Synthèse des méthodes de conduite de projets Big Data et des retours collectés lors de pilotes industriels

Christophe Ponsard [1], **Mounir Touzani** [2], **Annick Majchrowski** [1]

1. *CETIC - Centre de recherche, Gosselies, Belgique*
   {christophe.ponsard,annick.majchrowski}@cetic.be
2. *Docteur en Informatique, Toulouse, France*
   mtouzani64@gmail.com

RÉSUMÉ. *Un facteur déterminant du bon fonctionnement de l'entreprise, est sa capacité à traiter d'importantes quantités de données, désignées par l'appellation Big Data ou données massives. Nombre d'entreprises peinent à déployer des solutions techniques pour pallier cette problématique, en raison d'un manque de maturité dans la gestion de tels projets. Afin de les guider, cet article propose une série de briques méthodologiques identifiées dans la littérature depuis les méthodes de la fouille de données jusqu'à nos jours. Il est complété par des retours d'expériences issus de pilotes menés dans quatre domaines clefs (IT, santé, spatial, agroalimentaire). Ceux-ci donnent une vision concrète de la mise en œuvre des étapes de compréhension des besoins et des données en mettant l'accent sur l'identification de la valeur, la définition d'une stratégie adéquate et une démarche de type agile pour gérer la montée en maturité.*

ABSTRACT. *Companies are increasingly faced with the challenge of handling increasing and even massive amounts of digital data. Although Big Data technical solutions are available, many companies are struggling to deploy them because of a lack of maturity related to their management. This paper aims at improving guidance in this area based on a series of methodological brick documented in the literature from data mining projects to nowadays. It is complemented by lessons learned from pilots conducted in four key areas (IT, health, space, food industry). They give a concrete vision of how to implement the requirements gathering and data understanding steps with a focus on the identification of value, the definition of a relevant strategy and an agile follow-up to also manage the rise in maturity.*

MOTS-CLÉS : *gestion de projet, processus d'adoption, méthodes agiles, modélisation des données, Big Data, données massives, étude de cas.*

KEYWORDS: *projet management, adoption process, agile Methods, Big Data, case study.*

DOI:10.3166/ISI.23.1.9-33

1. Introduction

Notre monde est confronté à une explosion de l'information résultant d'une interconnexion globale et permanente des personnes mais également d'une croissance très rapide des dispositifs connectés via l'Internet des Objets (*en anglais Internet of Things, ou IoT*). De nombreuses statistiques attestent de ce phénomène : ainsi, 90 % des données stockées dans le monde ont été produites durant ces deux dernières années. Le volume des données créées par les entreprises, double tous les 1,2 année (Rot, 2015). D'ici 2020, plus de 40 zettaoctets (10^{21} octets) auront été produits.

Les organisations perçoivent bien le grand potentiel que les technologies Big Data peuvent leur apporter pour améliorer leurs performances et, dans le cas des entreprises, accroître leur avantage compétitif. Le mode devenu plus accessible, voire plus facile pour la collecte et le stockage des données, a incité un certain nombre d'entre elles à démarrer des projets Big Data, d'autant plus que des outils technologiques de stockage et d'analyse à grande échelle (notamment les bases de données NoSQL) sont de plus en plus accessibles.

Les caractéristiques et les défis posés par le Big Data sont souvent présentés au moyen d'une série de mots en « V » au Volume déjà mentionné, s'ajoutent notamment la Variété (diversité de formats structurés ou non), la Vélocité (aspect temps-réel du traitement), la Véracité (qualité des données), la Visualisation (facilité d'interprétation) et la Valeur (pour quel bénéfice ?) (Mauro *et al.*, 2016).

Cependant, le constat est que la plupart des organisations ne parviennent toujours pas à obtenir le dernier « V », c'est-à-dire produire une réelle valeur ajoutée à partir de leurs données. Un rapport de 2013 portant sur 300 entreprises Big Data, a révélé que 55 % des projets Big Data se sont terminés prématurément et que beaucoup d'autres, n'ont que partiellement atteint leurs objectifs (Kelly, Kaskade, 2013).

Ceci est confirmé par une étude en ligne conduite par Gartner en juillet 2016, qui a rapporté que de nombreuses entreprises restent bloquées au stade du projet pilote et que seulement 15% des projets Big Data ont été effectivement déployés en production (Gartner, 2016). En examinant la cause de tels échecs, il apparaît que le facteur principal n'est en réalité pas lié à la dimension technique, mais plutôt aux processus et aux aspects humains qui s'avèrent être aussi importants (Gao *et al.*, 2015). Une étude récente de la littérature montre que de nombreux articles se concentrent encore énormément sur la dimension technique, en particulier l'utilisation d'algorithmes qui permettent de réaliser des analyses approfondies, et que beaucoup moins d'attention est portée sur les méthodes et les outils qui pourraient aider les équipes à mener efficacement des projets Big Data à terme (Saltz, Shamshurin, 2016).

Il existe toutefois quelques travaux récents dans ce domaine, notamment en matière d'identification des facteurs clés de succès des projets Big Data (Saltz, 2015), aussi bien sur des problèmes de gestion de projets (Corea, 2016) que sur la manière dont les équipes s'organisent pour réaliser des projets Big Data, en pointant cependant l'absence de standard en la matière (Saltz, Shamshurin, 2016).

Notre article se situe dans la lignée de ces travaux et a pour objectif d'apporter des recommandations concrètes aux entreprises engagées dans un processus d'adoption de solutions Big Data. Par ce travail, nous souhaitons apporter quelques éléments de réponse à des questions telles que :

– Comment pouvons-nous être sûrs que le Big Data pourrait nous aider ?
– Quelles personnes devraient être impliquées et à quel moment ?
– Quelles sont les étapes clefs auxquelles il faut être attentif ?
– Est-ce que mon projet est sur la bonne trajectoire pour aboutir ?

Notre contribution se veut de nature pratique et s'appuie sur un ensemble de projets pilotes couvrant différents domaines (sciences de la vie, santé, espace, infrastructures informatiques). Ces pilotes sont répartis sur deux ans et sont réalisés dans le cadre d'un projet global commun, réalisé en Belgique. Le processus suivi est similaire et renforcé progressivement. Les travaux rapportés sont basés sur les quatre premiers pilotes et étend significativement notre travail précédent basé sur deux pilotes et un aperçu plus limité de la littérature (Ponsard *et al*., 2017). Il développe aussi la manière d'associer des indicateurs clefs de performance ainsi que les stratégies d'analyse de données à mettre en œuvre.

Le présent document est structuré comme suit. La section 2 donne une typologie des principales catégories de projets Big Data. La section 3 passe en revue les principales méthodologies concernant le déploiement du Big Data. Dans la section 4, nous présentons la méthodologie suivie pour mettre en lumière nos projets pilotes et dégager une guidance méthodologique. Ensuite, nous mettons l'accent sur les facteurs clés de succès du déploiement d'une solution Big Data. La section 5 propose un retour d'expérience en donnant des recommandations qui ciblent des étapes particulièrement importantes. La section 6 est une discussion qui met nos travaux en perspective avec d'autres approches actuellement proposées. Enfin, la section 7 tire quelques conclusions et extensions que nous envisageons de mener dans la suite de nos projets pilotes.

2. Typologie des méthodes d'analyse de données massives

L'analyse de données (*Data Analytics*) est un concept multidisciplinaire qui peut être défini comme tout moyen permettant d'acquérir des données depuis des sources diverses, de les traiter afin de découvrir des relations qui les relient et mettre des résultats à disposition des parties prenantes (Chen *et al*., 2012). L'application de ces techniques par des entreprises leur permet de mieux comprendre leur niveau de performance et de procéder à des améliorations de leur métier (*Business Analytics*). Trois catégories complémentaires d'analyse peuvent être distinguées et combinées pour atteindre les objectifs de compréhension des données et d'aide à la décision.

– *L'analyse descriptive* permet d'investiguer le passé afin de répondre à la question « Que s'est-il passé ? » Elle repose sur un ensemble de techniques permettant d'examiner les données pour comprendre et analyser les performances de l'entreprise. Il s'agit notamment de l'analyse statistique ainsi que de méthodes de classification et

de catégorisation. Elle comprend également le diagnostic pour répondre à la question : « Pourquoi est-ce arrivé ? », afin de comprendre les raisons des événements qui se sont produits dans le passé.

– *L'analyse prédictive* est tournée vers l'avenir et essaie de répondre aux questions « Que va-t-il se passer ? » et « Pourquoi cela risque-t-il de se produire ? » Elle utilise un ensemble de techniques d'analyse des données actuelles et passées pour découvrir ce qui est le plus susceptible de se produire (ou non). Les approches utilisées ici sont principalement basées sur des techniques statistiques (modèles de régression ou à choix discret) et l'apprentissage automatique (*Machine Learning*), notamment via des réseaux de neurones artificiels.

– *L'analyse prescriptive* examine également l'avenir, mais permet de mettre l'accent sur les recommandations et conseils afin de répondre aux questions « Que dois-je faire ? » et « Pourquoi devrais-je le faire ? » Les techniques spécifiques qui sont utiles ici, relèvent de l'optimisation, de la simulation du comportement futur, de systèmes de règles métier voire de systèmes experts permettant de proposer des actions contre les risques connus ou identifiés via l'analyse prédictive.

3. Revue des méthodes et processus existants

Cette section passe en revue les méthodes et processus existants pour la mise en œuvre de projets Big Data. Elle souligne certaines forces et limitations connues. Nous commençons par présenter des méthodes héritées du domaine de la fouille de données (*Data Mining ou DM*) et de l'informatique décisionnelle (*Business Intelligence ou BI*) avant d'envisager des approches plus spécifiques au Big Data avec une attention particulière aux méthodes agiles. Enfin, certaines méthodes complémentaires inspirées d'approches plus cognitives ou de gestion de la maturité seront également envisagées.

3.1. Méthodes liées à la fouille de données et l'informatique décisionnelle

La fouille de données a été développée dans le courant des années 1990 avec pour objectif d'extraire des données à partir d'informations structurées (bases de données relationnelles) pour découvrir des facteurs clés de l'entreprise à une échelle relativement petite. Elle s'appuie en particulier sur des méthodes statistiques (Vaillancourt, 2010). Le Big Data, quant à lui, opère sur une grande échelle : données structurées, semi-structurées (XML, JSON) et non structurées, et vise à dégager des indicateurs à vocation prédictive. Cependant, un point est commun aux deux types d'approches en termes de processus. Il est donc nécessaire de mettre en place une coopération étroite entre les experts techniques (données) et les experts métiers (Hoppen, 2015). De nombreuses méthodologies et modèles de processus ont été développés pour la fouille de données et la découverte de connaissances (Mariscal *et al.*, 2010).

L'informatique décisionnelle s'est également développée dans les années 1990 et vise essentiellement à produire des indicateurs clés de performance (en anglais, KPI : *Key Performance Indicator*) sous forme de tableaux de bord. Les techniques s'ap-

puient sur des données structurées et ne nécessitent que peu d'intelligence dans les traitements. Le Big Data permet d'élargir le champ de la BI aux données moins structurées. Inversement, la BI apparaît comme un prérequis permettant de mesurer précisément ce qu'on désire améliorer tandis que les techniques de Big Data apportent des possibilités d'analyse prédictive (Halper, 2014).

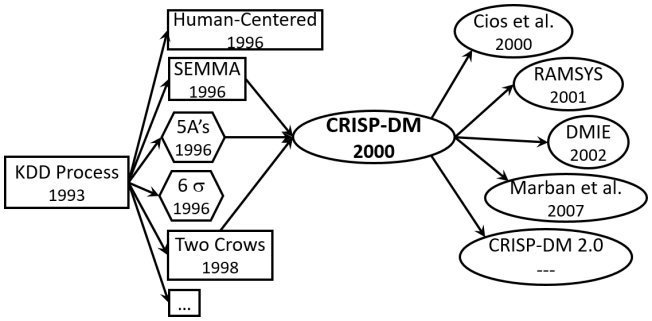

Figure 1. Evolution des méthodologies de traitement de données, adapté de (Mariscal et al.*, 2010)*

La figure 1 (Mariscal *et al*., 2010) illustre l'évolution des méthodologies de traitement de données. KDD (*Knowledge Discovery in Database*) est l'approche séminale en matière de fouille de données. Elle a été raffinée en plusieurs autres approches (SEMMA, Two Crows, etc.) avant d'être standardisée par CRISP-DM (*Cross Industry Standard Process for Data Mining*, ou processus standard pour la fouille de données, en français) (Shearer, 2000). Cette méthode est décrite dans la figure 2. Elle est composée de six phases, chacune étant décomposée en sous-étapes. Le processus n'est pas linéaire mais plutôt organisé en un cycle global, avec généralement des revues entre les phases. CRISP-DM est une méthode très utilisée depuis 20 ans, non seulement pour la fouille de données mais aussi pour l'analyse prédictive et des projets Big Data.

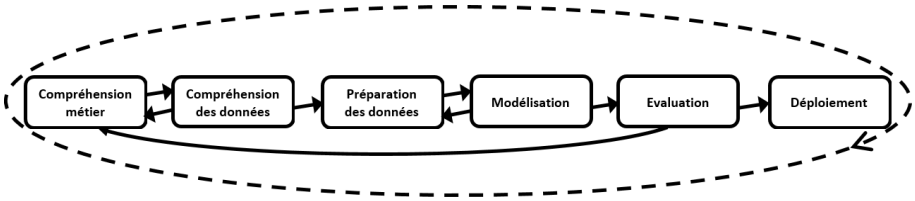

Figure 2. La méthode CRISP-DM, adapté de (Shearer, 2000)

CRISP-DM ne couvre pas la partie infrastructure/opérations pour la mise en œuvre d'un projet d'exploration de données/prédiction analytique. Cette méthode a très peu d'activités et de tâches de gestion de projet. Elle est légère sur les activités et les tâches dans la phase « Déploiement », et n'a pas de modèles ou de lignes directrices. Les méthodes liées à CRISP-DM souffrent toutefois des problèmes suivants :

– une visibilité limitée du management sur la communication ainsi qu'au niveau de la connaissance et sur les aspects projet.

– un manque de maturité au niveau du modèle pour mettre en évidence des étapes et des jalons plus importants qui peuvent être améliorés progressivement.

– un modèle itératif limité : les itérations prévues sont peu utilisées en pratique car elles n'impliquent pas l'aspect métier mais plutôt le contexte IT interne. Outre l'absence de contrôle sur la valeur produite, elles ont en plus tendance à être sans cesse reportées. C'est à cette fin que des modèles plus agiles, présentés à la section suivante, ont été introduits.

3.2. Vers plus d'agilité

Les méthodes agiles sont des méthodes itératives qui répondent au manifeste agile dont les principes mettent une interaction avec le client, l'adaptation aux changements et la production de valeur au centre du processus de développement (Alliance, 2001). Initialement mis en place pour le développement de logiciels, ces principes peuvent également répondre plus largement et en particulier à l'analyse des données afin de fournir une meilleure guidance et aboutir à la production de valeur. Une évolution de KDD et CRISP-DM vers l'agilité est assurée par la méthode AgileKDD (Nascimento, Oliveira, 2012). Celle-ci est basée sur le cycle de vie OpenUP qui répond aux principes du Manifeste Agile (Balduino, 2007). Les projets sont divisés en « sprints » planifiés avec des délais fixes, habituellement de quelques semaines. A l'issue de chaque sprint, les équipes doivent démontrer qu'elles sont parvenues à produire des résultats valorisables.

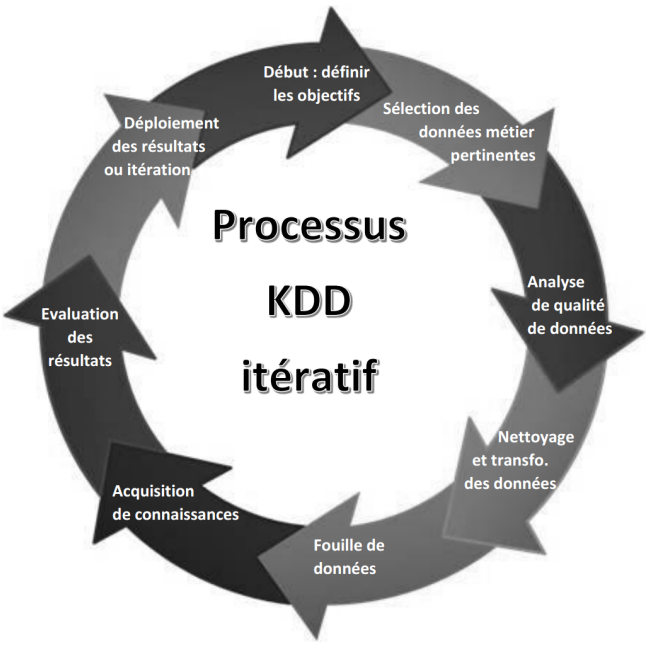

Figure 3. La méthode AgileKDD, adapté de (Nascimento, Oliveira, 2012)

Bien que les méthodes agiles semblent en adéquation avec les besoins, le déploiement de telles méthodes pour le Big Data peut se heurter à une résistance. C'est en particulier le cas dans les organisations de plus grande taille qui sont habituées à des processus assez rigides, plus aisés à planifier. Une enquête a révélé que, comme pour le logiciel, les entreprises ont tendance à accepter des méthodes agiles pour les projets Big Data de plus petite envergure, moins complexes et ayant peu d'exigences en matière de sécurité. Il s'agit aussi généralement d'organisations plus flexibles. En dehors de ces cas, l'approche préférée reste l'approche planifiée (Franková *et al*., 2016).

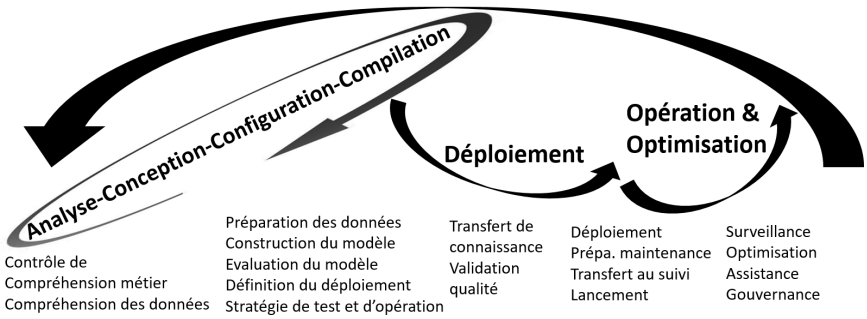

Figure 4. Méthode ASUM-DM, adaptée de (IBM, 2015)

IBM a également développé ASUM-DM, une extension et un raffinement de CRISP-DM qui combine la gestion de projets traditionnelle mais aussi des principes d'agilité (IBM, 2015). La figure 4 illustre les grands blocs et son principe itératif conduit par des activités spécifiques au niveau des dernières colonnes. Celles-ci incluent à la fois des notions de gouvernance et d'alignement avec la communauté.

3.3. Méthodes spécifiques pour le Big Data

La méthode AABA (Architecture-centric Agile Big data Analytics) répond aux défis techniques et organisationnels du Big Data (Chen *et al*., 2016). La méthode intègre à la fois la conception du système Big Data (BDD) et une architecture AAA (*Architecture-centric Agile Analytics*). Elle est centrée sur le modèle DevOps et orientée vers la découverte efficace et la livraison continue de valeur.

La méthode, illustrée à la figure 5, a été validée sur 11 études de cas dans différents domaines notamment dans les domaines du marketing, les télécommunications et la santé. Sur cette base, elle a émis les recommandations suivantes :

1) les analystes et experts en données doivent être impliqués tôt dans le processus ;
2) un soutien continu aux activités d'architecture est nécessaire ;
3) la définition d'une architecture de référence permet une plus grande flexibilité ;
4) les boucles de rétroaction permettent de traiter les exigences non fonctionnelles telles que la performance, la disponibilité et la sécurité ;
5) l'agilité permet aussi de gérer l'évolution rapide des technologies et des besoins.

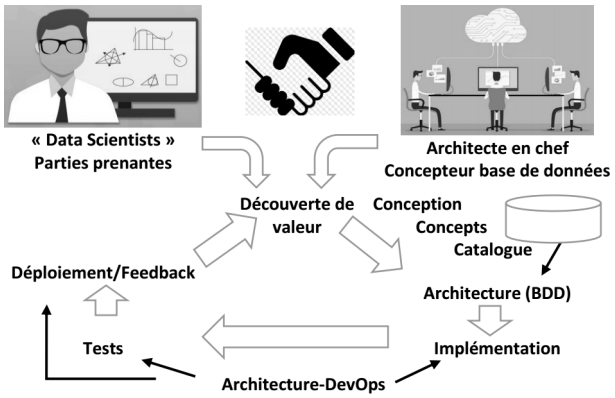

Figure 5. La méthode AABA, adapté de (Chen et al.*, 2016)*

Parallèlement, *Stampede* est une méthode proposée par IBM à ses clients. Son principal objectif est d'aider les entreprises à démarrer rapidement en validant le potentiel de génération de valeur à partir du Big Data. La méthode s'appuie sur la mise à disposition de ressources d'experts dans le cadre d'un projet pilote bien défini (IBM, 2013). La méthode s'appuie notamment sur un atelier d'une demi-journée permettant de définir le projet Big Data, identifier l'infrastructure nécessaire, établir un plan de travail mais surtout et avant tout établir la valeur pour l'entreprise. L'exécution du pilote est généralement répartie sur douze semaines et réalisée de manière agile avec un jalon important vers la neuvième semaine.

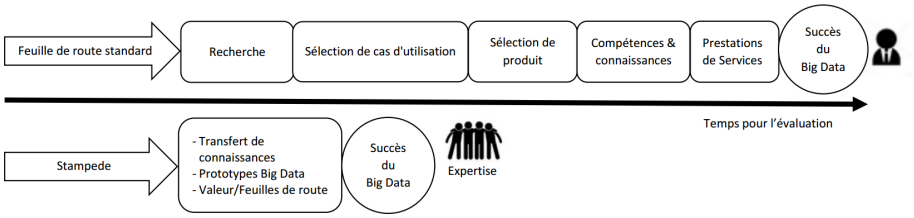

Figure 6. La méthode Stampede d'IBM, adapté de (IBM, 2013)

3.4. Vers des modèles de maturité du Big Data

Des solutions ont été également proposées afin de définir un *modèle de maturité de capacité* (*Capability Maturity Model* ou CMM) pour les processus de traitement des données dans le but de soutenir l'évaluation et l'amélioration de ces processus (Crowston, 2010) (Nott, 2014). Un tel modèle décrit les pratiques nécessaires à une gestion efficace en les organisant par paliers progressifs. Une échelle typique sur 5 niveaux est utilisée à la fois dans (Crowston, 2010) et (Nott, 2014). Le premier utilise les niveaux standards allant de « défini » à « optimisé » tandis que le second utilise une nomenclature plus spécifique allant de « ad hoc » à « breakaway ». Le tableau 1

détaille les principaux critères qui concernent la place de la donnée dans la stratégie métier, le type d'analyse utilisée pour les données, l'alignement de l'infrastructure IT ainsi que les aspects culture et gouvernance.

Tableau 1. Le modèle de maturité de (Nott, 2014)

Niveau	Ad hoc	Fondateur	Compétitif	Différentiateur	Libérateur (Breakaway)
Stratégie métier	Utilisation de reporting standard. Big Data juste évoqué	Identification d'un ROI lié aux données	Exploitation des données encouragée	Réalisation d'un avantage compétitif	Innovation métier conduite par les données
Analyse de données	Limité au passé	Détection d'événements	Prédiction de certaines probabilités d'évolution	Optimisation des décisions	Optimisation et automatisation possible de certains processus
Alignement IT	Pas d'architecture cohérente ni unifiée	Framework architectural présent mais adapté au Big Data	Définition de patrons architecturaux pour le Big Data	Architecture définie et standardisée pour la plupart des "V"	Architecture totalement alignée avec les besoins Big Data
Culture et gouvernance	Largement basé sur des individualités	Gestion fragmentaire, résistance au changement	Définition de politiques et de procédures, adoption partielle	Adoption large, utilisation quotidienne	Adoption et mise en œuvre généralisée

3.5. Approches complémentaires

De nombreux facteurs clés de succès, de guides pratiques et de listes de contrôle des risques ont été également publiés, principalement dans les blogs pour les directeurs des systèmes d'information, par exemple (Bedos, 2015). Une classification systématique des facteurs critiques de succès a été proposée par (Gao *et al*., 2015) en utilisant trois dimensions clés : les personnes, les processus et la technologie. Celle-ci a été étendue ensuite par (Saltz, Shamshurin, 2016) pour traiter aussi les dimensions outillage et gouvernance. Voici les principaux facteurs clés :

– pour les données : la qualité, la sécurité, le niveau de structure des données ;

– pour la gouvernance : une direction, une organisation bien définie, une culture axée sur les données ;

– pour les objectifs : la valeur de l'entreprise identifiée (KPI), la rentabilité, une taille de projet réaliste ;

– pour les processus : l'agilité, la conduite du changement, la maturité, la volumétrie des données ;

– pour l'équipe : des compétences en ingénierie des données, la pluridisciplinarité ;

– pour les outils : des infrastructures informatiques, le stockage, la capacité de visualisation des données, le suivi des performances.

4. Processus global suivi pour développer et valider la méthode

4.1. Aperçu du processus global

L'objectif global de notre projet est d'élaborer une méthode systématique pour aider les entreprises à valider les avantages potentiels d'une solution Big Data. Le processus global est représenté dans la figure 7.

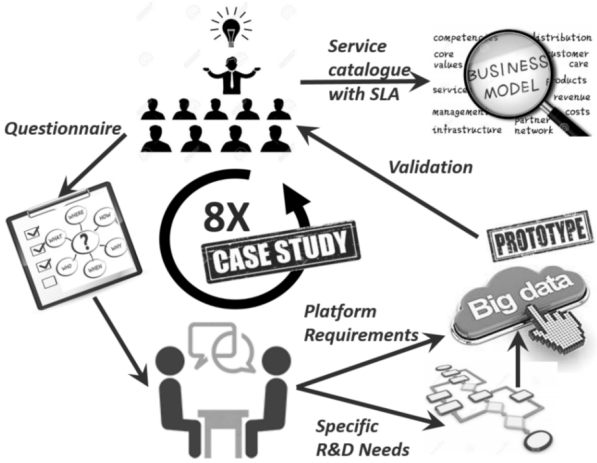

Figure 7. Développement itératif de la méthode et de l'infrastructure

La méthode proposée ici n'est pas totalement générique mais représente plutôt un canevas qui s'est dégagé lors des pilotes successifs afin de fournir un service reproductible de manière fiable aux entreprises ayant des besoins en traitement de données massives. Elle s'appuie sur les méthodes et processus décrits dans la section 3, plus particulièrement :

– le point de départ a été Stampede grâce à la plateforme d'IBM. Voici les principaux aspects retenus à partir des méthodes : l'atelier initial avec toutes les parties prenantes, la focalisation réaliste et un moteur de la valeur d'entreprise constant ;

– pour faire face à un manque de matériel de référence, nous avons défini un modèle de processus basé sur CRISP-DM, largement documenté ;

– les pilotes sont exécutés de manière agile, étant donné les disponibilités des experts (chercheurs universitaires) et planifiés sur des périodes plus longues que dans Stampede, soit 3-6 mois au lieu de 12-16 semaines. L'approche populaire SCRUM a été également utilisée car elle met l'accent sur la collaboration, le fonctionnement du logiciel, l'autogestion de l'équipe et la flexibilité pour s'adapter aux réalités de l'entreprise (Scrum Alliance, 2016).

La particularité de la méthode est abordée à des cas spécifiques dans le cadre de la section 5 et illustrée sur plusieurs de nos pilotes.

4.2. Caractérisation des projets pilotes

Les différents pilotes sont gardés confidentiels. Le tableau 2 en donne néanmoins les principales caractéristiques exprimées notamment au moyen des 3 premiers "V" du Big Data ainsi que de la typologie.

Tableau 2. Principales caractéristiques des 4 premiers projets pilotes

#	Domaine	Volume	Vélocité	Variété	Caractérisation
1	Maintenance IT	Environ 3000 serveurs	Haute (événements, logs, etc.)	Temps réel	Analyse prédictive pour maîtriser les coûts de maintenance
2	Santé	900 lits sur 3 sites	Temps réel	Nombreuses sources, formats divers	Analyse prédictive et prescriptive pour réduire la morbidité. Confidentialité.
3	Spatial	Maintenance infrastructure sol Galileo	Moyenne	Haute : messages, logs	Maintenance prédictive de matériel coûteux. Fiabilité 99,8 %
4	Agroalimentaire	20 Go/analyse, 2 To/semaine	Haute (à paralléliser)	Données métier et traçabilité (ex. agroalimentaire)	Essentiellement descriptive, au niveau de la qualité des produits

4.3. Schéma général appliqué au sein de chaque pilote

La méthodologie qui a émergé sur base des méthodologies existantes et des itérations sur nos 4 pilotes se compose des trois phases suivantes :

Phase 1. Contexte et sensibilisation au Big Data. Dans cette phase d'introduction, une ou plusieurs réunions sont organisées avec l'organisation participative. Une introduction générale est donnée sur les concepts du Big Data. La plateforme mise à disposition est présentée, de même que quelques applications représentatives dans différents domaines (éventuellement avec un focus sur le domaine de l'organisation). Les principaux défis et les étapes clés de la mise en œuvre sont également exposés. Lors des interactions, le niveau de maturité du client et certains facteurs de risque peuvent déjà être vérifiés (par exemple, l'implication de la direction, le niveau d'expertise interne, la formulation d'objectifs assez clairs).

Phase 2. Compréhension du métier de l'entreprise et du cas d'utilisation. Cette phase est largement alignée sur la première phase de CRISP-DM présentée à la section 3.1. Son objectif est d'identifier les besoins et problèmes pour lesquels une solution de type Big Data est envisagée. Il est aussi important de formuler un ou plusieurs cas d'utilisation qui peuvent démontrer l'apport de valeur à partir des données collectées et traitées. Il s'agit d'une phase très importante qui peut être soutenue par des activités spécifiques illustrées par le tableau 3 et détaillées dans la section 5.

Tableau 3. Liste des contrôles pour les ateliers, adapté de (IBM, 2013))

Compréhension du métier (*Business Understanding*)	**Compréhension des cas d'utilisation** (*Use Case Understanding*)
Stratégie et positionnement	**Évaluation de la maturité**
Stratégie globale	
Positionnement des produits/services	**Identification des objectifs**
Tableau de bord - Indicateurs (KPI)	**des cas d'utilisation**
Stratégie digitale	Valeur pour le client
Points de contact clients/prospects	Critères de succès métier
Recherche, e-Commerce,	
Médias sociaux, sites web,...	**Évaluation de la situation**
Compétiteurs directs	Exigences sur les ressources
Élément disruptifs	Hypothèses/contraintes
Modèles et technologies disruptifs	Risques et contingences
Comportements disruptifs	Coûts et bénéfices
Présélection de cas d'utilisation	**Raffinement de la fouille de données**
Objectifs	Buts de fouille de données
Priorités	Indicateurs pour la fouille de données
Coûts, ROI, Contraintes	
Valeur pour le client	**Production du plan projet**
Portée des cas d'utilisation	Approche du projet
Nouvelles sources de données	Livrables
	Planification
Faisabilité de haut niveau	Mitigation des risques (vie privée,etc.)
Buts de la fouille de données et indicateurs	Parties prenantes à impliquer
Disponibilité des ressources	Évaluation initiale des outils et techniques

Phase 3. Mise en œuvre d'un pilote pour un service ou un produit. Dans cette phase, les activités suivantes sont menées de manière agile :

– *Compréhension des données* : analyser les données pour en extraire les sous-ensembles les plus intéressants et assurer une bonne qualité des données.

– *Préparation des données* : sélectionner les données pertinentes, les nettoyer, les étendre et les formater selon les besoins.

– *Modélisation* : sélectionner une technique de modélisation spécifique (par exemple, arbre de décision ou réseaux de neurones). Le modèle est alors construit puis testé au niveau de sa précision et sa généralité (mais pas encore en relation avec les besoins de l'entreprise). Le respect des hypothèses de modélisation est également vérifié. A partir des résultats, les paramètres du modèle peuvent être revus ou d'autres techniques complémentaires peuvent être utilisées.

– *Évaluation* : évaluer dans quelle mesure le modèle répond aux objectifs de l'entreprise, en utilisant des données réelles et réalistes.

– *Déploiement* : transférer la solution validée en environnement de production et veiller à son utilisabilité au moyen d'outils de visualisation et d'un tableau de bord. Les activités de surveillance de performance/précision sont également mises en place.

5. Retour d'expérience et recommandations

Dans cette section, nous présentons quelques notions et des lignes directrices utiles pour appuyer l'ensemble du processus et augmenter les chances de succès. Nous illustrons également nos retours d'expérience sur la base de quelques types de pilotes tirés de 4 exemples : la maintenance informatique, les itinéraires cliniques, la maintenance dans le domaine spatial et l'agroalimentaire.

5.1. Définition d'objectifs progressifs dont la valeur est mesurable

Grâce au déploiement d'une solution Big Data, une entreprise s'attend à valoriser ses données. Les objectifs métiers doivent être clairement identifiés. Dans nos projets pilotes, nous avons intégré des techniques d'ingénierie des exigences orientées buts afin d'élucider, de structurer les objectifs de l'entreprise et de les relier ainsi aux processus et composants de traitement des données (Lamsweerde, 2009). De plus, ces méthodes incluent des techniques spécifiques qui permettent de vérifier que les buts sont complets, consistants, réalisables et robustes. Pour ce dernier point, il est possible d'analyser la présence d'obstacles et de les résoudre.

Une autre façon de relier les buts à la réalité consiste à définir la manière de mesurer la valeur qui devrait être définie dès la phase de compréhension des métiers, en s'appuyant généralement sur des indicateurs clés de performance. Les entreprises devraient déjà avoir défini leurs KPI et être en mesure de les mesurer. Si ce n'est pas le cas, elles devraient commencer à les améliorer, en d'autres termes, l'informatique décisionnelle (*Business Intelligence* ou BI) devrait déjà être présente dans les entreprises.

Sur cette base, différentes stratégies d'amélioration peuvent être identifiées et discutées pour sélectionner une bonne analyse. Dans ce processus, l'écart avec la situation actuelle doit également être pris en compte. Il est plus prudent de garder un premier projet avec des objectifs assez modestes que de risquer d'échouer en visant un projet trop complexe et qui serait sensé apporter plus de valeur. Une fois le projet pilote réussi, d'autres améliorations peuvent être planifiées afin d'apporter plus de valeur.

5.1.1. Étude de cas 1 - Maintenance informatique.

Le fournisseur de services informatiques considéré ici gère plus de 3000 serveurs hébergeant de nombreux sites Web, exécutant plusieurs applications et stockant une grande quantité de données clients. Actuellement, des procédures standards de gestion des incidents et de maintenance préventive sont appliquées. Cependant, les équipements IT (équipements réseaux, serveurs, disques) présentent toujours des risques de pannes, en particulier à des moments inattendus et coûteux à gérer, comme la nuit ou le week-end.

Afin de réduire le nombre d'événements réactifs coûteux et d'optimiser la maintenance préventive, l'entreprise souhaite développer une maintenance plus prédictive

en anticipant l'indisponibilité des serveurs afin qu'elle puisse réagir préventivement, et finalement, prévenir une telle indisponibilité. Dans le processus, le client peut diagnostiquer finement les causes des incidents et les résoudre afin d'en éviter d'autres en mode réactif pouvant se révéler cauchemardesques. L'objectif ultime est d'augmenter la disponibilité des services, la satisfaction des clients et réduire les coûts d'exploitation.

Le KPI résultant est appelé coût total de possession (*Total Cost of Ownership* ou TCO) et les coûts de pannes typiques à considérer peuvent être :

– la maintenance matérielle et logicielle pourrait être réduite si les prévisions réduisent le temps d'intervention grâce à une meilleure prédiction ;
– le travail effectué du personnel sur ces incidents ;
– toute pénalité liée aux SLAs (*Service Level Agreements*) des clients ;
– les effets indirects sur l'activité du client et son image de marque.

5.1.2. Étude de cas 2 - Itinéraires cliniques.

Les hôpitaux déploient de plus en plus des itinéraires cliniques, définis comme une vision pluridisciplinaire du processus de traitement requis pour un groupe de patients présentant la même pathologie, avec un suivi clinique prévisible (Campbell *et al.*, 1998). La raison est non seulement de réduire la variabilité des processus cliniques, mais aussi d'améliorer la qualité et les coûts de contrôle (Dam, 2013). Cela permet une analyse plus riche des données produites et donc le profilage des patients présentant des risques plus élevés, par exemple, en raison de multi-pathologie ou d'intolérances.

Un processus type de chimiothérapie est une séquence d'administration de traitements, généralement en hôpital de jour. Chaque cure est suivie d'une période de repos à la maison qui dure de quelques jours à plusieurs semaines. Un intervalle minimal entre les traitements est nécessaire parce que les médicaments de chimiothérapie sont toxiques et que le corps a besoin de temps pour se rétablir entre deux livraisons de doses. En suivant le protocole de traitement idéal, le nombre de cellules cancéreuses diminue progressivement, en espérant atteindre une guérison complète ou une rémission du cancer. Si les traitements de chimiothérapie, pour une raison quelconque, ne suivent pas de près la périodicité conseillée, ou si les doses sont significativement réduites, l'efficacité du traitement peut être insuffisante. Dans de telles conditions, les cellules cancéreuses peuvent se multiplier à nouveau, ce qui peut entraîner une rechute du cancer.

Afin de mesurer la qualité des soins, un indicateur quantifiable appelé « *Intensité de Dose Relative* » (*Relative Dose Intensity* ou RDI) (Lyman, 2009) a été défini. Il tient compte à la fois du fait que la dose requise est administrée et du moment de la délivrance, sur une échelle allant de 0 % (pas de traitement) à 100 % (conformité totale).

$$RDI = \frac{\text{dose planifiée}}{\text{dose delivrée}} \times \frac{\text{durée réelle}}{\text{durée planifiée}}$$

La littérature médicale a montré, pour un certain nombre de cancers, que la survie sans récidive est fortement corrélée au RDI. Par exemple, pour le cancer du sein, une valeur clé du seuil est de 85 % (Piccart *et al*., 2000). Par conséquent, cet indicateur peut être vu comme une jauge qui doit être soigneusement gérée à travers l'ensemble du cheminement clinique.

5.1.3. Étude de cas 3 - Maintenance dans le domaine spatial

La communication par satellite est un chaînon crucial dans les télécommunications. Toute indisponibilité engendre des impacts significatifs en termes économiques ou sur la sûreté de fonctionnement de certains systèmes. Les possibilités d'interventions sur satellite sont très limitées, par contre une maintenance corrective est possible au niveau des stations au sol. Cependant, cette maintenance engendre des interruptions de services qui sont très coûteuses en main-d'œuvre et en logistique d'acheminement des pièces de rechange et de matériaux, en particulier pour les sites distants. Par ailleurs, vu le caractère critique, toute interruption de service induit des pénalités qui sont spécifiées dans des clauses contractuelles très strictes relatives à la qualité de service (SLA).

L'objectif est de réduire les coûts de maintenance et de respecter le niveau de disponibilité attendu. Les principaux KPI sont les suivants :

- le coût des interventions (à minimiser) ;
- la disponibilité (maintenance sans interruption de service) ;
- le respect des SLA des clients, en minimisant les pénalités le cas échéant.

5.1.4. Étude de cas 4 - Agroalimentaire

L'analyse bactériologique est un élément clef des contrôles qualité dans le domaine agroalimentaire. Le développement des techniques de séquençage génétique, en particulier la méta-génomique, permet d'identifier la plupart des micro-organismes présents dans un échantillon en une seule analyse et sans passer par une mise en culture. Ces analyses souffrent cependant d'un coût élevé lié à une automatisation partielle et des temps d'analyse informatique (notamment via un logiciel de recherche de similarités tel que BLAST) trop longs. Les délais de réalisation peuvent aller jusqu'à une semaine, ce qui réduit fortement la valeur de l'analyse pour le client.

L'objectif est de réduire le temps et les coûts des analyses. Sur cette base, plus de clients pourront être satisfaits. Les principaux KPI sont les suivants :

- le taux de succès des méthodes automatiques de caractérisation ;
- le temps de calcul via l'infrastructure informatique ;
- le taux de fiabilité du procédé ;
- le temps d'analyse et d'interprétation.

5.2. Du réactif au préventif puis au prédictif

Dans plusieurs domaines, il est intéressant de mettre en place un schéma permettant d'évoluer vers une réaction immédiate à des caractéristiques identifiées à travers les données, vers plus d'intelligence afin d'anticiper des situations indésirables, voire les prévenir suffisamment pour pouvoir les éviter. Nous donnons ici des illustrations sur nos quatre études de cas.

5.2.1. Étude de cas 1 - Maintenance informatique

En matière de maintenance, un KPI est le coût total d'appartenance ou « *TCO* » (Total Cost of Ownership). Celui-ci inclut le coût d'achat, de maintenance et de réparation en cas de panne. Différentes stratégies peuvent être envisagées :

– *réagir* simplement aux problèmes après la survenue d'une panne. Ceci se traduit par un coût généralement important car il faut réagir rapidement afin de minimiser le temps d'indisponibilité. Par ailleurs, toute indisponibilité a un impact négatif en termes d'image si un SLA a été violé ;

– *anticiper* leur occurrence sur base de l'observation du système. Des stratégies simples peuvent être mises en place, par exemple, déclencher des alertes quand un stockage approche d'un seuil proche de la capacité maximale. Ceci ne permet cependant pas de prévoir des pannes qui résultent d'enchaînements complexes d'événements ;

– *tenter de prédire* les problèmes sur base d'historiques connus et d'observations du système. C'est à ce niveau que des techniques d'analyse de données permettent de mettre en évidence des relations de cause à effet entre des parties du système qui, en cascade, peuvent causer une indisponibilité. Par exemple, l'application d'un correctif mal validé peut affecter un service qui peut lui-même paralyser un processus métier ;

– *optimiser* l'étape ultime. Il faut constamment veiller à ce que le système opère dans des conditions optimales en éliminant les causes des pannes possibles à la source.

La solution prédictive est la meilleure à notre sens, mais elle ne devrait être envisagée que si l'étape préventive est réalisée. De même, les patrons temporels les plus fréquents doivent être identifiés et traités en premier, par exemple, les stockages risquent plus une saturation les jours où des sauvegardes sont effectuées, généralement de manière prévisible (fin de semaine ou fin de mois). Une anticipation permettrait d'éviter des interventions coûteuses, notamment le week-end.

5.2.2. Étude de cas 2 - Itinéraires de soins

En matière de soins de santé, la mise en place d'itinéraires permet de faire une analyse plus riche des données produites. Il faut notamment être très vigilant avec les patients ayant un risque de complications impactant la qualité de leur traitement (par exemple, lié à une autre pathologie ou des intolérances dont ils souffrent). Les capacités d'analyse peuvent être mobilisées pour anticiper au maximum l'occurrence de ces risques via certains indicateurs (prise de sang, état général, etc.) tandis que la qualité globale est caractérisée par l'indicateur de RDI décrit précédemment. L'automatisation de ces analyses est d'autant plus importante que le suivi est généralement

pluridisciplinaire. Certaines interactions peuvent être complexes à appréhender par un seul spécialiste et par conséquent potentiellement susceptible d'échapper à l'analyse humaine.

Pour atteindre ces objectifs, il faut disposer de suffisamment de ressources. Ainsi, la main-d'œuvre disponible est influencée par la disponibilité du personnel mais aussi à cause des jours fériés. La planification est donc ardue. Un opérateur humain peut difficilement trouver une solution répondant simultanément à tous les patients et à toutes les contraintes du service. De plus, la planification doit constamment être reconsidérée pour faire face aux événements imprévus et au flux de patients entrants/sortants. En revanche, une solution prédictive et prescriptive combinée est très intéressante car elle a la capacité d'assurer un fonctionnement optimal des soins et des services en prenant en compte le risque que certains patients puissent être retardés.

5.2.3. Étude de cas 3 - Maintenance dans le domaine spatial

La stratégie est similaire à celle de la maintenance d'équipements IT mais pour des types d'équipements différents (systèmes de contrôle, alimentation, communication, etc.). On utilise ici des indicateurs de maintenance (MTBF - temps moyen entre les pannes, MTTR - temps moyen d'indisponibilité). Au niveau des équipements à maintenir, des données sur le cycle de vie sont disponibles, de même que des informations de logs générées par les équipements.

Comme dans le premier cas d'utilisation, une approche prédictive permet de réduire les coûts et réaliser une maintenance planifiée sans interruption de service. Elle peut être mise en place en se basant sur des pannes passées et des données semi-structurées fournies par les logs des équipements. A plus long terme, un résultat escompté consiste à ce que le système fournisse des informations prescriptives sur l'évolution du système pour en améliorer la fiabilité.

5.2.4. Étude de cas 4 - Agroalimentaire

L'accélération des analyses est rendue possible au niveau du séquençage lui-même via des techniques de nouvelle génération (*Next Generation Sequencing* ou NGS). Ces méthodes génèrent cependant de gros volumes de données (de l'ordre du Go) dont le traitement est un défi que les technologies Big Data peuvent relever. Dans l'optique de réaliser une analyse dans un délai maîtrisé, plusieurs stratégies sont à combiner :

– plusieurs algorithmes permettent de caractériser automatiquement les bactéries. De meilleures performances peuvent être obtenues en confrontant les résultats produits par chaque algorithme exécuté de manière parallèle ;

– des frameworks Big Data (notamment Hadoop) offrent de nombreuses briques (MapReduce, Cloudburst, DistMap...) utiles pour implémenter efficacement différentes étapes du traitement NGS telles que le découpage d'adaptateur, la vérification de qualité, la lecture, l'assemblage de novo, la quantification, l'analyse des variantes et l'annotation (Tripathi *et al*., 2016) ;

– des techniques de machine learning peuvent aussi être mises en place afin que l'analyse soit configurée de manière optimale en se basant sur des analyses antérieures.

5.3. Guidance dans la phase de compréhension du métier et des données

Cette phase est critique pour le succès du projet car l'objectif n'est pas seulement d'aboutir à une compréhension des besoins et des données disponibles mais aussi de mettre en place le noyau de personnes qui seront porteuses de la suite du projet. A cette fin, nous recommandons de l'organiser sur la base d'un ou plusieurs ateliers impliquant un responsable commercial, un analyste des données et un architecte SI. D'autres experts peuvent aussi être impliqués plus ponctuellement, par exemple, un responsable de la sécurité informatique peut être consulté pour valider à un stade précoce les problèmes possibles de sécurité ou de confidentialité. Il faut prendre en considération aussi bien le système actuel que l'évolution future du système d'information. Afin de mener à bien cette phase, une liste de contrôles utiles est reprise au tableau 4.

Pour soutenir l'organisation d'une manière efficace, des outils spécifiques de ces ateliers sont décrits à la section 5. A la fin de cette étape, une planification de projet est également définie. La tenue d'un atelier exige de prêter attention à de nombreuses questions tout en concentrant la discussion sur les plus pertinentes. A cet égard, un questionnaire peut fournir un soutien efficace, d'une part pour préparer l'atelier et d'autre part, pour servir de liste de contrôle (*check-list*). Le tableau 4 illustre quelques questions utiles à la compréhension des données à traiter.

Tableau 4. Quelques questions d'atelier sur les données

Q.UD.1	Quelles sont les sources de données et les types de données utilisés dans vos processus métiers actuels ?
Q.UD.2	Quels outils/applicatifs sont utilisés pour traiter vos processus métier actuels ?
Q.UD.3	Vos processus métier actuels effectuent-ils un traitement complexe des données ?
Q.UD.4	Quelle est la disponibilité de vos données ? Que se passe-t-il si les données ne sont pas disponibles ?
Q.UD.5	D'autres utilisateurs ont-ils un droit d'accès différent sur vos données ?
Q.UD.6	Vos données contiennent-elles des informations sensibles (par exemple, des données personnelles ou confidentielles de l'entreprise) ?
Q.UD.7	Quelles sont les conséquences de l'altération des données ?
Q.UD.8	Connaissez-vous le niveau de qualité de vos données ?

Quelques retours intéressants sur les divers ateliers sont les suivants :

– Étude de cas 1 - Maintenance informatique. L'atelier a pris la forme d'une séance de modélisation sur tableau blanc du processus de traitement d'un incident avec divers scénarios possibles. Durant le processus, les données impliquées (requêtes, tickets, log systèmes) ont été identifiées et leur disponibilité à pu être discutée, avec notamment un représentant de l'équipe sécurité/confidentialité également présent.

– Étude de cas 2 - Itinéraires de soins. Un atelier initial a été organisé de représentants des différents acteurs : docteurs, infirmières prodiguant les soins et infirmières administratives. Le domaine étant complexe, une immersion dans le service a été réalisée pour comprendre les processus et l'ensemble des données manipulées. Sur cette base, un modèle de processus a été modélisé et même formalisé au moyen d'outils dédiés.

– Étude de cas 3 - Maintenance dans le domaine spatial. L'atelier s'est basé sur deux types de document déjà disponibles vu les contraintes de documentation dans ce domaine : le modèle des flux d'information du système de monitoring/contrôle des satellites et une caractérisation complète des données. Ceci a permis de passer rapidement à la phase de compréhension des données.

– Étude de cas 4 - Agroalimentaire. Le processus actuel (séquençage, analyse, interprétation) a été passé en revue durant l'atelier, de même que les outils actuellement utilisés et leurs limitations. Les contraintes les plus importantes de volume et de qualité ont pu être identifiées et étudiées en détail.

5.4. Mise en place de points de contrôle

L'approche agile permet au processus d'être flexible et incrémental sur les activités. Avant de commencer une activité, il faut cependant disposer d'un minimum de résultats des étapes précédentes.

Dans ce sens, le tableau 5 reprend quelques contrôles à consulter au démarrage d'une activité.

Tableau 5. Liste (partielle) de vérification, de la préparation à l'évaluation

R.EV.1	Êtes-vous capable de comprendre/utiliser les résultats des modèles ?
R.EV.2	Est-ce que les résultats du modèle vous semblent pertinents d'un point de vue purement logique ?
R.EV.3	Y a-t-il des incohérences apparentes qui méritent d'être approfondies ?
R.EV.4	D'après votre première vision, les résultats semblent-ils répondre au métier de votre organisation ?

Seuls deux projets pilotes ont déjà atteint la phase de validation. Des retours intéressants pour ces projets sont les suivants :

– Étude de cas 1 - Maintenance informatique. Plusieurs modèles intéressants ont permis de mettre en évidence des symptômes précurseurs de certaines défaillances de services voire de remonter à des causes racines (par exemple l'application d'un correctif entraînant des conséquences indésirables imprévues).

– Étude de cas 2 - Itinéraires de soins. La pertinence de l'utilisation du modèle a pu être évaluée de manière très précise grâce à la mise au point d'un outil de simulation du fonctionnement normal d'un service d'oncologie mais aussi de cas de comportement limites ou imprévus tels que la surcharge du service, l'absence de personnel ou tout simplement un patient qui ne se présente pas à son traitement.

6. Travaux connexes et discussion

6.1. Adoption des méthodologies

La section 3 a montré une perspective historique exhaustive de l'évolution des méthodologies pertinentes. Alors que les premières approches proposées basées sur

l'exploration de données étaient déjà de nature itérative (Shearer, 2000), leur évolution au fil du temps consiste à s'intéresser de plus en plus à la manière de faciliter l'adoption de telles méthodologies par les entreprises. La culture de l'agilité a été une étape clé pour mieux intégrer le client dans le processus et conduire ce processus vers la production de la valeur métier (Franková *et al*., 2016). Les méthodes commerciales comme IBM Stampede s'inspirent fortement de cette tendance (IBM, 2013). En complément de ces méthodes, la nécessité d'identifier les obstacles et les facteurs d'adoption a également été abordée par des travaux récents discutés précédemment, tels que les facteurs de succès critiques (Saltz, Shamshurin, 2016).

Des méthodologies consolidées de Big Data sont également publiées sous forme de présentations plus pratiques et simplifiées afin d'être attractives pour les entreprises. La méthode « *DISTINCT* » est basée sur seulement quatre étapes (acquérir, traiter, analyser, visualiser) et permet l'utilisation de boucles de rétroaction pour faire un raffinement répété du traitement des données (Erl *et al*., 2016). Bien que la phase d'analyse ne soit pas explicitement mentionnée, cette approche itérative peut être utilisée pour mettre en place un canal de rétroaction entre l'informatique et les métiers. Après chaque cycle de rétroaction, le système peut ensuite être affiné en améliorant la préparation des données ou les étapes d'analyse des données. La série bien connue « for Dummies » a également dédié un livre sur le Big Data (Hurwitz *et al*., 2013), qui contient une section sur la façon de créer une feuille de route pour une mise en œuvre basée sur des facteurs tels que l'urgence métier, le budget, les compétences et le niveau de risque. Une approche de gestion agile est recommandée. La disponibilité de la BI est également identifiée comme un facteur d'accélération.

Les travaux menés dans des domaines connexes sur la manière de relever les défis organisationnels méritent d'être étudiés. Par exemple, le CCBF (*Cloud Computing Business Framework*) aide les entreprises à réussir une bonne conception de la solution Cloud ainsi que son déploiement. À l'instar de notre approche, le CCBF est un cadre conceptuel et architectural reposant sur la modélisation, la simulation, les expériences et les études de cas hybrides (Chang, 2015).

6.2. Préoccupations éthiques concernant la confidentialité des données

L'interaction avec les entreprises a soulevé des préoccupations éthiques et des questions telles que : « Sommes-nous suffisamment prudents sur le phénomène Big Data ? » La technologie a un énorme potentiel pour améliorer la santé et les conditions de vie des humains. Néanmoins, nous sommes confrontés à un problème d'éthique face aux algorithmes prédictifs. Une intensification de la réglementation est donc nécessaire pour trouver un bon compromis entre l'utilisation des données personnelles et la protection de la vie privée.

Par exemple, dans le domaine de la santé, nous pouvons nous interroger sur la façon dont les gouvernements et organismes publics ont l'intention d'exploiter les données recueillies. En Belgique, les hôpitaux revendent d'ores et déjà ces données aux firmes pharmaceutiques (RTBF, 2017).

Parmi les enjeux éthiques du Big Data, la sécurité des données est primordiale contre le piratage et nécessite le développement de nouveaux systèmes de sécurité pour sécuriser les échanges en veillant à un contrôle strict des accès à la plateforme Big Data et garantir ainsi la confidentialité des données. Sécuriser une plateforme Big Data est un domaine à part entière car son architecture peut être de nature hétérogène et répartie sur plusieurs nœuds. Cela repose aussi typiquement sur des technologies Cloud qui suscitent des interrogations par rapport au traitement des données sensibles.

Grâce à l'utilisation de données massives dans la communauté médicale, les aspects juridiques et économiques changent à grande vitesse et défient les principes éthiques et les règles dans la relation entre un médecin et son patient. Cela perturbe l'équilibre entre confidentialité et transparence et crée un sentiment de perte de confiance dans l'environnement de la santé autour de la gestion et l'exploitation des Big Data. L'éthique de ce type de données nécessite un contrôle bien supervisé de l'utilisation de l'information médicale (Béranger, 2016 ; EESC, 2017).

Des études ont également démontré le pouvoir de segmentation de la modélisation prédictive et les avantages commerciaux qui en résultent pour les compagnies d'assurance vie (Batty *et al.*, 2010). Alors que certains clients à risque faible, pourraient bénéficier de meilleures conditions d'assurance, les clients présentant des risques anticipés plus élevés pourraient être exclus en raison de tarifs inabordables, réduisant ainsi l'effet de solidarité des assurances vie.

Les données véhiculent de plus en plus d'informations de localisation en raison du développement important d'applications mobiles et de l'émergence de l'Internet des objets. Une analyse par algorithmes prédictifs, peut être utilisée pour calculer le taux de probabilité qu'un événement se passe à tel endroit et à telle heure. Une branche spécifique du Big Data appelée analyse de localisation se concentre spécifiquement sur ce domaine et peut mettre en danger la vie privée si elle est appliquée sans garantie. Des directives et techniques spécifiques sont en cours d'élaboration à cette fin. Certaines lignes directrices sont publiées, par exemple, par la commission européenne pour l'administration publique (Bargiotti *et al.*, 2016). Des techniques de traitement des données et des algorithmes sont également développés pour préserver la confidentialité des services basés sur la localisation (Sun *et al.*, 2017 ; Liu, 2007).

À un niveau plus général, afin de mieux contrôler les énormes quantités de données traitées chaque jour, et de s'assurer que chaque personne est respectée, la commission européenne a publié le Règlement Général sur la Protection des Données ou RGPD (en anglais : *General Data Protection Regulation* ou GDPR) en 2016 pour entrer en vigueur progressivement en mai 2018 (European Commission, 2016). Un portail européen avec des ressources étendues est également disponible. Un élément important que le gestionnaire de données doit prendre en compte tout particulièrement avec beaucoup d'attention est d'obtenir le consentement explicite de l'utilisateur pour le traitement de ses données privées et pour toutes les finalités envisagées. Pour reprendre le cas des itinéraires cliniques, le traitement des données vise à garantir la qualité des soins du patient. S'il y a une finalité d'étude clinique, celle-ci doit être indiquée et consentie.

Notre recommandation basée sur nos pilotes, est d'aborder cette question tôt dans le processus, si possible, déjà lors de la phase de compréhension et impliquer des personnes compétentes comme le responsable de la sécurité de l'information ou même un délégué plus spécifique de la protection des données si ce rôle est défini. En fait, cela s'est produit tout naturellement dans la plupart de nos cas pilotes car les données doivent être traitées en dehors de l'organisation propriétaire. Cependant, l'accent était davantage mis sur la confidentialité que sur le but du traitement des données lui-même.

6.3. Problèmes de cybersécurité

Parmi les défis du Big Data, la sécurité des données est primordiale contre le piratage et nécessite la conception de systèmes capables d'assurer un contrôle très exigeant de l'accès à la plateforme Big Data, tout en garantissant la confidentialité des données. L'European Union Agency for Network and Information Security (ENISA) a produit un panorama des menaces Big Data et un guide de bonnes pratiques (Damiani *et al*., 2016). Ce document répertorie les actifs (*assets*) typiques du Big Data, identifie les menaces ainsi que les vulnérabilités et les risques associés. Sur la base de ces points, il suggère de bonnes pratiques émergentes et pointe les domaines de recherche.

Le Big Data nécessite des capacités de stockage considérables, donc le recours à des infrastructures distribuées, telles que le Cloud. En matière de sécurité, les problèmes de cette liés au Cloud ont été identifiés très tôt (Popovic, Hocenski, 2010). L'angle d'attaque est largement augmenté par plusieurs facteurs, tels que la multiplication aisée des serveurs (jusqu'à cinq fois plus nombreux que les serveurs physiques) ou encore des mouvements de données jusque 100 fois plus nombreux (CloudPassage, 2016). Au niveau SaaS, la présence d'API standards constitue aussi un risque accru de faille, causant des fuites massives de données voire la compromission de millions de comptes personnels (Lord, 2017). Bien sûr, une solution de Cloud privée ou hybride bien conçue permet de remédier à ce type de risque et garder une certaine maîtrise. D'autre part, les fournisseurs de solutions Cloud sont de plus en plus attentifs à ces aspects et intègrent la sécurité elle-même comme service supporté par la plateforme (Olavsrud, 2017).

Enfin, le stockage de données sensibles sur le Cloud n'est pas sans conséquence, car les réglementations ne sont pas les mêmes dans tous les pays. Un aspect sensible est la gestion des emplacements de stockage et de traitement de données, par exemple, la nécessité de traiter les données dans un pays donné. Cependant, étant donné que cette situation entrave également la compétitivité de l'Europe sur un marché mondial, l'UE travaille actuellement sur un cadre pour la libre circulation des données non personnelles dans l'UE (European Commission, 2017).

7. Conclusion et perspectives

Dans cet article, nous avons décrit comment aborder les défis et les risques liés au déploiement d'une solution Big Data au sein des organisations et entreprises sou-

haitant s'appuyer sur cette technologie pour soutenir leur développement. Sur base de différentes méthodes et études déjà rapportées dans la littérature, nous avons identifié une série de briques méthodologiques adaptées à nos besoins, en y intégrant des retours d'expérience de plusieurs pilotes.

Notre principale contribution qui continue à évoluer au fil des projets pilotes, représente le processus de mise en place d'un projet Big Data. Celui-ci vise à maximiser les chances de succès en s'adaptant aux besoins de l'organisation cible. Nous proposons en outre une série de recommandations soutenant cette mise en œuvre. Bien que centrée sur quelques pilotes, notre approche se veut donc générale et permet aux personnes confrontées aux mêmes défis de disposer de briques méthodologiques utiles pour déployer efficacement un projet Big Data et de bien en gérer les difficultés.

Jusqu'à présent, nous nous sommes focalisés davantage sur les phases de découverte et de compréhension des données. Dans la suite de nos travaux, nous explorerons plus en détail la phase d'exécution du projet au fur et à mesure que nos projets pilotes auront atteint leur terme ou des jalons importants. Un autre point utile à approfondir est le problème de modélisation, notamment au niveau conceptuel, des Big Data notamment par des approches MDA (Abdelhédi *et al*., 2017).

Remerciements

Ce travail a été financé en partie par le projet PIT Big Data de la région wallonne (no 7481). Nous remercions nos partenaires d'avoir partagé leurs cas d'étude.

Bibliographie

Abdelhédi F., Brahim A. A., Atigui F., Zurfluh G. (2017). MDA-Based Approach for NoSQL Databases Modelling. In *Proceedings 19th International Conference DaWaK, Lyon, France, August 28-31*, p. 88–102.

Alliance A. (2001). *Agile Manifesto.* http://agilemanifesto.org.

Balduino R. (2007). *Overview of OpenUP.* https://www.eclipse.org/epf/general/OpenUP.pdf.

Bargiotti L. *et al*. (2016). *European Union Location Framework Guidelines for public administrations on location privacy. JRC Technical Reports.*

Batty M. *et al*. (2010, April). *Predictive Modeling for Life Insurance Ways Life Insurers Can Participate in the Business Analytics Revolution.* Deloitte Consulting LLP.

Bedos T. (2015). *5 key things to make big data analytics work in any business.* http://www.cio.com.au/article/591129/5-key-things-make-big-data-analytics-work-any-business.

Béranger J. (2016). *Big data and ethics: The medical datasphere*. Elsevier Science.

Campbell H., Hotchkiss R., Bradshaw N., Porteous M. (1998). Integrated care pathways. *British Medical Journal*, p. 133-137.

Chang V. (2015). *A proposed cloud computing business framework*. Commack, NY, USA, Nova Science Publishers, Inc.

Chen H., Chiang R. H. L., Storey V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Q.*, vol. 36, n° 4.

Chen H., Kazman R., Haziyev S. (2016). Agile big data analytics development: An architecture-centric approach. In *Proc. hicss'16, hawaii, usa.*

CloudPassage. (2016). *Survey: Exponential Server Growth, Dynamics of Cloud Increase Attackable Surface Area and Risk.* http://bit.do/cloud-passage.

Corea F. (2016). *Big data analytics: A management perspective* (1st éd.). Springer Publishing.

Crowston K. (2010). A capability maturity model for scientific data management.

Dam P. A. van. (2013). A dynamic clinical pathway for the treatment of patients with early breast cancer is a tool for better cancer care : implementation and prospective analysis between 2002–2010. *World Journal of Surgical Oncology*, vol. 11, n° 1, p. 70.

Damiani E. *et al*. (2016). *Big data threat landscape and good practice guide.* https://www.enisa.europa.eu/publications/bigdata-threat-landscape.

EESC. (2017). *The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context.* European Economic and Social Committee.

Erl T., Khattak W., Buhler P. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall.

European Commission. (2016). *Regulation (EU) 2016/679 - General Data Protection Regulation (GDPR).* http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679.

European Commission. (2017). *A framework for the free flow of non-personal data in the EU.* http://europa.eu/rapid/press-release_MEMO-17-3191_en.htm.

Franková P., Drahošová M., Balco P. (2016). Agile project management approach and its use in big data management. *Procedia Computer Science*, vol. 83, p. 576 - 583.

Gao J., Koronios A., Selle S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. In *Amcis.*

Gartner. (2016). *Investment in big data is up but fewer organizations plan to invest.* http://www.gartner.com.

Halper F. (2014). *Predictive Analytics for Business Advantage.* The Data Warehousing Institute Best Practices Report, TDWI.

Hoppen J. (2015). *7 characteristics to differentiate BI, Data Mining and Big Data.* https://aquare.la/articles/2015/05/01/7-characteristics-differentiate-bi-data-mining-big-data.

Hurwitz J. *et al*. (2013). *Big Data For Dummies* (J. W. . Sons, Ed.).

IBM. (2013). *Stampede.* http://www.ibmbigdatahub.com/tag/1252.

IBM. (2015). *Have you seen ASUM-DM?* https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/.

Kelly J., Kaskade J. (2013). *CIOs & Big Data: What Your IT Team Wants You to Know.* http://blog.infochimps.com/2013/01/24/cios-big-data.

Lamsweerde A. van. (2009). *Requirements engineering - from system goals to UML models to software specifications*. Wiley.

Liu L. (2007). From data privacy to location privacy: Models and algorithms. In *Proc. of the 33rd international conference on very large data bases*, p. 1429–1430. VLDB Endowment.

Lord N. (2017). *The History of Data Breaches.* https://digitalguardian.com/blog/history-data-breaches.

Lyman G. (2009, Jul). Impact of chemotherapy dose intensity on cancer patient outcomes. *Journal of the National Comprehensive Cancer Network*, p. 99-108.

Mariscal G., Marban O., Fernandez C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Eng. Review*, vol. 25, n° 2, p. 137-166.

Mauro A. D., Greco M., Grimaldi M. (2016). A formal definition of big data based on its essential features. *Library Review*, vol. 65, n° 3, p. 122-135.

Nascimento G. S. do, Oliveira A. A. de. (2012). An Agile Knowledge Discovery in Databases Software Process. In *3rd Int. Conf., ICDKE, Wuyishan, Fujian, China, Nov. 21-23.*

Nott C. (2014). *Big Data & Analytics Maturity Model.* http://www.ibmbigdatahub.com/blog/big-data-analytics-maturity-model.

Olavsrud T. (2017). *Security-as-a-service model gains traction.* https://www.cio.com/article/3192649/security/security-as-a-service-model-gains-traction.html.

Piccart M., Biganzoli L., Di Leo A. (2000). The impact of chemotherapy dose density and dose intensity on breast cancer outcome: what have we learned? *European Journal of Cancer*, vol. 36.

Ponsard C., Touzani M., Majchrowski A. (2017). Amélioration des méthodes de conduite de projets big data : retour d'expérience de pilotes industriels multi-sectoriels. In *Actes du XXXVème Congrès INFORSID, Toulouse, France, 30 Mai-2 Juin, 2017*, p. 179–194.

Popovic K., Hocenski Z. (2010, May). Cloud Computing Security Issues and Challenges. In *The 33rd international convention mipro*, p. 344-349.

Rot E. (2015). *How Much Data Will You Have in 3 Years?* http://www.sisense.com/blog/much-data-will-3-years.

RTBF. (2017). *Vos données médicales sont revendues, vous le saviez?* https://www.rtbf.be/info/belgique/detail_vos-donnees-medicales-sont-revendues-vous-le-saviez?id=9728058.

Saltz J. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In *IEEE int. conf. on big data.*

Saltz J., Shamshurin I. (2016). Big Data Team Process Methodologies: A Literature Review and the Identification of Key Factors for a Project's Success. In *Proc. IEEE International Conference on Big Data.*

Scrum Alliance. (2016). *What is scrum? an agile framework for completing complex projects.* https://www.scrumalliance.org/why-scrum.

Shearer C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, vol. 5, n° 4.

Sun G., Liao D., Li H., Yu H., Chang V. (2017). L2P2: A location-label based approach for privacy preserving in LBS. *Future Generation Computer Systems*, vol. 74, p. 375 - 384.

Tripathi R., Sharma P., Chakraborty P., Varadwaj P. K. (2016). Next-generation sequencing revolution through big data analytics. *Frontiers in Life Science*, vol. 9, n° 2, p. 119-149.

Vaillancourt J. (2010). Statistical methods for data mining and knowledge discovery. In *Proc. of the 8th int. conf. on formal concept analysis*, p. 51–60. Springer-Verlag.