Amélioration des méthodes de conduite de projets Big Data : retour d'expérience de pilotes industriels multi-sectoriels

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

1. *CETIC - Centre de recherche, Gosselies, Belgique*
   {christophe.ponsard,annick.majchrowski}@cetic.be
2. *Académie de Toulouse, Toulouse, France*
   mounir.touzani@ac-toulouse.fr

RÉSUMÉ. *Afin de mener à bien leurs activités, les entreprises sont de plus en plus confrontées au défi de traiter des quantités croissantes de données provenant de dépôts numériques, d'applications d'entreprise, de réseaux de capteurs... Bien qu'un large éventail de solutions techniques soit disponible pour traiter ces données massives (Big Data), beaucoup d'entreprises peinent à les déployer en raison d'un manque de maturité lié à leur gestion. Cet article propose une guidance en la matière. Il s'ancre dans des méthodes documentées dans la littérature, trouvant leurs racines dans les projets de fouille de données. Nous avons également mené une série de pilotes Big Data dans différents domaines (IT, médical, sciences de la vie, spatial) qui nous ont permis de dégager un retour d'expérience et un guide pratique pour la conduite d'un projet Big Data. Ceci permet d'exploiter au mieux les méthodologies disponibles afin de traiter les problématiques relatives à la collecte des exigences, l'exploration et la préparation des nouvelles données, le phasage itératif de l'implantation de la solution et une montée en maturité.*

ABSTRACT. *Nowadays, in order to successfully run their business, companies are facing the challenge to process ever increasing amounts of data generated from digital repositories, enterprise applications, sensors networks... Although a wide range of technical solutions are available to deal with such Big Data, many companies fail to deploy them actually because a lack of maturity in process and management challenges. This paper aims at providing guidance in those matter. We report about lessons learnt when deploying a series of Big Data pilots in different domains. We provide feedback and some practical guidelines on how to organise and manage a project based on available methodologies, covering topics like requirements gathering, data understanding, iterative project execution and raising the level of 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*

1. Introduction

Notre monde est actuellement en train de vivre une explosion de l'information. De nombreuses statistiques attestent de la montée en puissance du phénomène Big Data. Par exemple, il est souvent rapporté que 90% des données dans le monde ont été produites seulement ces deux dernières années et que le volume des données créé par les entreprises double tous les 1,2 années (Rot, 2015).

Les organisations perçoivent bien le grand potentiel que les technologies Big Data peuvent leur apporter pour améliorer leur performance, et dans le cas des entreprises, pour accroître leur avantage compétitif. La facilité de collecter et stocker les données, combinée avec la disponibilité d'outils technologiques de stockage et d'analyse à grande échelle (notamment les bases de données NoSQL, MapReduce, Hadoop) a incité un certain nombre d'entre elles à démarrer des projets Big Data.

Les caractéristiques et 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 (afin de les interpréter facilement) et la Valeur (pour en tirer un revenu) (Mauro *et al.*, 2016).

Cependant, le constat est que la plupart des organisations ne parviennent toujours pas 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 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). Un examen de la littérature révèle qu'actuellement, 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 aux méthodes et aux outils qui pourraient aider les équipes à mener efficacement des projets Big Data à terme (J. 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 (J. S. Saltz, 2015), aussi bien sur des problèmes de gestion de projet (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 (J. 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 solution Big Data. A travers ce travail, nous souhaitons apporter quelques éléments

de réponses 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 quatre autres sont en phase de planification.

Ce document est structuré comme suit. La section 2 donne une typologie des principales catégories de projets Big Data. La section 3 passe ensuite en revue les principales méthodologies concernant le déploiement du Big Data. Dans la section 4, nous présentons la méthodologie suivie pour mener nos projets pilotes et dégager une guidance méthodologique. Nous mettons l'accent sur les facteurs clés de succès du déploiement d'une solution Big Data. La section 5 détaille notre retour d'expérience en donnant des recommandations ciblant des étapes particulièrement importantes. Enfin, la section 6 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 les moyens permettant d'acquérir des données depuis de sources diverses, de les traiter afin de découvrir des relations qui les relient et mettre des résultats à disposition des parties prenantes (H. Chen *et al.*, 2012). L'application de ces techniques par des entreprises ("Business Analytics") leur permet de mieux comprendre leur niveau de performance et de procéder à des améliorations. 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 l'exploration de données et l'apprentissage automatique ("machine learning")

– *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, des 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 '90 avec pour objectif d'extraire des données à partir d'informations structurées (bases de données) pour découvrir des facteurs clés de l'entreprise à une échelle relativement petite. Le Big Data, quant à lui, opère aussi sur des données non structurées, sur une plus grande échelle et vise à dégager des indicateurs à vocations prédictives. Cependant, un point commun aux deux types d'approches est qu'en termes de processus, il est 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 '90 et vise essentiellement à produire des indicateurs clé de performance (en anglais KPI : Key Performance Indicator) sous forme de tableaux de bord. Les techniques s'appuient 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 Big Data apportent des possibilités d'analyse prédictive (Halper, 2014).

L'approche séminale en matière de fouille de données est KDD (Knowledge Discovery in Database). 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 1. Elle est composée de six phases, chacune étant décomposée en sous-étapes. Le processus n'est pas linéaire, mais plutôt organisé comme un cycle global avec généralement des revues entre les phases. CRISP-DM a été largement utilisé depuis 20 ans, non seulement pour la fouille de données, mais aussi pour l'analyse prédictive et des projets Big Data.

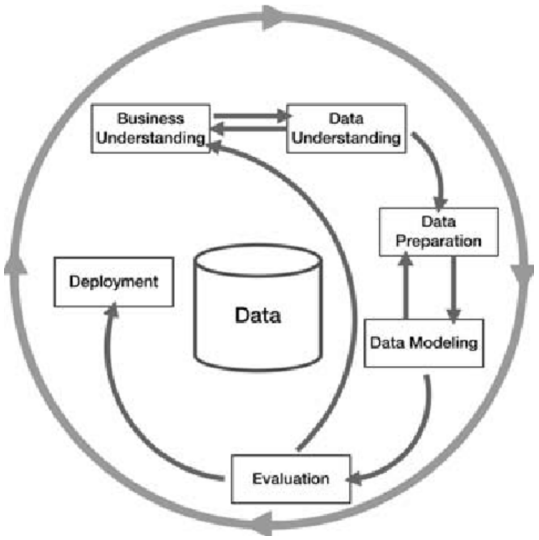

Figure 1. La méthode CRISP-DM

CRISP-DM et les méthodes similaires souffrent toutefois des problèmes suivants :

– elles ne fournissent pas une bonne vision du management du point de vue communication ainsi qu'au niveau de la connaissance et sur les aspects des projets.

– elles manquent d'une certaine forme de maturité au niveau du modèle pour permettre de mettre en évidence des étapes et des jalons plus importants, qui peuvent être améliorés progressivement.

– en dépit de la normalisation, elles ne sont pas très largement connues des entreprises qui peuvent donc difficilement les adopter pour mieux gérer la valeur de leurs données.

3.2. Vers plus d'agilité

Les méthodes agiles sont des méthodes itératives qui répondent au manifeste agile dont les principes mettent l'interaction avec le client, l'adaptation aux changements et la production de valeur au centre du processus de développement (Alliance, 2001). Initialement développées 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 en particulier pour 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. Dans chaque sprint, les équipes doivent produire de la valeur ajoutée aux parties prenantes de manière prédictive et démontrable.

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, tout comme c'est le cas dans le domaine du développement logiciel. 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 tout 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 liées à la 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).

3.3. Méthodes spécifiques pour le Big Data

La méthode AABA (Architecture-centric Agile Big data Analytics) répond aux défis techniques et organisationnels de Big Data (H.-M. Chen *et al.*, 2016). La méthode intègre à la fois une méthode de 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 livraison continue de valeur.

La méthode a été validée sur 11 études de cas dans différents domaines notamment en marketing, télécom et 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. des pics d'efforts en mode agile permettent de faire face aux évolutions rapides des technologies et des exigences
4. la définition d'une architecture de référence permet une plus grande flexibilité.
5. les boucles de rétroaction permettent de traiter les exigences non fonctionnelles telles que la performance, la disponibilité et la sécurité, mais aussi pour disposer d'un retour rapide des clients à propos d'exigences émergentes.

Parallèlement, *Stampede* est une méthode proposée par IBM à ses clients. Son principal objectif est d'encourager les entreprises et les aider à démarrer plus rapidement, afin de générer de la valeur à partir du Big Data. La méthode s'appuie sur la mise à disposition de ressources d'experts à un coût permettant d'aider les entreprises à se lancer dans le Big Data, dans le cadre d'un projet pilote bien défini (IBM, 2013). La méthode, illustrée à la figure 2, s'appuie notamment sur un atelier d'une demi-journée permettant de définir le projet Big Data, d'identifier l'infrastructure nécessaire, d'établir un plan de travail mais surtout et avant tout d'établir la valeur pour l'entreprise. L'exécution du pilote est généralement répartie sur 12 semaines et réalisée de manière agile avec un jalon important vers la 9ème semaine.

Figure 2. Méthode Stampede d'IBM

Des tentatives ont également été menées pour développer un *modèle de maturité de capacité de type CMM* (Capability Maturity Model) pour les processus de gestion des données scientifiques, dans le but de soutenir l'évaluation et l'amélioration de ces processus (Crowston, 2010) (Nott, 2014). Un tel modèle décrit les principaux types de processus ainsi que les pratiques nécessaires à une gestion efficace. Un CMM caractérise les organisations au moyen d'un niveau de maturité qui représente leur capacité à exécuter des processus de façon fiable. Une échelle classique sur 5 niveaux est typiquement utilisée à la fois dans (Crowston, 2010) et (Nott, 2014). Le premier utilise les niveaux standard allant de "défini" à "optimisé" tandis que le second utilise une nomenclature plus spécifique allant de "ad hoc" à "breakaway". Le tableau 1 en détaille les principaux critères qui concernent la place de la donnée dans la stratégie métier, le type d'analyse de données utilisée, l'alignement de l'infrastructure IT, ainsi que des aspects de culture et de gouvernance.

Tableau 1. Modèle de maturité de Nott et Betteridge (IBM)

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'événement	Prédiction de certaines probabilité d'évolution	Optimisation des décisions	Optimisation et automation 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 oeuvre généralisée

3.4. Approches complémentaires

Sensemaking est également une approche itérative, mais en rapport avec les processus cognitifs réalisés par les humains afin de se construire une représentation mentale de l'information pour atteindre l'objectif visé. Elle met l'accent sur les défis de la modélisation et de l'analyse en intégrant les modèles cognitifs afin d'analyser les

caractéristiques des données et de détailler les activités des utilisateurs (Lau *et al.*, 2014).

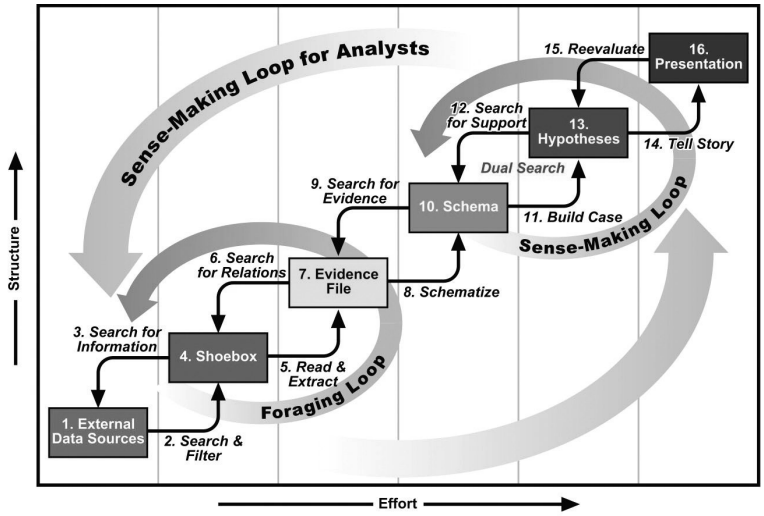

Figure 3. Méthode SenseMaking

De nombreux facteurs clés de succès, guides pratiques et listes de contrôle des risques ont été également publiés, principalement dans les blogs pour les directeurs des systèmes d'information, p. ex. (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 (J. Saltz, Shamshurin, 2016) pour traiter aussi des dimensions de l'outillage et de la gouvernance. Les principaux facteurs clés sont les suivants :

– 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 de changement, la maturité, la volumétrie des données

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

– 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 4, il est guidé par huit pilotes successifs qui sont utilisés pour affiner la méthode et rendre plus technique les briques disponibles à travers l'infrastructure proposée. Le résultat final attendu est de fournir un service reproductible de manière fiable aux entreprises ayant de tels besoins.

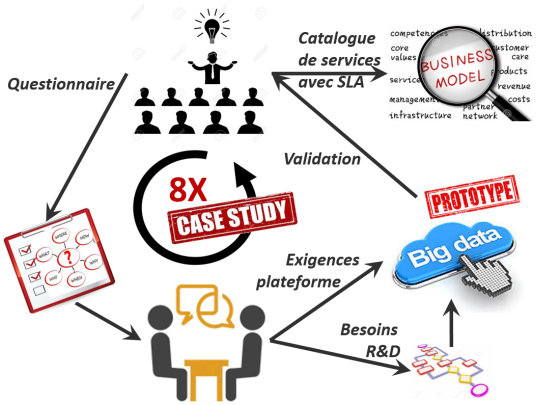

Figure 4. Développement itératif de la méthode et de l'infrastructure

La méthode choisie est fortement inspirée des méthodes et processus décrits dans la section 3 :

– le point de départ a été Stampede grâce à la plate-forme d'IBM. Les principaux aspects retenus à partir des méthodes sont : 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 qui est 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 : 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).

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 premier "V" du Big Data ainsi que de la typologie

Tableau 2. Principales caractéristiques des 4 premiers projets pilotes

#	Domaine	Volume	Vélocité	Variété	Caractérisation
1	Sciences de la vie	20 Go/analyse, 2 To/semaine	Haute (à paralléliser)	Données métier et de traçabilité (ex. agro-alimentaire)	Essentiellement descriptive, au niveau de la qualité des produits
2	Spatial	Maintenance infrastructure sol Galileo	Moyenne	Haute: messages, logs	Maintenance prédictive de matériel coûteux. Fiabilité 99.8%
3	Santé	900 lits sur 3 sites	Temps-réel	Nombreuses sources, formats divers	Analyse prédictive et prescriptive pour réduire la morbidité. Confidentialité.
4	Maintenance IT	Environ 3000 serveurs	Haute (événements, logs,...)	Temps-réel	Analyse prédictive pour maîtriser les coûts de maintenance

4.3. Schéma général appliqué au sein de chaque pilote

La méthodologie qui a émergé sur base des méthodologies existantes et sur base des itérations sur nos 4 pilotes se compose de 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 participante. Une introduction générale est donnée sur les concepts du Big Data. La plate-forme mise à disposition est présentée, de même que quelques applications représentatives dans différents domaines (éventuellement avec déjà 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 de l'entreprise et du cas d'utilisation.

Cette phase est largement alignée avec 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 et des outils méthodologiques concrets pour aider à la conduire sont détaillés dans la section 4.

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 réalistes ou même réelles.

– *Déploiement* : transférer la solution validée à l'environnement de production et veiller à ce que l'utilisateur puisse l'utiliser (par exemple, au moyen de bons outils de visualisation et d'un tableau de bord). Les activités de surveillance de performance et de précision sont également mises en place.

5. Retour d'expérience et recommandations

Dans cette section, nous présentons nos principaux retours d'expérience ainsi que des recommandations méthodologiques permettant d'augmenter les chances de succès d'un projet de déploiement Big Data.

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

Par le déploiement d'une solution Big Data, une entreprise s'attend à gagner de la valeur de ses données. La façon de mesurer cette valeur doit être définie dès la phase de compréhension des données de l'entreprise, généralement en s'appuyant sur les indicateurs clés de performance (KPI). Ces KPI doivent déjà être clairement définis par l'entreprise et celle-ci doit être déjà en mesure de les mesurer.

Sur cette base, différentes stratégies d'amélioration peuvent être identifiées et discutées pour aboutir à la sélection d'un bon projet pilote de mise en œuvre. Dans ce processus de sélection, l'écart avec la situation actuelle et le niveau de maturité doivent également être pris en considération. Il est plus sûr de garder un premier projet avec des objectifs assez modestes que de risquer l'échec en visant un projet trop complexe, même s'il pourrait apporter plus de valeur. Une fois que ce premier projet pilote réussi, d'autres étapes peuvent être planifiées pour mettre en place des traitements plus complexes amenant plus de valeur à l'entreprise.

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 deux illustrations respectivement dans le domaine de la maintenance et de la santé.

Étude de cas du domaine de la maintenance informatique. En matière de maintenance, un KPI est le coût total d'appartenance (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 voire de pénalité si un SLA (Service Level Agreement) a été violé.

– *anticiper* leur occurrence sur la 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 résultants d'enchaînements complexes d'événements.

– *tenter de prédire* les problèmes sur base d'historique connu et d'observation 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 veiller constamment 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.

Étude de cas dans le domaine des itinéraires de soins. En matière de soins de santé, les hôpitaux déploient de plus en plus des trajets de soins, définis comme une vision pluridisciplinaire du processus de traitement requis par un groupe de patients présentant la même pathologie avec un suivi clinique prévisible (Campbell *et al.*, 1998). Il peut par exemple s'agir d'un pontage cardiaque ou d'une chimiothérapie. Ceci permet de non seulement de réduire la variabilité des processus cliniques mais aussi d'améliorer la qualité et mieux en maîtriser les coûts (Dam, 2013). La mise en place d'itinéraires permet aussi de faire une analyse plus riche des données produites : on peut ainsi détecter des patients ayant un profil qui pourrait impacter la qualité de leur traitement (par exemple lié à une autre pathologie dont il souffre ou des intolérances).

L'automatisation de ces analyses est d'autant plus importante que le suivi est généralement pluridisciplinaire et que certaines interactions peuvent être complexes à appréhender par un seul spécialiste et peuvent donc potentiellement échapper à l'analyse humaine. Ceci est particulièrement critique dans la cas de traitement tel que le

cas des chimiothérapies où le respect du temps et des doses est important. Un indicateur médical défini en la matière est le RDI (Relative Dose Indicator). Dans le cas du cancer du sein, il a été montré que la dégradation de cet indicateur avait un impact direct sur la courbe de rémission (Piccart *et al.*, 2000). La surveillance du RDI par le système et l'analyse prédictive des facteurs qui l'impacte est donc primordiale.

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 sera porteuse de la suite du projet. A cette fin, on recommande de l'organiser sur la base d'un ou plusieurs ateliers impliquant un responsable commercial, l'analyste des données et l'architecte SI. D'autres experts peuvent aussi être impliqués plus ponctuellement, par exemple, le responsable de la sécurité informatique peut être consulté pour valider à un stade précoce les problèmes possibles de sécurité ou confidentialité. A la fois le système actuel et l'évolution future du système d'information doivent être considérés. Afin de mener cette phase, une liste de contrôles utiles est représentée à la figure 5.

Figure 5. Compréhension de l'entreprise et des cas d'utilisation

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 à la fois comme préparation avant l'atelier et comme une

liste de contrôle (check-list) pendant celui-ci. Le tableau 3 illustre quelques questions utiles à la compréhension des données à traiter.

Tableau 3. 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 métier 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	Des utilisateurs autres 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 ?

5.4. Utilisation de notations pour la modélisation

L'utilisation de notation de modélisation est utile comme outil pour inventorier les données, comprendre leur structure et comprendre les différents flux d'information. Il ne faut cependant pas la confondre avec l'étape technique de modélisation qui est ultérieure. Pendant les ateliers, un tableau blanc peut être utilisé pour esquisser des modèles dans un mode collaboratif avec les participants.

Selon notre expérience, les modèles de flux de données aident à comprendre quel processus génère, modifie, stocke ou extrait des données. Les modèles d'entités-relations (ou diagrammes de classe ou ontologies) aident à capturer la structure du domaine

Par contre, les cas d'utilisation doivent être évités car ils se focalisent sur une fonctionnalité spécifique mais ne permettent pas de mettre en évidence les liens entre les données. Ils ne peuvent donc pas fournir une image globale du problème.

5.5. 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 but, le tableau 4 reprend quelques contrôles à consulter au démarrage d'une activité.

6. 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 d'organisations et en particulier d'entreprises souhaitant s'appuyer sur cette technologie pour soutenir leur développement. Sur base de différentes méthodes et études déjà rapportées dans la littérature, nous

Tableau 4. 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 ?

avons élaboré de manière itérative une méthode adaptée à nos besoins en y intégrant des retours d'expérience de plusieurs pilotes. Au delà de cette méthode qui continue à évoluer au fil de projets pilotes, notre principale contribution est centrée sur le processus suivi pour mettre en place un projet Big Data qui maximise les chances de succès et qui s'adapte 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 même défis de disposer de briques méthodologiques utiles pour déployer efficacement un projet Big Data et bien en gérer les difficultés et pièges.

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 détails la phase d'exécution du projet au fur et à mesure que nos projets pilotes auront atteint leur terme ou des jalons importants.

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 leur cas d'étude.

Bibliographie

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

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

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.

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

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

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

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, nº 1, p. 70.

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

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.

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

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.

Lau L., Yang-Turner F., Karacapilidis N. (2014). Requirements for big data analytics supporting decision making: A sensemaking perspective. In *Mastering data-intensive collaboration and decision making.* Springer Science & Business Media.

Mariscal G., Marban O., Fernandez C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Eng. Review*, vol. 25, nº 2, p. 137-166.

Mauro A. D., Greco M., Grimaldi M. (2016, 04 04). A formal definition of big data based on its essential features. *Library Review*, vol. 65, nº 3, p. 122-135.

Nascimento G. S. do, Oliveira A. A. de. (2012). An agile knowledge discovery in databases software process. In *Data and knowledge engineering: Third international conference, icdke, wuyishan, fujian, china, nov. 21-23.* Springer Berlin Heidelberg.

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

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

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

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.*

Saltz J. S. (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.*

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, nº 4.