiels sans MDM

Avec du MDM, une base centrale regroupe toutes les données de référence et les différents services viennent piocher leurs informations de clients, produits ou employés directement dans cette base.

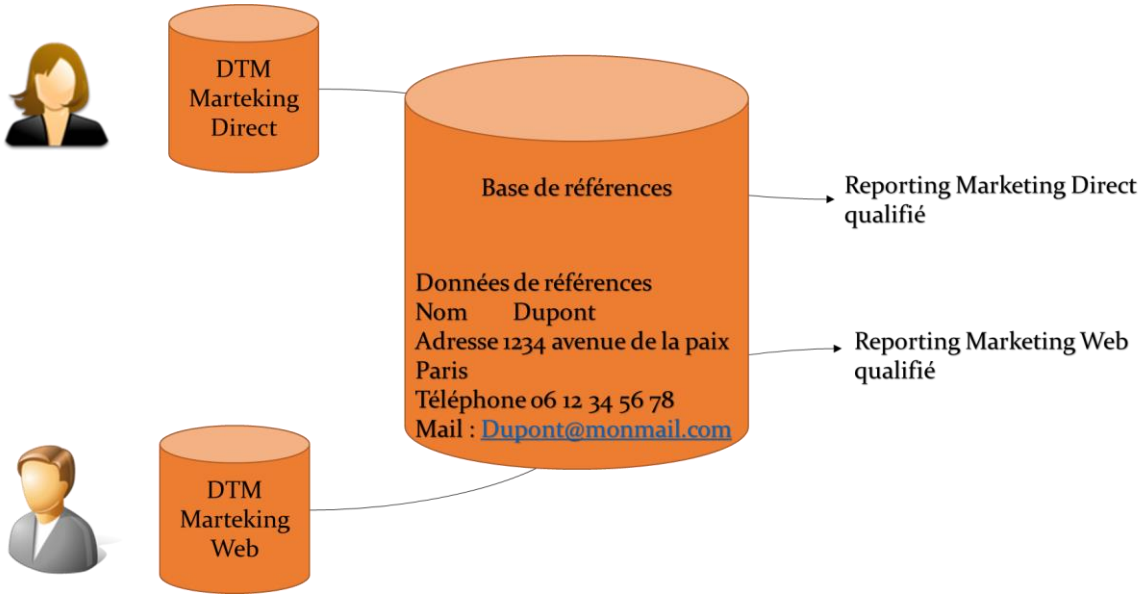

Figure 48: Gestion des référentiels avec MDM

Les Masters Data et les métadonnées sont toutes les deux utilisées pour assurer un traitement sur des données pour le Big data.

4.3. Traitement des métadonnées

Pour traiter des données non structurées de type Big Data, il faut un nouveau type métadonnée qui peut être associée à cette donnée pendant un traitement. Nous avons vu précédemment que le data management du Big Data se fait en cinq étapes : Collecte, Analyse, Structuration, Transformation et Extraction. La Métadonnée est créée dès la seconde phase. Lors de l'analyse, la donnée est :

- Taguée, à ce moment on assigne une métadonnée à l'information.
- Classifiée, les données sont regroupées par famille (exemple : données client d'un côté, fournisseur de l'autre). La classification permet d'optimiser le traitement de la donnée non structurée par la suite.
- Modélisée pour la préparer à sa structuration.

Lors de la phase de structuration et de transformation de nombreux changements sont effectués à la métadonnée. En effet, à ces étapes des contextes sont attribués ; il est très important d'associer le bon contexte à la bonne donnée car il va permettre de faciliter les analyses en fin de chaine une fois que la donnée sera stockée. La contextualisation peut s'apparenter à l'indexation de mot clé dans un moteur de recherche. Par exemple, si une entreprise souhaite être visible sur internet, elle va définir des mots clés à indexer. Quand un utilisateur effectuera une recherche, si les mots clés sont ceux de l'entreprise, elle aura des chances d'être en tête des résultats. Par conséquent, si le contexte n'est pas bien défini, lors d'une analyse, une donnée pourra être utilisée comme référence à un sujet qui n'est pas le sien.

Lors de l'extraction, des liens sont créés entre les métadonnées et certaines données de références. Ces liens permettent de créer des relations entre des données non structurées et structurées.

Si un travail sur les métadonnées et le MDM est essentiel pour assurer la cohérence des traitements du Big Data, une gouvernance sur la donnée est indispensable et peut impacter l'organisation de l'entreprise.

4.4. Gouvernance des données et transformation de l'organisation de l'entreprise

Dans une entreprise qui manipule des données pour un usage stratégique, connaitre d'une information de type « qui en est responsable ; a le droit d'y accéder » ou encore savoir à quoi elle correspond, est critique. La gouvernance des données peut se définir par toutes les solutions qui régissent la prise de décision et l'autorité autour des données (Data governance Institute, 2004).

La gouvernance

Toute entreprise utilisant les architectures BI pour améliorer sa prise de décision possède une gouvernance autour de la donnée. Cette gouvernance traite des sujets relatifs à la

signification des données, leurs règles, qui sont les propriétaires, les utilisateurs habilités, etc.

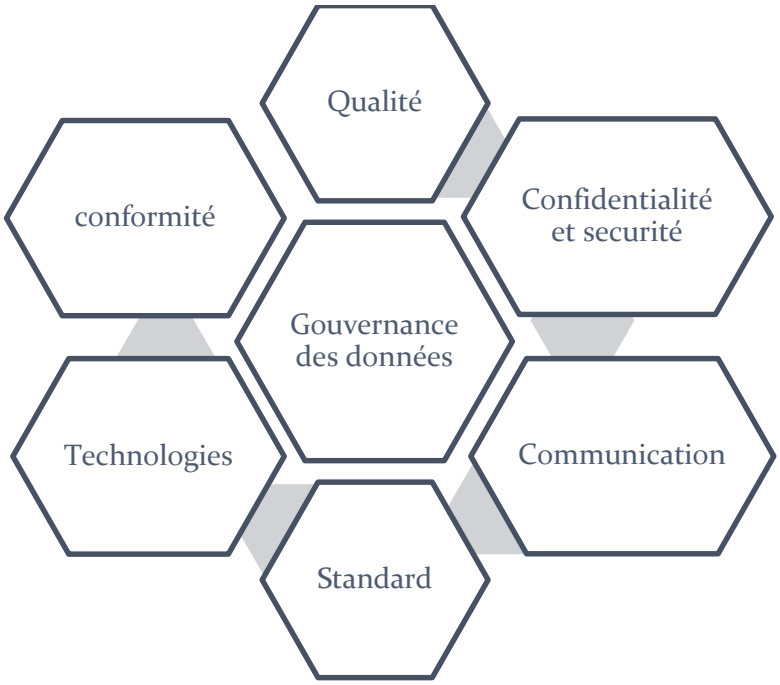

Figure 49: Gouvernance des données

Dans le cas de la BI, la gouvernance des données est facilitée par la structure des données. La chaine de traitement BI implique que l'ETL travaille une donnée pour la stocker. Cette donnée est associée à une métadonnée qui permet à la gouvernance de tracer et d'assurer sa qualité et sa sécurité. Le fait qu'elle soit stockée en un endroit précis et connu comme le DWH ou les DTM assure l'intégrité et le contrôle des accès. L'architecture centralisée de la BI assure une gouvernance fiable.

Cependant, dans le cas du Big Data, l'architecture distribuée implique que les données sont éparpillées et le plus souvent répliquées à des endroits multiples. De plus, les volumes de données utilisés étant beaucoup plus important que ceux de la BI, il est nettement plus difficile de gérer les données et leurs accès. Pour pouvoir gérer, tracer et donc répondre aux contraintes de sécurité et confidentialité des données, l'entreprise doit changer son organisation.

La transformation

Le Big Data étant essentiellement développé aux Etats-Unis, l'inspiration est à prendre dans leurs organisations. Les données étant stratégiques, très nombreuses et difficilement traçables, une nouvelle entité interne à l'entreprise doit être responsable à 100% de la gouvernance des données.

Dans une organisation classique, les responsabilités sont séparées entre la Direction des Systèmes d'Information et la Direction Administrative et Financière. Dans la nouvelle organisation, une entité dédiée à la donnée doit apparaitre avec les mêmes niveaux de

responsabilités que les deux autres cités ci-dessus. A la tête de cette entité, le Chief Data Officer CDO est un membre de l'équipe de direction de l'entreprise et a pour responsabilité de gérer ce qui est en rapport avec la donnée. (Messatfa & IBM, 2013), (JG, 2013), (JPL, 2013).

CONCLUSION

Le volume des données augmente constamment et n'est pas prêt de ralentir. Les données sont pour la plupart générées sur le Web, lors de la navigation, de partage d'informations ainsi que pour des transactions. Elles peuvent également être émises par des machines que ce soit des logs d'exécution ou des capteurs.

La Business Intelligence vient traiter les données transactionnelles d'une entreprise, en les structurant, les stockant et les analysants à l'aide d'outils multidimensionnels. Elle se base sur un traitement en quatre étapes : les données sont extraites des sources puis transformées et stockées dans un DWH ou DTM, puis à l'aide de la technologie OLAP, une analyse multidimensionnelle facilite la partie analyse pour visualiser les résultats sur des tableaux de bord ou des rapports. La BI fournit des indicateurs pour piloter l'activité et optimiser les revenus d'une entreprise. Cependant, seules les données structurées venant des BDD relationnelles sont exploitées.

Le Big Data ajoute des méthodes permettant de traiter les données non exploitées par la BI. Les 3V du Big Data apportent des réponses aux problématiques de Volume, Vitesse et Variétés des données en rencontrées par les géants du Web comme Google, Facebook ou Amazon. Ils ont alors participé au développement des technologies permettant d'exploiter toutes les données. Ces technologies sont pour la plupart Open Source, appartiennent aux BDD non relationnelles et actuellement se répartissent autour de l'écosystème Hadoop et du mouvement NoSQL. A la différence des architectures relationnelles traditionnelles qui sont dites centralisées (une seule machine ultraperformante traite les données), les architectures Big Data sont distribuées (beaucoup de machines aux performances faibles traitent de plus petits volumes de données).

Le Big Data n'aborde pas les données comme la BI. En effet, les données sont d'abord collectées de toutes les sources disponibles, puis elles sont explorées pour déterminer les données utiles. Ensuite, ces données sont transformées et structurées pour pouvoir être extraites. De plus, les technologies n'étant pas les mêmes que pour la BI, il est difficile de faire du Big Data avec de la BI. Les deux approches sont donc complémentaires et pouvoir les fusionner représente une opportunité pour créer davantage de valeur.

Deux approches sont à définir avant de se lancer dans l'aventure Big Data. L'approche « Make » permet à l'entreprise de fabriquer sa propre solution via les technologies Open Source. L'approche « Buy » consiste à acheter une solution fonctionnelle pour faire du Big Data. Dans les scénarios étudiés, deux approches Make ont été proposées et une Buy. Les approches Make visaient dans un premier temps à développer une solution Big Data en parallèle des solutions BI existantes, puis les informations venant des architectures BI et Big Data sont échangées grâce à des connecteurs, à différents niveaux des chaines de traitements. Dans un second temps, le deuxième scénario généralise l'approche hybride en passant par un Cloud dédié au Big Data offrant des niveaux de service en interne aux divisions métiers. L'approche Buy étudiée consiste à passer par une appliance, c'est-à-dire un éditeur de solution/BDD comme Oracle, SAP ou Microsoft pour disposer d'une architecture Big Data qui serait non modifiable mais disponible. Dans notre étude, l'accent a été porté sur le scénario 1, avec l'approche « Make » Hybride.

Grâce à une enquête de terrain et une étude documentaire, les besoins ont montré que les problématiques sont surtout liées aux volumes de données disponibles trop faibles et souvent des temps de réponse trop longs. Les compétences nécessaires pour maitriser le Big Data se positionnent autour de trois composantes : technique, analytique et métier. Le Big Data étant encore assez jeune, les interfaces de traitement sont souvent brutes (lignes de commande) et donc inadaptées à des utilisateurs non-initiés. Des personnes capables de traiter les données structurées et non structurées doivent donc avoir une formation type ingénieur en informatique. Pour l'analyse de ces données, des statisticiens et data mineur sont indispensables. Enfin pour comprendre les données et les problèmes de l'entreprise, des experts métiers doivent être sollicités. Ces trois personnes doivent donc collaborer pour créer de la valeur. Dans l'idéal, elles peuvent être une seule et même personne : le Data Scientist.

Il est encore trop tôt pour estimer les ROI sur les projets Big Data. Cependant, une méthode agile facilite l'adoption des technologies. En effet, une approche Test & Learn dans laquelle les employés sont incités à découvrir les concepts, technologies et surtout où le droit à l'erreur est encouragé, assurera une meilleure mise en place. De plus, un travail en parallèle du scénario doit être fait sur les métadonnées, les données de référence ainsi que la gouvernance. Ceci assurera un meilleur contrôle, une qualité supérieure et une définition claire des responsabilités vis-à-vis des données.

Pour aller plus loin, le Big Data apporte deux opportunités, d'une part la valorisation et d'autre part la monétisation. La valorisation des données consiste à tirer profit du Big Data en déterminant quelles données une entreprise possède et/ou pourrait exploiter ; lesquelles seraient potentiellement capables de générer des revenus et enfin d'être au courant de ce qui est faisable avec du Big Data. Les données pouvant être utiles proviennent généralement des sources de données internes mais que l'entreprise n'utilise pas. Par exemple, lors de l'étude des besoins, beaucoup d'informations comme la navigation sur le site internet ou les commentaires sur les forums d'assistance n'étaient pas exploitées et représentaient une perte de valeur. La valorisation continue avec le choix d'une technologie adaptée pour l'exploitation de ces données (JB, 2013). La monétisation intervient comme une source de revenus supplémentaire. Il s'agit de la revente d'informations en accord avec la règlementation de la CNIL. La monétisation existait déjà avant le Big Data, mais les nouvelles données qui peuvent être exploitées apportent une exploitation une valeur pour d'autres sociétés. Par exemple, l'étude du nombre de personnes présentes sur une antenne mobile, permet de revendre aux commerces aux alentours de l'antenne ou aux espaces publicitaires des informations sur qui sont les personnes circulant dans le quartier ou combien ils sont (JB, 2013), (JPL, 2013).

Ces deux opportunités sont des points forts pour les entreprises. Le Big Data nécessite de travailler sur les données des clients pour pouvoir valoriser et monétiser ses informations. Cependant, en se basant sur les données des clients et plus particulièrement en fouillant de plus en plus leurs publications en ligne pour déterminer avec plus de précision ce qu'ils recherchent, les clients peuvent se demander quelles informations les entreprises possèdent sur eux. Le Big Data en exploitant toutes les informations à disposition va se heurter à des contraintes légales propres à la propriété des données et le droit à l'exploitation. Les clients doivent être sensibilisés à la confidentialité des contenus en ligne et les entreprises se devront certainement de suivre une charte d'exploitation de l'information.

TABLE DES FIGURES

TABLE DES TABLEAUX

BIBLIOGRAPHIE

Antoine Crochet-Damais; JDN, 2013. *SFR ouvre ses données clients grâce au Big Data.* [En ligne]
Available at: http://www.journaldunet.com/solutions/dsi/projet-de-big-data-en-france/
[Accès le 11 2013].

Apache, 2005. *Nutch.* [En ligne]
Available at: http://nutch.apache.org/
[Accès le 11 2013].

Apache, 2013. *Hive.* [En ligne]
Available at: http://hive.apache.org/
[Accès le 11 2013].

Apache, 2013. *Pig.* [En ligne]
Available at: http://pig.apache.org/
[Accès le 11 2013].

Apache, 2013. *SQOOP.* [En ligne]
Available at: http://hive.apache.org/
[Accès le 11 2013].

Apache, 2013. *ZooKeeper.* [En ligne]
Available at: http://zookeeper.apache.org/
[Accès le 11 2013].

Aurélien Foucret; SMILE, 2011. *NoSQL une nouvelle approche du stockage et de la manipulation des données,* s.l.: s.n.

CIGREF, 2012. *Les fondamentaux du Cloud computing,* s.l.: s.n.

Data governance Institute, 2004. *Definitions of Data Governance.* [En ligne]
Available at: http://www.datagovernance.com/adg_data_governance_definition.html
[Accès le 11 2013].

DeWitt, D., 2008. *MapReduce: A major step backwards.* [En ligne]
Available at:
http://homes.cs.washington.edu/~billhowe/mapreduce_a_major_step_backwards.html
[Accès le 11 2013].

etechnoforte, 2009. *BI and Analytics.* [En ligne]
Available at:
http://www.etechnoforte.com/products/qlikview.html#/images/stories/gallary_img/qlikview01.jpg
[Accès le 10 2013].

Fernandez , A., 2008. *Qu'est-ce qu'un ERP ? Enterprise Ressource Planning.* [En ligne]
Available at: http://www.piloter.org/techno/ERP/ERP.htm
[Accès le 10 2013].

Fernandez, A., 2008. *CRM Customer Relationship Management qu'est-ce que c'est ?.* [En ligne] Available at: http://www.piloter.org/techno/CRM/index.htm [Accès le 10 2013].

Gartner, 2013. *Big Data.* [En ligne] Available at: http://www.gartner.com/technology/topics/big-data.jsp [Accès le 09 2013].

Goglin, J.-F., 2001. *Construction du datawarehouse.* 2nd éd. s.l.:Hermes Lavoisier.

Google; Sanjay Ghemawat; Howard Gobioff; Shun-Tak Leung, 2001. *The Google File System,* s.l.: s.n.

HARTANI, N., 2012. *Application de l'approche Agile au projet de Business Intelligence,* Grenoble: s.n.

Hurwitz, J., Nugent, A., Halper, D. F. & Kaufman, M., 2013. *Big Data for DUMMIES.* s.l.:John Wiley & Sons, Inc..

Iafrate, F., 2013. *Master Data Management* [Interview] (04 2013).

Inmon, W., 2002. *Building the Data Warehouse.* 3rd éd. s.l.:Robert Ipsen.

JB, 2013. [Interview] (22 10 2013).

JG, 2013. [Interview] (26 11 2013).

JPL, 2013. [Interview] (22 09 2013).

Kimball, R., 2004. *The Data Warehouse ETL Toolkit.* s.l.:Wiley Publishing, Inc.

Krishnan, K., 2013. *MK.Data.Warehousing.in.the.Age.of.Big.Data.May.2013.* 1st éd. s.l.:Morgan Kaufmann; Elsevier.

Laney, Doug; Gartner, 2001. *3D Data Management: Controlling Data Volume, Velocity and Variety,* s.l.: s.n.

Luhn, H. P., 1958. A Business Intelligence System.

Messatfa, H. & IBM, 2013. *Big data.* ESIEA, s.n.

MM, 2013. [Interview] (10 09 2013).

MN, 2013. [Interview] (2 09 2013).

Nielsen, J., 2007. *Fancy Formatting, Fancy Words = Looks Like a Promotion = Ignored.* [En ligne]
Available at: http://www.nngroup.com/articles/fancy-formatting-looks-like-an-ad/ [Accès le 11 2013].

NIST, 2010. *Cloud Computing.* [En ligne] Available at: http://www.nist.gov/itl/ [Accès le 10 2013].

Orange Business Services, 2013. *what can businesses do to capture the full potential of big data?,* s.l.: s.n.

Oxford Dictionnary, 2013. *Oxford Dictionnary.* s.l.:s.n.

Parageaud, C., 2013. *Big Data : La jungle des différentes distributions open source Hadoop.* [En ligne]
Available at: Big Data : La jungle des différentes distributions open source Hadoop
[Accès le 11 2013].

Perriault, N., 2008. *Methodologies de Developpement Agiles : Scrum et XP.* [En ligne]
Available at: http://fr.slideshare.net/nperriault/mthodologies-de-dveloppement-agiles-presentation
[Accès le 10 2013].

Rajkumar Buyya; James Broberg; Andrzej M. Goscinski, 2010. *Cloud Computing: Principles and Paradigms.* s.l.:John Wiley & Sons.

Schmidt, E., 2013. *techonomy.* Lake Tahoe, Califorine, s.n.

Strauch, C., 2011. *NoSQL Databases ,* s.l.: s.n.

w3resource, 2012. *NoSQL.* [En ligne]
Available at: http://www.w3resource.com/mongodb/nosql.php
[Accès le 11 2013].

ANNEXE I GLOSSAIRE

BDD	Base de données	5-10,15,19,23,29-33,38,39,45,49,52,58,64,68,70-75
BI	Business Intelligence	5-11,18-20,22-27, 32,55-61,64-67,70-74,76,79-82
BO	Business Object	58,61
CA	Chiffre d'Affaires	57, 61, 62,73
CCBI	Centre de Compétences BI	58
CDO	Chief Data Officer	81
CRM	Customer Relationship Management	9, 10, 57,58
DAAS	Data As A Service	36
DN	Data Node	51, 42, 44
DTM	Data Mart	13, 15, 19, 23, 57, 58, 66, 67, 76, 80,82
DWH	Data WareHouse	13,15,16,18-23,40,44,45,56-58,6667,70,76,80,82
ERP	Entreprise Ressource Planning	9,10
ETL	Extract Transform and Load	10-13, 20, 22, 57, 58, 64,66
GFS	Google File System	39,40
HDFS	Hadoop Distributed File System	40-44
IAAS	Infrastructure As A Service	36,38 56-59, 62,64
JT	Job Tracker	41, 42,44
MDM	Master Data Management	60, 63, 75,77-79

MOA	Maitrise d'Ouvrage	74
MOE	Maitrise d'Œuvre	74
NN	Name Node	41, 42,44
ODS	Operational Data Store	13,20
OLAP	On-Line Analytical Processing	14, 15,82
OLTP	On-Line Transactional Processing	15
PAAS	Platform As A Service	36,68
POC	Proof Of Concept	74
ROI	Return Of Investment	72, 74,83
SAAS	Software As A Service	36,68
SPOF	Single Point Of Failure	40
SQL	Structured Query Language	6, 9,19
SSD	Solide State Drive	5,23
TDB	Tableau de Bord	61,62
TT	Task Tracker	42,44
XP	EXtrem Programming	25,26

ANNEXE II QUESTIONNAIRE MN ET MM

Section 1 information utilisateur

1) Nom, Prénom et Fonction
2) Connaissance du Big Data ?
    a. Description par l'utilisateur
    b. Explication du Big Data

Section 2 Sources

1) Quelles sont les sources de données que vous utilisez et pour quels buts ?
2) Il y a-t-il des données non structurées ?
3) De quels ordres sont les volumes extraits
4) Comment sont les temps de réponse

Section 3 Besoins actuels

1) Quels sont les besoins
2) Il y a-t-il des besoins satisfaits par d'autres sources ?
3) Les volumes de données fournis sont-ils suffisant ?
4) Les temps de réponse sont-ils satisfaisants ?
5) Il y a-t-il des besoins ne pouvant être satisfaits par vos sources ?
    a. Si oui pourquoi ?

Section 4 Besoins liés au Big Data

1) Certains types de données pourraient-ils être utiles ?
    a. Photo ?
    b. Vidéo ?
    c. Son ?
    d. Texte ?
    e. Réseaux sociaux ?
    f. Autres ?
2) Connaissez-vous des sources disposant de ces types de données ?
3) Les problématiques rencontrées pour exploiter les données sont-elles d'ordre
    a. Volume ?
    b. Vitesse ?
    c. Variété ?
    d. % ?
4) Quels seraient les gains générés s'il était possible d'exploiter ces données
5) Quels seraient les gains générés s'il était possible de réduire les problématiques rencontrées ?

ANNEXE III QUESTIONNAIRE JPL

Le vendredi 13 septembre durée 2h.

Introduction présentation du sujet de thèse et des questions de recherches.

Q1 En quoi consiste le projet ?

Q2 Le Cloud sera-t-il privé ou public ?

Q3 Quelles démarches et compétences sont nécessaires

Q4 les compétences à voir

Q5 De quelles données disposez-vous pour les études Big data