

EPSI BORDEAUX

73 rue de Marseille,

33000 Bordeaux



ASAPE

347 Avenue Thiers

33100 BORDEAUX



Le Datamining et la gestion de projet

Ikram CHRAIBI KAADOUD

Directeur de Recherche

Sylvain Labasse

Promotion 2013

Remerciements

Je tiens à remercier dans un premier temps, M. Sylvain Labasse pour son suivi et son aide. Ses précieux conseils, sa patience et son expérience m'ont permis d'aborder sereinement la rédaction de ce mémoire.

Je remercie également l'équipe pédagogique et administrative de l'EPSI, ainsi que les intervenants professionnels pour avoir assuré la partie théorique de ma formation au cours de cette année.

Je remercie tout particulièrement et témoigne toute ma reconnaissance à Madame Isabelle DUFFAUD, à Monsieur Laurent PAITREULT et à Monsieur Stéphane LEYMARIE, les responsables d'ASAPE, pour m'avoir accompagné au cours de cette année et de m'avoir permis, de vivre une expérience aussi enrichissante humainement que professionnellement.

À mes collègues Asapiens et Asapiennes, j'adresse mes plus sincères remerciements, pour la qualité de leur encadrement, leur patience face à mes nombreuses questions, leur confiance ainsi que pour l'expérience enrichissante et pleine d'intérêt que j'ai vécue pendant ces mois au sein de l'entreprise ASAPE

Je remercie tous les interlocuteurs que j'ai eus. Ils ont eu la patience de répondre à mes questions, m'ont permis de recueillir leurs expériences et ainsi approfondir ma maîtrise des sujets abordés dans ce mémoire.

Enfin, je remercie ma famille pour leur soutien moral durant toute la période de cette année en alternance et la période de stage de cet été. Grâce à eux, j'ai su garder le cap et ne jamais me décourager.

Sommaire

Sommaire.....	1
Introduction.....	3
I. Les Prérequis en gestion de projet et limites	12
A. Prérequis en gestion de projet	12
B. Exemple des gestions de projets	21
C. Les outils informatiques.....	32
D. Les Limites : causes d'échec.....	34
II. Les éléments de solution : DATAMINING.....	39
A. Le Datamining : Théorie et Origine.....	39
B. Le Datamining : Notions de base nécessaire à la compréhension	42
C. Le Datamining : Étape de fonctionnement.....	47
D. Le Datamining : Domaines d'applications.....	54
E. Les outils informatiques existants : étude et comparatif	57
F. Le Datamining : Exemple d'utilisation dans un projet	59
III. Mise en place de l'architecture informatique : Interconnexion des outils de GP et de Datamining.....	64
A. Interconnexion d'outils de gestion de projet et de datamining	64
B. Informations et données	65
C. Solutions d'extraction et de stockage des données	81
D. Mise en place de l'ensemble	97
E. Manipulation, Résultats et Travaux connexes	105
F. Et si on ajoutait des informations extérieures ?	106
G. Généralisation à d'autres domaines	107
Conclusion	110

Glossaire	118
Liste des tableaux, schémas et illustrations	126
A. Liste des illustrations	126
B. Liste des tableaux	127
Table des matières	128
Bibliographie	134
A. DATAMINING	134
B. Élément Techniques lié à l'environnement de développement	136
C. Gestion de projet informatique et Bâtiment	137
Annexes.....	138
ANNEXE 1 : Synthèse en Anglais.....	139
ANNEXE 2 : Liste d'indicateurs de suivi de projet.....	144
ANNEXE 3 : REDMINE et les données	145
ANNEXE 4 : Les modèles de données	148

Introduction

« Être toujours plus performant en toujours moins de temps » tel semble être la perfection que cherchent à atteindre les entrepreneurs et jeunes actifs de nos jours.

De plus en plus connectées, grâce aux nouvelles technologies, les entreprises sont confrontées à une concurrence mondiale de plus en plus féroce. Pour mener à bien leurs projets, que ce soit pour un marché national ou non, ils ont besoin de réaliser une veille informationnelle* intense à tous les niveaux : surveillance des concurrents, dialogue avec des acteurs multiples de tous horizons, étude du marché et de ses tendances, etc.

Cette veille génère, ainsi, un amas de données à évaluer et à analyser très conséquent. Beaucoup plus important qu'il y a 10 ans, notamment grâce à Internet qui permet un échange rapide, ce volume doit être traité et décortiqué assez rapidement afin de laisser le temps de réflexion aux décideurs.

En résumé, pour mener un projet à bien en limitant les risques, une entreprise a besoin d'une part de centraliser ses données en les mettant régulièrement à jour et d'autre part, elle a besoin d'outils de filtrage, de calculs et d'analyse performants et efficaces.

Les logiciels de gestion de projet possèdent ces différentes fonctions et représentent une solution adaptée et adaptable aux dirigeants d'entreprises.

Dans le cadre de mon alternance au sein de l'entreprise ASAPE, j'ai eu l'occasion de collaborer à l'amélioration de logiciels et progiciels. L'une des solutions développées par mon entreprise, et sur laquelle je suis intervenu, se nomme ASAPE-GAP. J'ai ainsi choisi d'aborder ce sujet à travers ce mémoire afin d'analyser l'impact de ce type de logiciel sur le fonctionnement d'une entreprise et plus particulièrement sur la gestion des données.

La gestion de projet est une démarche qui permet de structurer et d'organiser différents éléments d'un projet, aussi bien humains que matériaux, afin que celui-ci atteigne son objectif final.

Limité dans le temps, un projet fait l'objet d'une planification à son début et d'un bilan à sa fin afin de savoir si les buts ont été atteints et s'il est survenu d'éventuels problèmes.

Ainsi, le processus pourra être amélioré pour les projets futurs.

De manière générale, un projet, quel que soit son type, son cadre ou son objectif est soumis à 3 catégories de contraintes à des conditions bien précises : le temps, les ressources et les spécifications techniques.

Selon son importance, de plus, il peut avoir plusieurs dimensions : technologique, marketing, juridique, technique (innovation), etc. Cependant il a toujours une visée stratégique.

Pourquoi cela ? Tout simplement parce que selon la manière dont il est mené, un projet peut impacter le présent et l'avenir d'une entreprise. Voici, quelques exemples :

XEROX est une entreprise américaine réputée principalement pour l'invention du photocopieur. Fait moins connu, il s'agit aussi de l'inventeur de la souris et de l'interface graphique* à fenêtre telle que nous la connaissons aujourd'hui.

En 1972, les scientifiques de cette entreprise inventent, entre autres, le réseau Ethernet* et l'interface graphique* tel que nous la connaissons aujourd'hui. Cependant, pour les décideurs, ces découvertes sont trop coûteuses et lentes pour l'époque. De ce fait, elles ne représentent pas de plus-value particulière. Ils les montrent alors à Steve Jobs dans le but de pouvoir négocier un accord avec Apple. Ce dernier, voyant le potentiel de l'interface, convainc les dirigeants de laisser son entreprise exploiter ces recherches, mais ne signe pas de partenariat.

Aujourd'hui, Apple est l'une des entreprises les plus cotées en bourses avec un chiffre d'affaires de 156 milliards de dollars contre 22.4 milliards de dollars pour XEROX pour l'année 2012 (source LesEchos.fr).

Ainsi nous pouvons nous demander, si la direction XEROX avait reconnu le potentiel de ces innovations, et ne les avait pas jugés comme de simples gadgets, aurait-elle pu connaître une autre évolution ?

Une autre entreprise ayant souvent souffert de sa gestion de projet est NOKIA.

Fondé en 1965, c'est une entreprise finlandaise. D'abord conglomérat "touche-à-tout" estimé d'envergure modeste, elle finit par se spécialiser en 1992 dans le secteur de la téléphonie mobile et en devient en 1998, le leader mondial. Fort de son succès, le géant finlandais s'essaie alors à d'autres domaines :

En 2003, il lance un téléphone-console nommé la « N-Gage ». Malgré 3 millions d'unités vendues, cette console n'a pas su s'imposer sur le marché notamment à cause de problèmes techniques et de son système complexe d'achat de jeu. L'entreprise tente alors de lancer une version améliorée en 2004 : la N-gage QD. Cependant, face à l'arrivée des consoles de jeu la Nintendo DS et de la console de jeu Sony PSP en 2005, cette nouvelle version séduit peu le public. Le groupe finlandais décide alors d'arrêter sa production en 2010.

Un autre domaine ayant attiré NOKIA est celui des tablettes numériques. En 2005, la multinationale tente sa chance avec le NOKIA 770 : une tablette tactile dotée de la technologie Wi-Fi qui se distingue par l'utilisation d'un système d'exploitation Linux. Plusieurs nouveautés ont suivi depuis, mais la marque n'a jamais réussi à se positionner face aux leaders du marché Apple et Samsung.

Ces projets, réelles déceptions pour les utilisateurs NOKIA, ont abouti en 2010 à une perte nette de 559 millions d'euros en lieu et en place des 350 millions de bénéfices estimés par les analystes. Il s'agit alors du premier déficit trimestriel de Nokia en 10 ans. L'entreprise perd alors son rang de leader mondial.

Dans ce cas encore, à deux reprises, la mauvaise gestion de projet a été un échec et a coûté son classement mondial à l'entreprise.

Enfin, dans un domaine différent et plus d'actualité : la reprise d'une part des activités industrielles de la société française ARKEMA.

Figure incontournable au niveau mondial, cette société est le premier chimiste français et emploie près de 9000 personnes réparties sur 30 sites industriels et centre de recherches.

En juillet 2012, l'entreprise française décide de se délester d'une de ses filiales, Kem One, en manque de vitesse. Composée de 22 usines de chlore, soude et PVC*, il s'agit d'un choix stratégique du chimiste qui souhaite se recentrer sur des activités plus spécialisées. Elle choisit alors un homme d'affaires américain Gary Klesch comme acheteur.

Initialement décrite comme opération modèle, sans « casse sociale » au vu du contexte socio-économique français, cette cession tourne vite au « fiasco » : La filiale renommée Kem One est en phase de déposer le bilan et ARKEMA est en conflit juridique avec l'américain G.Klesch.

En effet, Kem One déclare un déficit d'exploitation de 56 millions d'euros en six mois : ni les fournisseurs, ni les Urssaf* ne sont payés et les employés craignent pour leurs postes. Le nouveau propriétaire affirme que la société française ne lui a pas communiqué les bonnes informations sur la santé financière des sites alors que l'accusé affirme l'avoir fait.

Aujourd'hui, les médias, d'une part, s'interrogent sur la véracité des propos de l'américain, surtout connu comme « requin des affaires », voire même « de prédateur » pour les industries en situation difficile. Grâce à ses acquisitions, il possède d'ailleurs une fortune, estimée 2013, à 5 milliards d'euros. D'autres parts, les journalistes reprochent au chimiste de ne pas avoir été plus regardant sur le profil du repreneur.

Toujours en cours de négociation, cet exemple met en exergue l'importance des données dans la prise de décisions : si M.Klesch est réellement une victime et qu'il n'a pas eu connaissance des informations avant la transaction, aurait-il repris la filiale s'il les avait eues ? Si l'entreprise ARKEMA avait mieux enquêté sur son acquéreur, aurait-elle réalisé la cession avec lui ?

Toutes ces différentes situations auraient pu être évitées si les entreprises avaient eu, à l'époque, toutes les informations en main lors de la planification ou de la réalisation du projet. C'est d'ailleurs afin d'éviter de tels échecs que les outils de gestion de projet ont dû constamment évoluer. Aujourd'hui, les décideurs disposent d'un grand panel de solutions informatiques dédiées à la gestion de projet et au décisionnel dans le but de limiter le risque d'échec au maximum dans leur domaine.

Afin d'explicitier et bien comprendre le besoin des entreprises actuelles en outils de gestion de projet et l'impact de ces derniers dans l'économie actuelle, nous allons d'abord aborder l'évolution et l'historique de ces outils.

La gestion de projet, tel que nous la connaissons aujourd'hui, a été formalisée dans les années 1950. Cette époque est marquée par la sortie de la Seconde Guerre mondiale et la nécessité de tout reconstruire. Les états mettent alors en place des projets complexes avec une main d'œuvre fortement réduite. Pour cela, ils adoptent un mode de fonctionnement

proactif. Ce dernier marque une rupture majeure avec les modèles antérieurs qui jusque-là étaient dans le réactif.

Des solutions et outils ont été développés afin d'offrir aux dirigeants de meilleurs moyens de contrôle sur des projets d'envergures. En classant les tâches et décisions selon les ressources nécessaires et leur importance stratégique, les responsables anticipaient celles qui étaient « légère à mettre en place » et retardaient au maximum celles qui étaient plus lourdes et avaient le plus d'impact. Le modèle de PERT* et celui du chemin critique* ont d'ailleurs vu le jour à cette époque.

Au vu de leur succès, ces concepts se sont rapidement répandus dans les milieux de l'industrie. En effet, les entreprises, afin de rester concurrentielles, avaient besoin de développer leur projet de plus en plus rapidement, afin d'avoir un « temps de mise sur le marché » de plus en plus court entre deux mêmes projets et ainsi toujours rester dans l'actualité.

Le temps, devenu ainsi un axe majeur de la concurrence, a entraîné les entreprises à adopter un rythme d'innovation assez soutenu et donc à changer, améliorer voir réviser totalement leurs méthodes de gestion.

Cette stratégie a eu deux impacts :

- le premier est de dresser une barrière importante pour l'entrée de la concurrence sur leur marché. Peu d'entreprises avaient alors les ressources nécessaires (matérielles et financières) pour suivre la cadence.
- Le second est le déclassement de certaines sociétés de leur propre produit. Afin d'être le premier sur le marché, il faut être le premier à proposer aux prospects ce qu'ils voudraient éventuellement. Leur logique n'est alors plus de répondre à un besoin, mais d'en créer en proposant des nouveautés. Leur démarche est donc à ce niveau-là purement proactif et surtout très risquée, car rien ne garantit le succès du nouveau produit.

Cette course à l'innovation a entraîné parallèlement une course à la connaissance : connaissance de la tendance actuelle, des attentes des consommateurs et de l'état de lieux des marchés locaux, nationaux et internationaux. Or la connaissance est issue de données.

Ces dernières doivent être collectées, triées, filtrées puis analysées. Et selon l'importance du projet, le volume peut être très important.

Si au début avoir du personnel et de bons outils de gestion et de production suffisait, aujourd'hui ce n'est plus le cas : à l'heure où l'information traverse le globe en une milliseconde, l'être humain a besoin d'être aidé d'outils particuliers pour tenir la cadence.

En effet, les entreprises, tous secteurs confondus, se sont vite heurtées aux limites du genre humain face à la rapidité d'échange et d'évolution du monde qui les entoure.

Ainsi, à la fin du XXe siècle, parallèlement à l'évolution de la gestion de projets, il y eut celle des outils de restitution et de collecte de données, soit les outils d'informatique décisionnelle.

L'informatique décisionnel ou Business Intelligence est l'ensemble des moyens, méthodes et outils, destiné aux décideurs, qui permettent de collecter, modéliser et restituer les données. C'est une informatique dédiée à la prise de décision : elle fournit une vue d'ensemble d'une activité de l'entreprise afin que les responsables puissent faire des choix stratégiques.

C'est un élément qui a évolué depuis sa création et qui, aujourd'hui, est dominant dans une majorité des systèmes d'informations.

Les premiers outils d'informatique décisionnelle ont vu le jour dans les années 80. Peu évolutifs, mais très coûteux, ils monopolisaient beaucoup d'informaticiens. En effet, ces derniers travaillaient à plusieurs afin de fournir toutes les semaines des tableaux de données assez complexes, mais malheureusement pas assez détaillés pour répondre au besoin des décideurs. Néanmoins premier du genre, ils apportent une vision globale de l'activité de l'entreprise.

Dans les années 90, la technique s'améliorant, le « Reporting* » apparaît. Il s'agit d'un ensemble d'outils permettant une collecte et un accès simplifié des données. Le mode de restitution est alors principalement graphique (histogrammes et graphes), car jugé alors plus explicite.

Le but alors est de rendre compte périodiquement à la direction, des performances (ou de l'avancement) de projets et ainsi valider des stratégies ou au contraire effectuer des changements conséquents.

À partir des années 2000, le besoin change. Avec la démocratisation d'internet (la population de la toile passe de 250 millions à 500 millions entre 1999 et 2002), les outils évoluent afin de mieux gérer le grand volume de données. Les sociétés commerciales ont alors besoin de nouveaux indicateurs face aux nouvelles tendances (achat en ligne, consultation d'avis, création de communautés en ligne, etc.) afin de mieux cibler leur clientèle.

Les entreprises, devant être sur plusieurs fronts à la fois, ont eu besoin d'outils de pilotage adaptés. De nouveaux modes de restitution de données ont été développés apportant à la fois une vision globale de l'entreprise (ex. : planning des collaborateurs sur toute l'année, tous projets confondus), mais aussi une vision détaillée pour un projet donné (ex. : Liste des collaborateurs disponible pour collaborer ou ayant collaboré sur une tâche précise durant une période déterminée).

Le système d'information et la manière dont il restitue l'information doivent être de plus en plus agiles. Ainsi, il faut pouvoir suivre l'activité de son organisation, son évolution avec des critères que l'on n'avait pas avant, tout en continuant à réaliser des projets rentables. Il faut d'une part pouvoir identifier les points qui posent problème à tout instant afin d'y remédier rapidement et d'autre part pouvoir piloter son activité et ses projets par la mesure des performances (grâce à des indicateurs).

La Business Intelligence a été pendant ces 20 dernières années l'outil de prédilection des décideurs. En effet, grâce à ses outils analytiques de statistique, il fournit une information managériale sur la situation de l'entreprise et permet ainsi de dégager des axes de progression pour cette dernière.

Description et explication sont les mots d'ordre pour des cycles de décision et d'actions asynchrones*.

La problématique aujourd'hui n'est plus d'être dans le descriptif et l'explicatif, mais surtout dans le prédictif. Les données analysées doivent permettre de prédire des tendances avec une marge de risque respectable. C'est le rôle du Datamining.

Dans la gestion de projet, comme nous l'avons vu précédemment à travers les exemples de XEROX, NOKIA et ARKEMA, les mauvaises décisions faites au mauvais moment ont des conséquences importantes : carrière ruinée, perte d'emplois et de capitaux, etc.

À la lumière de ces réflexions, nous pouvons ainsi nous demander :

« Dans quelle mesure le Datamining peut-il permettre de limiter les risques dans la gestion de projet ? »

Définissons d'abord cette notion :

Le Datamining, ou fouille de données est un processus* d'étude et d'analyse de grand volume de données. Il vise à découvrir de nouvelles corrélations, caractéristiques ou tendances que l'on n'aurait pas forcément vues avec des méthodes traditionnelles.

Créé en 1980, il est une conséquence de l'évolution du paysage informatique et économique, à savoir : la multiplication de bases de données volumineuses et difficilement exploitables pour les entreprises n'ayant pas assez de moyens.

Si l'informatique décisionnelle permet de constater un fait tel que le chiffre d'affaires réalisé sur une période pour un produit donné, le Datamining permet quant à lui de classer (voir prévoir) des faits et déterminer des explications en révélant des paramètres.

Sa plus-value réside dans le fait qu'il est indépendant du secteur étudié : il peut être appliqué aussi bien dans les secteurs de la grande distribution, les finances, les ressources humaines, l'informatique ou encore le bâtiment.

Pour répondre à notre problématique, nous allons tout d'abord, nous pencher sur la gestion de projet : qu'est donc ? Quels sont ses buts ? Dans quels cas est-elle utile ? Nous étudierons la notion de projet tel qu'elle existe dans les entreprises actuelles afin de bien comprendre les enjeux qui ont découlé et surtout ses impacts sur le monde entrepreneurial.

Nous aborderons aussi ses limites : pourquoi NOKIA, et XEROX n'ont-elles pas rencontré le succès dans leur projet ? Quels ont été les éléments possibles qui ont abouti aux échecs que l'on a cités plus tôt ?

Cela nous amènera ensuite à réfléchir sur les solutions possibles pour pallier à ces limites : s'il en existe qu'elles sont –elles ? Dans quelle mesure le Datamining en fait-il partie ? Et surtout comment un domaine aussi jeune peut-il être porteur d'espoir ? Nous tacherons d'éclaircir cela en présentant d'abord les fonctions du Datamining et son mode de fonctionnement. Ensuite, nous proposerons une infrastructure alliant les deux concepts et permettant de résoudre notre problématique.

I. Les Prérequis en gestion de projet et limites

La gestion de projet est spécifique à chaque domaine, cependant, il est possible de la décrire dans les grandes lignes. Après un rappel de sa définition et de son fonctionnement, nous en aborderons des exemples dans des domaines bien précis.

Une fois les notions clés éclaircies, nous comparerons deux solutions informatiques dédiées à cela. D'une part, nous mettrons ainsi en évidence la multiplicité des offres et donc des demandes liées au métier. Et d'autres parts, nous éclaircirons la gestion informatique des entreprises aujourd'hui. Cela nous permettra ainsi de déterminer d'où viennent les contraintes et limites actuelles.

A. Prérequis en gestion de projet

1. Rappel de la définition

« La gestion de projet est une démarche visant à organiser de bout en bout l'ensemble d'un projet », tel est la définition que nous avons donnée précédemment. Mais qu'elle en est la signification ? Et surtout pourquoi ce domaine est-il le sujet de tant de recherche, d'articles et d'études de nos jours ?

À la différence des opérations classiques dans une entreprise qui sont répétitives, un projet est innovant et unique.

En effet, l'achat de cartouche d'imprimantes est une opération assez classique : à chaque fois qu'elle sera exécutée, elle suivra un schéma classique :

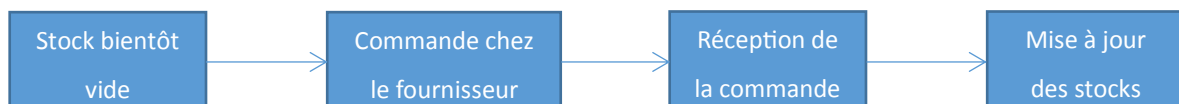


Illustration 1 : Opération d'achat de cartouche d'imprimante

Un projet est différent : il comporte toujours une notion de nouveauté et de changement.

Exemple : une fusion de service, une réorganisation interne, une construction d'immeuble, etc.

Il revêt plusieurs dimensions et peut en impacter plusieurs. Délimité dans le temps, il possède donc un début et une fin. Quelle que soit sa nature, il fait l'objet d'une

budgétisation et la mobilisation de ressources matérielles et humaines. Il nécessite parfois des personnes aux compétences multiples et souvent étrangères les unes aux autres.

Si elle est si importante dans les entreprises aujourd'hui, c'est parce que la gestion de projet intervient et impacte tous les niveaux : de la hiérarchie aux employés.

Elle permet de formaliser et tracer la mise en place des processus* et phases des projets : elle organise et contrôle le travail de tout le monde.

De plus, c'est un outil décisionnel : selon les éléments qu'il met en exergue, les décideurs ne prendront pas les mêmes décisions et c'est justement là que se trouve le « nerf de la guerre ».

2. Buts de la gestion de projet : Importance et enjeux

« Innover plus, innover vite et innover mieux », tel est la règle à suivre pour les entreprises aujourd'hui. En effet, quel que soit leur domaine d'activité, face à la concurrence, les entreprises se doivent d'avoir une grande capacité d'adaptation. Pour cela, leur gestion de projet doit être optimale : temps de réaction rapide, gestion des risques, etc.

Le respect de ce "tryptique*" fait partie des enjeux de la gestion de projet : cette notion correspond au pourquoi est réalisé le projet, sa finalité. Il peut être financier, commercial, voire même humain, mais il est toujours essentiel.

Comprendre les enjeux d'un projet, permet de mieux adapter la gestion qui sera appliquée. Par exemple, travailler avec une filiale d'une multinationale, garantit, si le client est satisfait, une publicité gratuite auprès des autres filiales de la société et ainsi un panel de prospect non négligeable. Si l'entreprise cherche à s'étendre alors, l'enjeu dans ce cas est important.

Selon la taille des projets, il peut être facile ou non de coordonner le travail de tous. À l'heure où le télétravail et les déplacements clientèles sont légions, centraliser les données relève parfois du miracle. C'est là aussi qu'intervient la gestion de projet : elle gère les données, permet une centralisation et surtout une mise à jour de ces mêmes données.

Or sur quoi se basent les décideurs au moment de prendre leurs décisions ? Sur ces données-là justement. Si elles sont obsolètes, elles entraîneront forcément de mauvais choix et entraîneront donc de graves conséquences.

Les experts estiment aujourd'hui que 60% des coûts des projets relèvent juste de la synchronisation entre acteurs : cela implique que seuls 40% des coûts sont réellement utilisés pour le projet en soi.

Les raisons sont essentiellement liées au manque d'outils adaptés : la plupart des entreprises utilisent les mails seuls et les outils bureautiques pour piloter leur projet. Or cela n'est plus suffisant, surtout avec la multiplicité des supports (Tablette, Smartphones, PC HYBRIDE) qui permettent une mobilité et un échange continu.

En effet, ces outils n'offrent pas de vision transverse des activités de l'entreprise facilement. De plus, l'information est difficile à consolider, à récolter, elle est non accessible à tous les acteurs et surtout répartie sur plusieurs outils (serveur, machine locale). De ce fait, le format des données est hétérogène. Cela ne peut qu'induire l'inefficacité du comité de pilotage sans compter sur la saturation des boîtes mails, la perte de document et d'information possible, les problèmes de traçabilité et de fiabilité, la non-gestion des versions, etc.

Ainsi les buts de la gestion de projets et de ses outils sont de permettre aux entreprises d'atteindre leurs objectifs et de remporter leurs enjeux stratégiques, de pouvoir faire de bonnes estimations selon des données mises à jour et suivre l'évolution des projets en temps réels tout en diminuant les coûts de synchronisation des acteurs et améliorant la pertinence du pilotage.

3. Les acteurs principaux

Dans la gestion de projet, il y a deux acteurs principaux le Maître d'ouvrage (MOA) et Le Maître d'œuvre (MOE). Nous allons les définir.

Le MOA est la personne pour qui le projet est réalisé. Il s'agit donc du futur propriétaire de l'objet du projet. À ce titre, c'est donc lui qui spécifie le besoin et qui garantit la stabilité de ce dernier. Il valide donc le produit lorsque ce dernier est réalisé et le paye. De ce fait, c'est à lui que revient le choix du MOE et des entreprises prestataires. Il est le premier maillon du projet.

Le MOE est celui qui réalise : responsable et garant des moyens et de l'organisation nécessaires au projet, il en supervise le déroulement et possède les compétences techniques nécessaires pour la réalisation du projet.

Il travaille pour le MOA et lui fournit l'objet de sa commande à la fin du processus* de réalisation. Il peut aussi être une personne morale (société) ou une personne physique (autoentrepreneur)

Un autre acteur principal est le chef de projet : il peut être le MOE lui-même ou un de ses employés. Il dirige l'équipe en charge du projet : en charge de la synchronisation, de l'avancement, du respect des délais et budgets, il possède à la fois des compétences techniques et de management. Il se focalise sur le traitement et l'anticipation des difficultés rencontrées et s'assure que toutes les parties réalisent leur travail correctement.

Il est aussi l'intermédiaire entre l'équipe et le comité de direction à qui il fait son rapport concernant les projets à sa charge.

Les acteurs secondaires, mais non de moindre importance, sont l'équipe du MOA et l'équipe du MOE (assistants, conseillers, artisans, etc.), les sous-traitants ou entreprises partenaires dont on a besoin pour un domaine spécifique (conseil expertise dans une spécialité donné) et dont la qualité de travail influe directement sur le projet.

4. Fonctionnement : Données et indicateurs

Un projet possède un fonctionnement par étape : il peut être découpé en Lot, sous projets ou encore par tâches. Cela facilite la maîtrise de la complexité et de la planification. Confié au chef de projet, il est soumis à des contraintes en termes de temps et de budgets. Globalement, un projet peut être découpé en 4 étapes /phases :

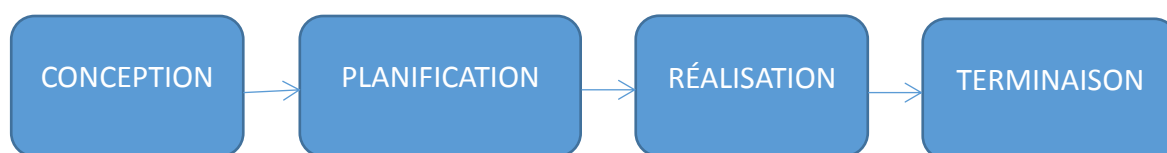


Illustration 2 : Étapes de gestion de projet

a) CONCEPTION

Cette première phase correspond à l'étape de l'étude du projet et à son lancement. C'est à ce niveau qu'est menée la première réflexion sur le projet en lui-même : pourquoi ? Qu'elles sont les enjeux stratégiques ? Quelles sont les opportunités qu'il offre ? Etc.

S'il est jugé intéressant, alors il lui sera attribué des ressources humaines, financières et techniques : on détermine alors l'organisation à suivre, le planning, l'environnement technique, etc.

C'est aussi à ce stade que l'on a l'expression du besoin et son étude : c'est la phase de définition des attentes à combler, des objectifs à atteindre, du périmètre, des éléments importants et de moindre importance. Cette phase nécessite la collaboration du MOE et MOA.

b) PLANIFICATION et RÉALISATION

Une fois les besoins éclaircis et déterminés, le MOE se charge de coordonner son équipe en attribuant les tâches et en fixant les délais.

La phase de planification passe par de nombreuses étapes, d'abord une planification globale, puis l'étude des coûts et des délais dans le détail, puis l'attribution des tâches, voir même le recrutement de ressources extérieures s'il le faut.

La réalisation est considérée comme la phase générant le plus de coûts : elle consiste à la mise en place de l'organisation choisie au préalable, l'exécution du travail et aussi la résolution de problèmes lorsqu'ils surviennent. Cette phase est aussi connue sous le nom de pilotage coûts-délais-spécifications, le « tryptique »* ou « triangle fou »* du chef de projet.

Sur toute la période de vie d'un projet, selon le secteur, il existe des moments clés qui influent directement sur la réussite ou l'échec du projet. Ce sont des phases qui, du fait de mauvaises décisions, peuvent entraîner des surcoûts, voir des malfaçons et ainsi des mauvaises performances. Il est donc important que le client contrôle les livrables* qu'il reçoit, et que le maître d'œuvre garde un œil sur la qualité du travail de ses équipes.

La phase de réalisation nécessite une collaboration proche de toutes les parties afin d'éviter toutes les causes possibles d'insuccès du projet.

c) TERMINAISON

C'est la livraison au client. Connue sous le nom de recette dans certains domaines, cette phase correspond à la remise du produit entre les mains du commanditaire. Propriétaire alors, ce dernier peut contrôler, critiquer le rendu ou l'accepter.

C'est alors la fin du projet.

Dès les premières livraisons, le client se doit de réaliser des contrôles continus afin de s'assurer de la conformité du résultat avec la commande de départ et les spécifications établies.

Une fois tous les éléments de désaccord réglé, un procès –verbal de réception définitive est signé : le contrat de commande prend alors fin.

d) BILAN

Cette dernière phase ne concerne que le prestataire : il s'agit d'un bilan post-projet qui permet de mettre en avant les réussites, mais aussi les échecs.

Grâce à ce moment de remise en question et d'introspection, le prestataire permet à son équipe de partager leur expérience et leur ressenti. Tout le monde profite de ça et ainsi l'entreprise améliore ces méthodes de travail et prépare mieux ses futurs projets.

5. INDICATEURS DE SUIVI DE PROJET : KPI ou ICP

Les Indicateurs clés de performance, dit ICP ou KPI pour Key Performance Indicateur, sont un élément important pour la direction. Témoinant de l'avancée et du taux de réussite d'un projet, ils mettent aussi en relief les « points rouges » à analyser : quelle est la raison de l'échec ? Pourquoi ? Comment aurait-on pu l'éviter ? Et surtout comment éviter que cela se reproduise ?

a) DÉFINITION

Un indicateur est une information à un instant t qui permet de mesurer une situation. C'est un outil d'aide à la prise de décision puisqu'il va apporter des éléments de réflexion au décideur qui va ensuite prendre une décision en conséquence de cela.

Les KPI sont des indicateurs de mesure de la performance des projets. Utilisés par le chef de projet, ils concernent généralement le suivi budgétaire, la consommation des ressources et le respect des délais.

Ils permettent un état des lieux, à savoir, comparer les progrès réalisés avec les objectifs et le planning initiaux. Grâce à cela, le comité de pilotage possède une connaissance précise de l'état du projet ce qui améliore leur prise de décisions. En effet, comme les indicateurs représentent un ratio ou une moyenne, ils permettent de comparer le prévisionnel et le réalisé en terme de temps, de consommation de budget et de ressources.

Associés à une stratégie, ils représentent un élément de poids lors des comités de pilotage grâce à leurs deux fonctions principales : **aide à la décision et analyse de performance**.

b) CHOIX et ENJEUX DES KPI

Bien qu'importante, leur sélection est difficile pour la gestion de projet, car cela nécessite plusieurs éléments. Tout d'abord, il faut fixer honnêtement l'objectif stratégique du projet, c'est-à-dire sa vocation principale : a-t-il pour but de développer le chiffre d'affaires ? Ou améliorer la notoriété ou encore conquérir des parts de marché bien spécifique ?

Ensuite, il faut déterminer les objectifs secondaires qui vont permettre d'atteindre les objectifs principaux.

C'est grâce à cela que la sélection des KPI est pertinente puisqu'elle se fait en concordance avec les objectifs véritables de la société : cela permet de se focaliser sur les axes importants du projet qui préoccupent réellement les décideurs.

Une fois la sélection réalisée, il faut que le calcul soit juste. Un indicateur est dit juste et bon s'il délivre une information en temps réel et s'il incite à l'action. Connaître l'état des lieux d'une situation passé est trop tard. Les données doivent être mises à jour, vérifiées, validées et classées afin qu'on puisse y accéder rapidement lorsque l'on en a besoin.

Les KPI sont rarement présentés tel quel : ils présentent certes un état des lieux pertinent, mais cela ne suffit pas, car ils ne portent chacun que sur un seul élément à la fois (délai, budget, qualité, etc.). Ce qui est intéressant, c'est la vision transverse qu'ils offrent tous ensemble à un instant t. Regroupés ensemble dans des graphiques et tableau de bord*, les KPI alimentent donc le Reporting* constituant ainsi une part essentielle des outils de pilotage de projet.

Selon la politique de l'entreprise considérée, ils peuvent être utilisés comme outils de communication et d'information auprès d'investisseur ou de prospect, ou encore comme outils de motivation auprès des équipes. Ce sont aussi des outils d'évaluation et de diagnostic qui permettent de faire le bilan d'une action menée au sein d'un service ou organisme et d'en tirer un axe de progression.

Multi-usages, les KPI sont donc les acteurs incontournables du progrès continu de l'entreprise.

c) EXEMPLE : le ROI, Return On Investment ou Retour Sur Investissement RSI

C'est l'un des indicateurs primordiaux et communs à toutes les entreprises, car il permet d'évaluer le rendement d'une somme investie. Dans un contexte mondial de crise, il est important que les investissements soient rentables afin d'assurer la pérennité des entreprises et investisseurs. Le ROI permet justement d'évaluer cela. En déterminant quels sont les gains apportés par une opération donnée, nous pouvons calculer son pourcentage de rentabilité grâce à la formule suivante :

$$\text{ROI} = (\text{Gains} - \text{Montant investissement}) / \text{Montant investissement}$$

Cet indicateur peut aussi bien s'utiliser à priori d'un projet ou à postériori. Dans le premier cas, il revêt une dimension prédictive. Il peut alors servir de critères de choix pour un investisseur donné qui cherche à sélectionner un projet parmi tant d'autres. Dans le second cas, il permet de mesurer la rentabilité réelle d'un projet et de trouver les points à améliorer pour une meilleure performance. En gestion de projet, le ROI prédictif peut aussi être comparé au ROI effectif afin de savoir si le processus* de prédiction est efficace, si la marge de différence est acceptable ou non et surtout comment améliorer la performance pour les projets futurs.

C'est un ratio économique important donc pour l'entreprise, quel que soit le type de projet, car il influe directement sur la rentabilité de l'entreprise. C'est donc un indicateur de mesure de réussite financière d'un projet.

d) EXEMPLE : KPI de COÛT

L'un des axes majeurs d'un projet est le coût de celui-ci. Le respect du budget est important en théorie, mais cela diffère souvent de la réalité, où il est souvent dépassé.

Pour contrôler cet axe et l'ajuster, un indicateur est calculé : c'est l'indicateur « Écart de coût de projet ».

Sa formule est la suivante :

$$\text{KPI_Cout} = (\text{Coût réel} - \text{coût prévisionnel}) / \text{coût prévisionnel}$$

Grâce à cela, le responsable peut ajuster ses investissements en réallouant les montants là où ce serait bénéfique pour le projet ou, si besoin, demander une augmentation du budget à la direction.

e) EXEMPLE : KPI de DELAI

« L'écart de durée » est un indicateur tout aussi important. Mesurant la consommation réelle du temps par rapport au temps prévu, il se calcule comme ce qui suit :

$$\text{KPI_Délais} = (\text{Durée réelle} - \text{Durée initiale}) / \text{Durée Initiale}$$

Ce calcul peut aussi bien s'appliquer à une tâche spécifique, à une phase complète (Conception, Réalisation) ou encore au projet dans son intégralité.

Grâce à cet indicateur, les responsables peuvent plus simplement identifier alors les tâches qui ont consommé le plus de temps et aboutis à un dépassement des délais et ainsi identifier les « points rouges » à améliorer dans leur gestion.

f) EXEMPLE : KPI de RESSOURCES

Le KPI de délai donne une information objective à un instant t : lorsque l'on observe une consommation de 110% du temps imparti, on ne sait pas si le problème à une cause humaine ou matériel.

Ainsi afin d'apporter une vision plus complète, il est important de coupler cet indicateur avec celui de la « productivité des ressources humaines de projet » soit le KPI de Ressources.

Ce dernier offre une mesure de la productivité des membres d'une équipe. Il mesure donc en temps réel, le nombre de jour-homme (unité de mesure correspondant au travail d'une personne pendant une journée) consacrés à une tâche donnée au regard du pourcentage de réalisation de cette même tâche.

Ce rapport se fait en deux temps. Le ratio est d'abord calculé comme ce qui suit :

$$\text{KPI_Ressources} = \text{nombre de jour-homme consacrés} * \% \text{ de réalisation de la tâche}$$

Puis il est comparé au nombre de jour-homme planifié pour atteindre ce même % de réalisation de la tâche.

Il permet ainsi d'évaluer soit un retard soit une avance sur le planning ou au contraire un respect de ce dernier.

Ainsi, si une tâche possède un indicateur "écart de durée" et un "indicateur productivité des ressources", tous deux importants, elle est alors considérée doublement pénalisante pour le projet. Les décideurs et chefs de projet peuvent ainsi plus se focaliser sur la recherche de solution pour optimiser le travail pour cette tâche-là, plutôt que perdre leur temps à trouver d'où vient le problème.

Les KPI sont donc des éléments de gestion facilitateur puisqu'ils permettent de cibler rapidement les problèmes. Alliant une dimension technique et business, ils permettent en effet d'avoir à la fois une vue d'ensemble des objectifs et des mesures des actions menés dans ce but. Indispensable au chef de projet, ils constituent surtout le cœur du plan d'amélioration continue des entreprises.

B. Exemple des gestions de projets

Comme abordée précédemment, la gestion de projet possède un fonctionnement standard. Cependant, selon le métier de l'entreprise considéré, elle peut subir quelques variantes. Nous allons ainsi aborder l'application de la gestion de projets dans deux métiers éloignés l'un de l'autre : l'informatique et le Bâtiment.

1. Gestion de projet informatique

a) FONCTIONNEMENT

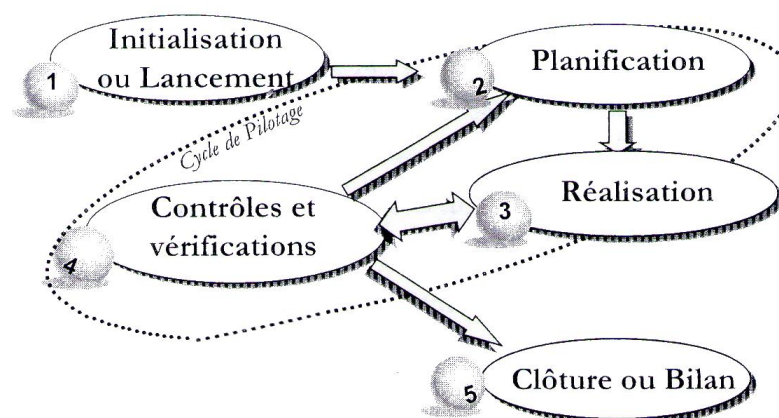


Illustration 3 : les Phases d'un projet

Dans son mode de fonctionnement, la gestion de projet informatique (GPI) est similaire au standard détaillé précédemment, à quelques différences près.

Les 4 phases appliquées au domaine informatique se présentent comme qui suit :

- **La Conception dite aussi « phase d'initialisation »** : les objectifs et besoins sont formalisés sous la forme d'un cahier des charges*. Ce dernier est un document qui spécifie les besoins fonctionnels de l'application à développer. Essentiel, il est la référence au cours de la vie du projet. Il peut être soit élaboré par le prospect lui-même s'il en a les compétences, soit issue de la collaboration entre le prospect qui exprime son besoin et le prestataire qui apporte ses conseils en terme de faisabilité technique et financières.

Cette première phase est aussi appelée « l'avant –vente » : il s'agit de l'étape de négociation avec le prospect ou alors du traitement d'un appel d'offres. Le projet n'est pas signé, mais initié. Concrètement cela implique, que le prestataire a su proposer une stratégie qui met en œuvre des facteurs clés de réussite pour combler les attentes du client :

- Les charges sont estimées, le budget ventilé et le compte d'exploitation prévisionnel établi.
- Les listes des acteurs et des ressources ont été fixées et les domaines de responsabilités de chacune des parties aussi
- Le planning prévisionnel de réalisation et de livraisons a été réalisé et validé
- Les jalons structurants le projet ont été posés
- Techniquement, l'architecture, les méthodes, les normes qui seront utilisées ont été choisies
- Le projet a été découpé en une liste de modules

En résumé, cette phase d'initialisation est celle où l'on présente globalement le projet afin d'en poser les bases. Tous les détails sont abordés afin de réaliser une solution informatique répondant parfaitement au besoin du client.

- **La Phase d'encadrement de la réalisation :**

Cette étape importante a pour objectifs de contrôler l'avancement du projet au jour le jour et de veiller à ce qu'il atteigne les objectifs fixés dans le cahier des charges*. Cela passe par une multitude d'actions notamment, veiller au respect des exigences et des

engagements, attribuer des missions adaptées à chaque collaborateur et gérer les activités de l'équipe projet.

Cette phase comporte trois sous-phases :

- La première est la Planification : le cahier des charges* contient les dates de livraison intermédiaires au client et leur contenu. Il offre donc une planification première des modules et leur ordre de réalisation. Cependant, ce n'est pas suffisant. C'est par la suite au chef de projet d'analyser avec son équipe les modules et donc de les découper en tâches. Chaque tâche fait l'objet d'une planification en interne : qui va réaliser quoi ? Pour quand ? Qui va tester ? Cette phase peut faire l'objet de la réalisation d'un diagramme de Gantt afin que toute l'équipe soit consciente des échéances et qu'elle ait une vision globale, mais détaillée du temps et des tâches à réaliser dans les délais impartis.

- La seconde la Réalisation : Dans les métiers de l'informatique, c'est la phase de développement. Les équipes développent le logiciel ou module commandé, testent leur travail et finissent par livrer une version au client à une date préalablement fixée. Il s'agit, pour le chef de projet, d'une phase de coordination et de suivi des équipes et des travaux.

- la troisième est le Contrôle et Vérification :

Souvent inclus dans la phase de réalisation, cette étape peut-être considérée comme une phase à part entière en informatique. La qualité absolue* voudrait que l'on teste tout ce que l'on a développé. Or malheureusement, ce n'est pas toujours le cas, car les développeurs sont souvent focalisés sur le côté technique et oublient le côté fonctionnel (ergonomie, graphisme). De plus, ils ne sont pas du métier du client, ce qui leur rend difficile de comprendre l'importance de certains éléments par rapport à d'autres. C'est justement pour pallier à cela, que cette phase existe : à chaque livraison d'un module ou d'une version, c'est un groupe d'utilisateurs finaux qui vont tester l'application. Ils vont ainsi déterminer si elle correspond à leur besoin, si les temps de chargement sont raisonnables. En résumé, ils permettent de vérifier l'adéquation entre le besoin réel, celui théorique décrit dans le cahier des charges* (point de référence lors des développements) et le rendu du prestataire.

Ces tests peuvent intervenir pendant la réalisation, ou de manière ponctuelle à chaque livraison d'une version. C'est alors ici qu'apparaissent de nouvelles tâches à planifier auquel on n'avait pas pensé au début. C'est la boucle du cycle de pilotage.

Cette phase de vérification ne revêt pas qu'une dimension technique. Elle permet plus globalement de situer l'avancement réel du projet à intervalles réguliers, au regard des objectifs (charges, délais et qualité des livrables*). De plus, c'est l'occasion aussi de mesurer les écarts, en analyser les causes et ainsi mettre en œuvre des plans d'action correctifs et préventifs nécessaires.

- La Terminaison dit aussi « Phase de BILAN ou Clôture »:

Lorsque le projet a été réalisé dans son intégralité, tous les besoins comblés et que tous les bugs (dysfonctionnements informatiques) corrigés, la recette peut être signée.

C'est la phase de « recettage » : elle a pour but de valider le bon fonctionnement de l'application par une série de tests clients. Ces tests seront menés dans un environnement similaire à l'environnement technique final afin de rester au plus près du cadre client.

Si tout s'est bien passé du côté du client, il s'ensuit un basculement en production de l'application : le produit commandé est installé chez le client afin que celui-ci puisse l'utiliser tous les jours. C'est une étape sensible et très importante surtout si le logiciel est destiné à un grand nombre de personnes. En effet, c'est ici que l'on voit si tout a bien été pensé pour une sollicitation quotidienne et massive de l'application (plusieurs utilisateurs).

Enfin, la dernière étape intervient, il s'agit de la Gestion du changement. Le prestataire informatique assiste et accompagne les utilisateurs dans leur prise en main du nouvel outil.

En interne, le projet n'est pas fini. En effet, il reste à faire un bilan interne de ce qui s'est passé. Sous la forme d'une réunion avec toute l'équipe, cette étape est cruciale pour l'entreprise, car elle a pour objectif principal d'apporter un axe de progression à cette dernière. En effet, le but de cette phase bilan est d'abord de synthétiser les acquis sur le projet pour capitaliser les astuces et les procédures. Ensuite, c'est l'occasion de trier, regrouper et classer les documents représentatifs et les enregistrements qualité. Enfin, elle permet de rédiger le dossier de bilan qui servira ultérieurement pour une diffusion auprès d'autre équipe ou des présentations.

b) PROBLÉMATIQUE PROPRE À LA GPI

Une des problématiques propres à la GPI est que selon le type de projet informatique, les intervenants peuvent être très nombreux. En effet, si l'on prend l'exemple d'un projet qui vise à mettre en place un portail intranet pour toute l'entreprise, les profils intervenants seront d'une grande diversité (département financiers, RH, communication, etc.). A chacun ses besoins propres, ses exigences et ses informations. Il faut donc être capable d'adapter sa communication en fonction de chaque interlocuteur et surtout, en tant que prestataire, être capable de gérer ce grand amas d'informations et donc de données. La gestion doit être optimale afin de garder une trace des informations (qui a dit quoi ? quand ?) et surtout afin de les analyser et déterminer les besoins communs et les besoins propres à chaque fois.

Dans ces cas-là, le MOE demande à avoir un responsable MOA. Ce dernier sert d'intermédiaire entre le prestataire et tous les interlocuteurs. Cela permet une centralisation de l'information et des besoins au niveau d'une seule personne. Cette dernière pourra ainsi vérifier qu'il n'y a pas de contradiction entre les demandes des services et surtout coordonner la réalisation des modules selon les bonnes consignes en prenant compte les objectifs de chacun et les difficultés éventuelles.

c) INDICATEURS

Les indicateurs de coût, de délai et de ressources citées précédemment sont très utilisés en informatique. Cependant, ils sont rattachés à une dimension globale et liés au chef de projet directement : avons-nous dépassé le budget prévu ? Sommes-nous dans les délais ? Avons-nous consommé trop de ressources par rapport à ce qui était prévu pour cette tâche ?

Il existe d'autres indicateurs tout aussi déterminants, mais liés à la production directement et donc à l'équipe : les indicateurs de suivi de projet.

Ce sont des indicateurs qui portent sur les tâches attribuées aux développeurs. Par exemple, les charges consommées ou encore le reste à engager (RAE). Ce dernier correspond à l'estimation de la charge du travail restant pour finir la tâche. Il permet de savoir, si à telle date, l'équipe est dans les temps ou en retard ou en avance. À la différence du KPI de ressources qui se base sur les nombres de jour-homme, le RAE ne prend en compte que la tâche à réaliser (ni les ressources, ni le temps).

D'autres indicateurs de suivi de projet sont présentés en Annexe 2 : que ce soit pour une tâche, un projet, une équipe ou un individu, ils permettent de comparer les prévisions à ce qui a été effectivement réalisé.

Ainsi, ce qui est important pour les décideurs et le chef de projet c'est non seulement de suivre l'avancement du projet selon les délais impartis, mais aussi déterminer les sources d'échec possibles ou du moins les causes des ralentissements dans la production. Ils peuvent, par exemple, cibler le développeur qui peine le plus dans son travail et voir avec lui si c'est un problème de motivation, de matériel, etc. A contrario, les décideurs peuvent ainsi connaître l'équipe la plus performante et ainsi la récompenser.

De la vision globale du décideur à celle détaillée du développeur, les indicateurs ont une place prédominante dans la gestion de projet informatique et cela durant toute la vie du projet.

2. Gestion de projet BTP

a) FONCTIONNEMENT

Tout comme dans l'informatique, on retrouve dans le domaine du bâtiment, les 4 phases globales de la gestion de projet, avec aussi des variantes propres au domaine.

- **Programme :**

Cette première phase est spécifique au Maître d'ouvrage. Il s'agit en fait de la définition précise du projet de construction : choix du type de bâtiments (bureau ou appartement) et choix des solutions architecturales et techniques (forme du bâtiment, orientation des façades, organisation des espaces intérieurs, systèmes d'isolation, toiture, planché, Alimentation en énergie, etc..). Cette phase permet de faire ressortir toutes les exigences du MOA qui seront déterminantes pour la suite selon leur faisabilité technique et financière.

C'est aussi là que le MOA va déterminer le budget qu'il souhaite allouer à son projet : toute une approche financière et bancaire parfois complexe est envisagée afin de bâtir un dossier solide pour le projet.

À l'issu de cette phase, le MOA aura constitué un dossier – appelé Programme – comportant toutes les pièces administratives (mairie, etc.) nécessaires. Ce dossier est le cahier de charges du futur bâtiment : il valide les exigences et surtout les traduits en objectifs.

- **Conception :**

C'est à partir de cette étape que le MOE intervient. En effet, une fois le programme constitué, il faut que le MOA trouve le prestataire qui va l'accompagner dans son projet. Soit via un appel d'offres soit par un contact direct (selon le type de l'entreprise et de projet), une fois trouvé, le MOE peut prendre connaissance du projet et la collaboration peut alors commencer.

Le rôle premier du MOE lorsqu'il va étudier le projet est de le préciser : dans le programme le client a décrit ce qu'il attendait au niveau architectural et technique. Le MOE va accompagner le client afin de l'aider à cerner la faisabilité du projet dans les conditions initiales techniquement et surtout financièrement. Il apporte donc un complément d'information voir des réponses au MOA.

Cependant, avant d'en arriver à la sélection du MOE, il y a d'abord l'étude de la faisabilité du projet à l'aide d'un architecte. Cela se fait au cours des premières étapes de la phase de conception :

▪ **L'Esquisse :**

Il s'agit de l'étape de réalisation d'une série d'esquisses et de plans peu détaillés. Le but est d'obtenir une première visualisation du projet inséré dans le site. Ces esquisses se font en fonction du terrain, des options de la construction demandée par le MOA, et des contraintes financières. C'est donc une première réponse architecturale et technique au programme. Cette étape génère aussi une première estimation des coûts des différentes options qui répondent aux attentes du MOA.

Si le projet concerne la rénovation d'un bâtiment existant, au lieu d'une construction nouvelle, cette étape est alors dédiée non pas aux esquisses, mais au diagnostic technique et architectural pour réaliser les modifications souhaitées et les travaux requis.

À ce niveau, rien n'engage encore le MOA. Ce n'est que s'il est satisfait des premières esquisses et des prédictions de l'architecte, qu'il peut se lancer dans les études avant-projet.

- **Avant-projet sommaire (APS) :**

Cette première étude va être déterminante, car elle va permettre d'établir la demande de permis de construire. À la suite des esquisses et sur la base du programme, l'APS permettra de sélectionner les possibilités techniques les plus adaptées aux caractéristiques du projet selon les options retenues par le projet. L'architecte va donc fournir une description détaillée des différentes options retenues pour le projet et une estimation de leurs coûts. Ce document servira d'offre de service de l'architecte.

- **Avant-projet définitif (AVD)**

Cette troisième phase a pour but de préciser encore plus le projet afin de pouvoir le chiffrer au plus près. En effet, c'est durant cette phase que tous les choix techniques se font (choix des matériaux et des prestations techniques) et que les documents détaillant les caractéristiques définitives du projet architectural sont rédigées de manière formelle. C'est la formation du contrat. Point par point, l'ensemble des services fournis par l'architecte au MOA est décrit minutieusement.

En dehors de son assistance pour le projet et sa réalisation, l'architecte joue aussi un rôle administratif et juridique. C'est à lui que reviennent la constitution du dossier de permis de construire, et sa remise à l'autorité compétente. Il suit ainsi l'instruction du dossier et apporte les compléments nécessaires jusqu'à l'obtention du permis.

Une fois l'architecture sélectionnée, il faut choisir les entreprises sur le marché de la construction. L'architecte assiste le MOA dans cette étape aussi. Il l'aide à constituer le dossier de consultation des entreprises. Ce dernier est un document qui détaille l'ensemble des caractéristiques techniques de chaque lot du projet, ainsi qu'une estimation des budgets de chaque lot. Les entreprises peuvent ainsi proposer des devis au MOA qui va pouvoir les choisir en adéquation avec les buts du projet (basse conso ou écologiques), l'expérience requise, mais aussi le budget prévisionnel.

- **La réalisation ou Construction/Exécution :**

Cette étape représente la concrétisation d'objectifs papier en réalité. Les fondations sont mises en place et c'est donc le début de la construction.

Contrairement à l'informatique, où l'on peut contrôler à la fin d'un module, la réalisation des tests dans la construction doit, certes se faire au fur et à mesure de l'avancement, mais surtout à des étapes clés, sans quoi le projet risque d'être invalidé. Par exemple, un test adéquat doit être fait avant la pose des parements* et la fermeture des gaines techniques afin de vérifier les normes de sécurité, l'étanchéité, la perméabilité et l'isolation.

- **Réception/VISA :**

La réception est le moment de vérifier l'adéquation entre les résultats et les objectifs. Les dispositifs de régulation sont alors réglés et ajustés. Les installations de chauffage et de production d'eau chaude sanitaire sont programmées. Si le MOA découvre des malfaçons à cette étape, il peut encore faire des recours et obtenir des corrections finales.

Cette étape peut s'illustrer aussi par la rédaction d'un guide du bâtiment à l'attention du MOA. Il contiendra notamment les renseignements techniques permettant un entretien régulier des équipements.

- **Phase utilisation et entretien du bâtiment / exploitation - usage :**

À ce stade, le bâtiment est en fonctionnement, il est entre les mains des usagers. Bien qu'elle ne fasse plus partie du processus* de construction, cette phase est très importante surtout pour le MOA. Elle comprend la prise en main des usagers, l'entretien et l'instrumentalisation. Ainsi, si en informatique un groupe réduit d'utilisateurs finaux peuvent tester l'application avant sa mise en production, dans le bâtiment, il n'est pas possible de « tester » la construction dans son intégralité avant sa mise à disposition aux usagers. C'est donc une étape importante validant l'ensemble des tests précédents.

b) PROBLÉMATIQUE PROPRE À LA GPB

Une des problématiques majeures dans le Bâtiment est la problématique environnementale : ces dernières années la gestion du « durable » est devenue un enjeu incontournable.

Quatre points majeurs sont soulevés, et cela tous types de projet confondus :

- La gestion des ressources en eau
- La protection du milieu naturel
- L'émission de gaz à effet de serre
- La consommation énergétique.

Ces problématiques se présentent sous la forme de pratiques et de réglementations bien concrètes qui sont directement créatrices de complexité pour les projets.

En effet, la réglementation en vigueur amène les entreprises du bâtiment à rechercher les nouveaux équipements et technologies qui leur permettent de respecter la loi. Elles se doivent d'avoir un suivi beaucoup plus soutenu et précis de leurs travaux. Cela est source d'investissements supplémentaires financiers, humains et organisationnels. Cela peut même les amener à revoir tout leur processus* de construction selon l'environnement naturel du projet (faune, flore, sous-sols, bruit, patrimoine, paysage, etc.)

Les entreprises du BTP sont donc continuellement à la recherche de solutions de gestion efficaces pour l'ensemble de ces nouvelles informations. Elles cherchent donc à limiter la complexité des projets et à mieux inclure les données supplémentaires, afin de conserver des marges suffisantes pour leur pérennité.

c) INDICATEURS

Parmi les indicateurs de performance clé qui peuvent être abordés dans le bâtiment, en voici trois spécifiques à ce domaine :

Le KPI de Productivité de la main-d'œuvre :

Particulièrement important pour les sous- traitants, cet indicateur influe directement le budget main d'œuvre et les marges bénéficiaires. En effet, c'est selon l'information issue de cet indicateur, que les responsables sont capables de prédire les coûts nécessaires à la finalisation des travaux.

Calculé quotidiennement, cet indicateur peut aussi servir à établir des objectifs de performance quotidienne et des estimations futures d'avancement du projet.

Le KPI des Coûts engagés* :

Les entreprises de Bâtiment sont souvent face à un dilemme au vu du contexte économique actuel. Elles se retrouvent, en effet, souvent vulnérables et exposées financièrement, lorsqu'elles doivent avancer des sommes pour les travaux alors qu'elles n'ont pas été elles-mêmes payées. Les charges sociales et les prix des matériaux engendrant beaucoup de frais, l'erreur de jugement n'est donc pas permise.

Tout projet se doit d'être rentable surtout dans un domaine où les ressources sont quasiment toutes consommables (ciment, pavé, tapisserie, etc.)

Ainsi, il est essentiel pour l'entreprise de suivre l'évolution de ce KPI, afin d'une part augmenter la proportion des coûts engagés* et d'autre part inclure les facteurs tels que la hausse des prix et les termes de frais divers.

KPI Satisfaction des clients/carte de performances.

La concurrence dans le monde du bâtiment est féroce, il est donc important de développer un avantage concurrentiel et surtout de le préserver.

Le KPI de satisfaction des clients est un indicateur pertinent dans la recherche de cet avantage : en examinant les capacités de l'entreprise à satisfaire les clients et leur exigence, l'entreprise entreprend une analyse des remarques qualitatives. Cet examen rétrospectif lui permet ainsi de mettre en avant ses faiblesses potentielles et donc de déterminer ses axes des progressions. Ainsi, selon cet indicateur et l'information qu'il fournit, les projets futurs seront gérés différemment. Une refonte de la GPB sera même envisageable si l'indicateur est très mauvais.

Outre son utilité interne, ce KPI est l'exemple même du type d'indicateur qui peut être communiqué extérieurement pour une campagne de publicité par exemple.

3. GP et KPI : Identification des forces potentielles

Dans la gestion de projet en général, les KPI sont déterminants. Permettant d'une part un suivi du projet (et donc la mise en avant d'axes de progression), ils permettent surtout, combinés tous ensemble, l'identification des forces potentielles de l'entreprise et donc de ses avantages concurrentiels potentiels.

Le problème qui surgit alors est la qualité et la fiabilité des indicateurs. En effet, face à la multitude de systèmes informatiques incompatibles, les entreprises sont parfois contraintes à réaliser de lourdes manipulations de données. Cela implique donc une possibilité de perte d'information voire même des calculs erronés. Ces erreurs peuvent alors induire de mauvais choix stratégiques et orienter les décideurs dans les mauvaises directions.

Il est donc important pour une entreprise d'adopter la solution logicielle adéquate lui permettant de contrôler ses KPI. Ces derniers seront ainsi totalement conformes au plan stratégique de l'entreprise.

C'est donc grâce à une gestion méthodique et à des KPI performants qu'une entreprise peut détecter les opportunités et résoudre les problèmes qui portent atteinte à ses projets. Une perception claire de ses performances et de ses besoins actuels est ainsi possible, ouvrant la voie à une réussite future de la société.

Après avoir défini la gestion de projets en général, nous l'avons étudiée dans deux domaines différents : l'informatique et le Bâtiment. Le point commun qui en est ressorti est l'importance d'avoir des outils informatiques performants. Souvent cité, mais jusque-là pas encore présenté, il s'avère ainsi légitime d'aborder ces outils dans la partie suivante. C'est la première étape vers notre solution finale.

C. Les outils informatiques

Les logiciels de gestion de projet sont des outils informatiques pensés et conçus pour faciliter le travail des chefs de projets.

Il existe différents outils sur le marché. Certains sont propriétaires, d'autres sont open source*.

La première catégorie implique que le logiciel n'est accessible qu'après l'achat d'une licence. Celle-ci donne alors le droit d'utiliser, pendant une durée déterminée, ledit logiciel.

Le code* du logiciel, n'étant pas accessible, nécessite de contacter le prestataire, s'il y a un problème, ou une demande d'évolution. Duplication, modification et usage sont alors très limités.

La seconde catégorie est complètement à l'opposé : les logiciels open source *sont totalement libres, donc redistribuables* et utilisables sans aucune contrainte. Ils sont aussi gratuits.

Leur intérêt vient du fait que le code* est accessible. Ainsi, si un problème survient, ou qu'une évolution est souhaitée, il suffit d'avoir les compétences et le matériel. Il est même possible de créer des travaux dérivés à partir de ces logiciels.

Duplication, modification et usage sont, dans ce cas, illimités. Ce type d'outils sont privilégiés dans beaucoup d'entreprises et de pays, car cela confère une indépendance technologique indéniable et donc une plus grande liberté.

Tous les logiciels ne permettent pas de faire la même chose, le choix du chef de projet se fait généralement en fonction de ce qui le préoccupe le plus dans sa gestion.

Certains prennent en charge la gestion de version, alors que d'autres sont centrés autour de la planification (Gantt*, Pert*, etc.)

Voici, ci-dessous un comparatif de deux outils différents afin de se rendre compte des possibilités : le premier, édité par Microsoft est le logiciel propriétaire MS-Projet, le second, Open source* est Redmine. Ces deux logiciels sont très prisés dans le monde du développement informatique.

Logiciel	Logiciel collaboratif	Système de suivi des problèmes	Planification	Gestion de portefeuille de projets	Gestion de ressources	Pilotage par les livrables*	GED	Accessible en ligne	Licence
Microsoft Projet	X	X	✓	X	✓	X	X	X	Propriétaire
REDMINE	✓	✓	✓	✓	X	NA	✓	✓	LIBRE

Tableau 1: Comparatif d'outil de gestion de projet

1. OPEN SOURCE : REDMINE

Redmine est une application web libre de gestion considérée comme l'un des outils open source* les plus aboutis.

En effet, il propose de multiples services tels que la gestion multiprojet, la gestion électronique des documents (indexation de documents techniques), la gestion d'un wiki* intégré, la gestion des droits et des profils des intervenants, et celle des demandes d'intervention.

De plus, sous la forme d'un diagramme de Gantt, il fournit une console de l'état d'avancement des projets, des tâches et des recettes. S'interfaçant avec d'autres outils notamment les services de messagerie, il permet aussi l'envoi de notifications par email aux différents collaborateurs.

Au niveau des données, il est possible de configurer cet outil pour qu'il s'interface avec son serveur. Les intervenants peuvent alors s'y connecter depuis n'importe où grâce à un

navigateur internet (Google Chrome, Internet Explorer). La centralisation des données est garantie grâce à ce système.

En résumé, simple et efficace, cet ERP est très répandu en France et même à l'international, car il se décline en 48 langues depuis la version 2.3.0

2. PROPRIÉTAIRE : MS-PROJET

Édité par Microsoft, MS-Project est un logiciel de gestion de projets multitâches. En effet, il permet de planifier les projets, les tâches, de gérer les ressources et budgets, mais aussi de réaliser une analyse des données des projets.

Il est l'un des outils les plus répandus au niveau mondial. En effet, en 2011, il comptabilisait 20 millions d'utilisateurs (des chefs de projets surtout).

Au niveau de la gestion de projet, en plus de la planification, il permet le pilotage complet d'un projet : via cet outil, on peut créer une équipe puis lui affecter des tâches et ensuite saisir l'avancement du projet.

Une fois les données saisies, elles peuvent être utilisées pour une communication auprès des autres équipes ou pour une utilisation marketing : en effet, il comporte un module de reporting* qui lui permet donc d'afficher des tableaux, graphiques et diagrammes croisés dynamiques.

Tantôt dédiées aux chefs de projet, tantôt dédiées aux équipes mêmes, les solutions informatiques de gestion de projet sont multiples et parfois très différentes. Cependant, il existe un point commun indéniable à tous ces produits : la génération d'un gros volume de données. Nous verrons par la suite, lors de présentation de notre infrastructure open source*, comment, avec REDMINE, nous pouvons récupérer ses informations dans un but décisionnel.

D. Les Limites : causes d'échec

La gestion de projet n'est pas une science exacte, elle possède ses contraintes et subit souvent des échecs. Certains peuvent être préjudiciables et porter atteinte à l'entreprise, d'autres moins.

Dans un contexte où l'information et les données s'échangent de plus en plus rapidement, nous allons aborder les différentes causes d'échecs possibles afin de saisir les réelles problématiques.

1. CAUSE ÉCHEC AU NIVEAU HUMAIN ET ENTREPRISE

Que ce soit dans l'informatique ou le bâtiment, lorsque l'on échange avec des managers et chefs de projet, les problèmes les plus récurrents sont le dépassement de délais et de budget. En effet, malgré une planification précise et un cahier des charges* détaillé, il arrive que les projets prennent du retard sur leur planning.

Les causes d'échecs possibles sont multiples : tout d'abord, il y a le fait que les spécifications (demandes) du client sont souvent incomplètes ou imprécises.

De ce fait, toute l'estimation est erronée. Ensuite, certaines entreprises afin de décrocher un contrat n'hésitent pas à sous-estimer les charges et les délais. Le client conquis pense ainsi être livré en 3 mois, et en fait 6 mois plus tard ce n'est toujours pas le cas. Ce problème en particulier souligne le fait que ceux qui vendent un projet (commerciaux, patrons, etc.) n'ont pas souvent les compétences techniques requises pour estimer correctement la complexité d'une tâche technique et donc le temps ou les ressources humaines nécessaires à déployer pour la réaliser.

Une autre cause possible est l'imprévu : que faire si un fournisseur est en rupture de stock ? Que faire si l'outil de développement ne permet pas de faire de la 3D alors que cela a été vendu ? Bien que les responsables sont sensés avoir prévu toutes les possibilités, il est toujours possible qu'un imprévu survienne et ralentisse le rythme de production.

Au niveau humain, les causes d'échecs sont tout aussi nombreuses : la première est le travail en équipe. Cela nécessite de savoir échanger et communiquer. Plus il y a d'intervenants, sur un projet, plus il y a d'informations et donc de données à gérer. Il faut donc veiller à ce que tout soit centralisé, tracé et surtout accessible à tous en même temps. La coordination est un élément crucial pouvant faire défaut dans certaines équipes. Cela est alors préjudiciable pour leurs projets : les ressources manquent là où elles sont nécessaires, des produits non fonctionnels sont livrés, les risques sont mal pris en compte ou mal gérés, perte de temps sur la recherche d'informations, mauvaise répartition des rôles, etc.

Le capital social de l'entreprise et sa rentabilité sont alors directement impactés par cela.

2. OUTIL INFORMATIQUE NON ADAPTE : EXPLICATIF ET NON PAS PRÉDICTIF

En dehors du paramètre humain, il y a l'outil de gestion de projet qui peut être considéré comme une cause d'échec de projet.

Pour les décideurs, il est important de connaître l'avancement des projets de l'entreprise : où en est-en ? Combien de ressources financières sont déjà engagées ? Combien devons-nous engager encore, etc. ?

Malheureusement, tous les logiciels de gestion de projet n'offrent pas de vision transverse de l'ensemble des activités de l'entreprise. Et lorsqu'elles le font, le format n'est pas toujours assez explicite pour les décideurs. Les outils classiques sont donc inadaptés.

Par ailleurs, la gestion de projet nécessite un échange d'informations entre les parties prenantes. Cet échange peut se faire sous la forme d'emails, de fichier Word, Excel, PDF, Odt*, etc. Le format des données est donc totalement hétérogène. Ajouté à cela, le fait que l'information est souvent soit cloisonnée soit éclatée sur plusieurs outils, l'accès à la donnée devient alors plus problématique. D'autres problèmes sont aussi régulièrement soulevés tels que les problèmes de traçabilités, de gestion de version des documents.

C'est donc à cause de tout cela, que les outils classiques ne sont plus suffisants à l'heure où le volume de données ne cesse de croître et que les besoins en décisionnel deviennent de plus en plus pressants.

3. INDICATEURS

Le problème principal qui découle d'outils inadaptés est la fiabilité des données. En effet, si les outils ne peuvent gérer en temps réel un grand volume de données, il va alors de soi que les indicateurs de mesure ne sont plus aussi pertinents. Les décideurs prennent alors des risques énormes puisqu'ils peuvent prendre des décisions erronées.

Ce problème met en exergue deux autres problèmes : le premier est le caractère explicatif des indicateurs et le second est leur choix préfixé.

Si l'on considère le premier point, on prend conscience que les KPI ne sont ni plus ni moins que des états des lieux. Se basant sur des données passées ou actuelles, ils représentent à un instant t ce qui est.

C'est ensuite aux décideurs d'analyser les données et « suivre leur intuition » afin de prendre des décisions. Ainsi, bien qu'ils permettent de faire des prédictions, ils ne sont pas prédictifs : les choix et prédictions sont issus d'éléments abstraits (intuition, feeling et expérience du décideur), ils ne sont pas basés sur du factuel ou des calculs tangibles. Le risque est donc omniprésent.

Or aujourd'hui cela n'est plus suffisant : le monde évolue rapidement et des échanges et contrats sont conclus à chaque instant. Dans ce contexte, la veille concurrentielle et informationnelle* se doit d'être la plus efficace possible et donc le système de prise de décisions aussi.

Le second problème mis en exergue précédemment est le choix préfixé des indicateurs : à partir du moment où l'on choisit de se focaliser sur un élément, on ignore le reste ou du moins, on y prête moins attention. De ce fait, on peut laisser passer des informations qui pourraient se révéler cruciales par la suite.

Par ailleurs, il est important de noter que les KPI se calculent sur des données tangibles (temps passé, jour homme). À aucun moment, ils ne prennent en compte l'immatériel (motivation des équipes, ambiance de travail et du marché, image de marque de l'entreprise). Or ce dernier a une place importante dans l'économie, il est donc important de l'inclure dans le système de décision.

Enfin, en terme de calcul, le problème des indicateurs est leur simplicité. En effet, les règles de calculs, plus ou moins simples, peuvent porter préjudice à leur pertinence, car ils masquent certains paramètres environnementaux.

Bien qu'indispensable en gestion de projet, les outils informatiques aujourd'hui ne sont plus suffisants. À l'heure où les volumes de données explosent et les échanges sont quasi instantanés, on s'aperçoit que les outils sont incomplets : ils ne prennent pas du tout en compte ces aspects-là. Plutôt descriptif, ils ne permettent pas de faire de l'hyper vision ou du prédictif en se basant sur des faits, mais seulement sur de l'intuition. Le risque est alors considérable et surtout non maîtrisé.

Dans la suite de ce mémoire, nous allons aborder un domaine nouveau et jeune qui promet d'être la solution à tous les problèmes cités précédemment : la fouille ou l'exploration de données, connus aussi sous le nom de DATAMINING.

II. Les éléments de solution : DATAMINING

A. Le Datamining : Théorie et Origine

Le Datamining ou forage de données est l'exploration de grand volume de données afin d'en extraire des informations pertinentes pour les décideurs. Il s'agit donc d'un processus* de découverte de connaissance (Knowledge discovery in data, KDD) issu de la jonction de trois domaines scientifiques : l'intelligence artificielle, les Statistiques et les traitements liés aux bases de données.

Son principal avantage est qu'il n'y a pas besoin de préfixer ce que l'on cherche : en effet, grâce à un système de recherche semi-automatisé, les outils de Datamining permettent d'établir des corrélations invisibles entre des données a priori sans aucun lien.

L'utilité même du Datamining peut être comprise par l'exemple de la collaboration d'IBM et de la police de Memphis. En mettant en place un système d'analyse prédictif, l'application arrive à déterminer les lieux où pourraient avoir prochainement des crimes tels des cambriolages ou vols de voitures. La police renforce alors sa surveillance en postant des patrouilles dans ces zones-là.

Après quelques mois d'expérimentation, une baisse de 15,8% des crimes a été observée par la police sans aucune augmentation du nombre d'officiers employés. L'application avait eu raison dans 40% des cas grâce à la détection de facteurs anormaux ou de corrélations qui avaient échappé à la police.

Cet exemple de l'utilisation du Datamining est très éloquent, car il met en exergue les points essentiels de ce domaine : tout d'abord, l'application dispense l'utilisateur de chercher des réponses à des questions en lui fournissant des axes de réflexions. De plus, les grands volumes de données ne sont plus un obstacle dans la recherche d'informations, mais un atout. Enfin, les résultats obtenus nécessitent d'être analysés, afin d'une part de déterminer les types de relation et leurs causes (cause à effets, résultante d'une cause) et d'autre part afin d'aboutir à une préconisation concrète qui sera utile pour les décideurs.

Ainsi la pertinence des résultats du Datamining repose en partie sur l'analyse effectuée.

Ce qu'il faut souligner surtout, c'est le caractère polyvalent du Datamining, car certes il peut être utilisé dans un objectif prédictif, mais il possède aussi des fins explicatives.

Prenons un autre exemple, celui de l'entreprise américaine Wall Mart spécialisée dans la grande distribution. Cette dernière fut l'une des premières entreprises à utiliser les premières techniques de Datamining sur ces produits. Elle réalisa alors que le samedi après-midi les ventes de couches augmentaient proportionnellement aux ventes de bières. Apparemment sans liens directs, il s'avéra, après analyse que le samedi après-midi, pour les couples ayant un ou plusieurs enfants en bas âges, c'était les pères qui faisaient les courses. Ainsi ces derniers prenaient des couches pour leurs enfants et des bières pour eux-mêmes. L'entreprise choisit alors de ré-agencer les rayons, plaçant à proximité les couches, des bières. Ses ventes grimperent alors en flèches.

Le Datamining est donc issu de la nécessité de trouver des informations riches et pertinentes. Son objectif est de découvrir des modèles implicites dans les données en exploitant, par exemple, les données de l'historique d'une entreprise, afin d'améliorer et sécuriser le processus* de prise de décisions.

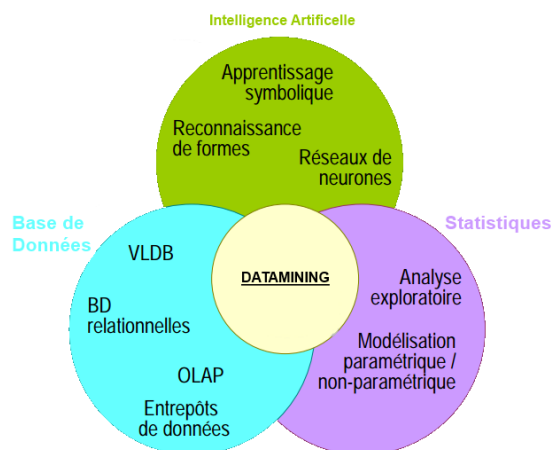


Illustration 4: Datamining à la jonction de l'IA, les statistiques et les bases de données

Les bases de données ou entrepôts des grandes entreprises atteignent aujourd'hui des volumes allant de plusieurs téraoctets (1 téraoctet = 10^{12} octets) à plusieurs pétaoctets (1 pétaoctet = 10^{15} octets). SFR, EDF ou encore Orange, collecte annuellement plusieurs grands volumes de données liés à la consommation de leur bien. Ils utilisent souvent ces données dans un but marketing (édition de profil de consommateur). Cependant, parmi ce grand amas, tout n'est pas utilisé.

En effet, dans chaque entreprise, il existe un volume de données non exploitées. Ce dernier souvent volumineux ne représente pas d'intérêt direct pour l'entreprise, mais peut contenir une mine d'information. Le Datamining permet justement d'extraire ces informations et cela de manière fiable.

Ainsi, le but du Datamining est de transformer des données volumineuses et multiformes en connaissance. Celle-ci peut être sous la forme de tableau, d'un concept, d'un rapport ou d'un graphique ou tout simplement d'un modèle mathématique ou logique pour la prise de décision. Mais, quelle que soit leur forme, ces nouvelles connaissances peuvent, à l'initiative de l'utilisateur, être injectées dans la base de connaissance afin d'être analysées à leur tour dans un nouveau processus* de Datamining. Ce dernier est donc itératif et interactif.

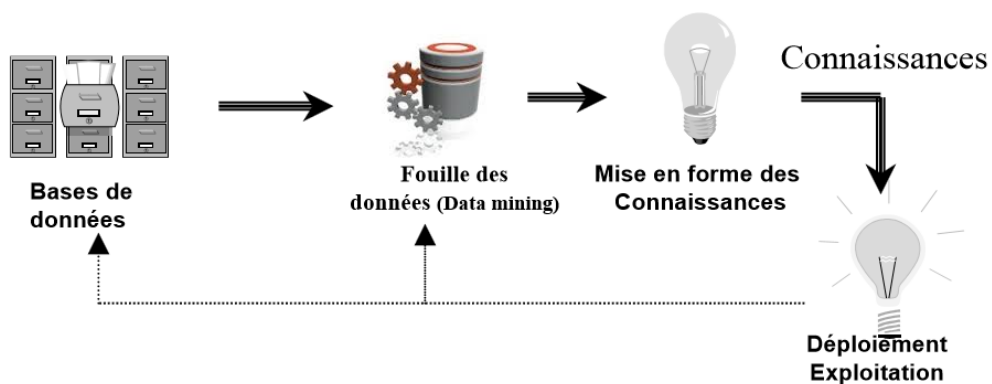


Illustration 5 : le Datamining, un processus itératif et interactif

Traditionnellement, que ce soit via des bases de données relationnelles ou des cubes* OLAP, l'utilisateur a l'initiative : il choisit donc ce qu'il veut observer ou analyser. C'est souvent un expert dans son métier afin de réellement pouvoir apprécier les informations qu'il percevra. Dans le cadre du Datamining, il en est de tout autre. Comme cité précédemment, le système est celui qui a l'initiative, puisque c'est lui qui découvre les associations qu'il soumet à l'utilisateur. Ce dernier n'a pas besoin d'être un expert pour apprécier l'information, il lui suffit de connaître le métier. Ainsi, il a la capacité de compréhension et d'analyse nécessaire.

Le Datamining a donc démystifié l'analyse de données en donnant les clés à toute personne du métier.

Enfin le dernier point, mais non pas des moindres, un outil de Datamining appliqué à un même ensemble ne donnera pas toujours les mêmes résultats.

En effet, au fur et à mesure que la base de données sera alimentée en informations, le système les inclura dans son traitement. Les résultats seront donc adaptés aux données et donc probablement différents.

B. Le Datamining : Notions de base nécessaire à la compréhension

Il n'existe pas un seul outil ou une seule technique de Datamining, mais plusieurs reposants chacune sur des algorithmes mathématiques bien spécifiques. Afin d'être parés pour la suite, nous allons d'abord aborder des notions nécessaires à la compréhension et que nous retrouverons par la suite.

1. LES STATISTIQUES

À la base de tout raisonnement sur les données, les statistiques sont omniprésentes tout au long du processus* de Datamining, car ils permettent de synthétiser un grand nombre de valeurs pour une variable donnée.

Ainsi pour chaque variable, il y a deux indicateurs importants : le premier permet de mesurer la tendance centrale, et le second permet de mesurer la dispersion.

a) *Indicateur de tendance centrale*

Les mesures des tendances centrales permettent de déterminer, sur une série de données, la valeur centrale autour de laquelle les données ont tendance à se rassembler. En voici les trois exemples les plus connus :

Moyenne arithmétique : Il s'agit de l'indicateur le plus couramment utilisé.

Ex : Soit \bar{X} la moyenne arithmétique de la quantité

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

L'inconvénient principal de cet indicateur est d'être sensible à la présence de valeur aberrante.

Médiane : Plus robuste que la moyenne, cet indicateur va diviser la population (liste de valeur) en deux parties égales. D'une part, il y aura l'ensemble des valeurs qui sont

inférieures ou égales à la médiane et de l'autre côté, l'autre moitié des valeurs supérieures. S'il y a un nombre impair de valeurs, la valeur centrale sera mise des deux côtés.

Mode : Le mode d'une série statistique est la valeur la plus fréquente. Dans le cas d'une répartition en classe, la plus fréquente sera dite "classe modale".

b) Indicateur de dispersion

Comme leur nom l'indique, les indicateurs de dispersions permettent de mesurer comment les données se «répartissent» autour de la moyenne.

Prenons l'exemple ci-dessous :

Soit un Etudiant A avec les notes suivantes : 9 - 10 - 10 - 10 - 10 – 11

Et un Etudiant B avec les notes suivantes : 0 - 0 - 0 - 20 - 20 – 20

Ces deux étudiants auront tous deux une moyenne de 10, cependant ils ne se valent pas.

Afin de mesurer la dispersion entre les valeurs, il existe plusieurs indicateurs :

Étendue : nommé aussi l'amplitude ou la dimension, il correspond à l'écart entre sa plus grande valeur et sa plus petite

$$\text{Étendue} = \text{Valeur Sup.} - \text{Valeur Inf.}$$

Bien qu'intéressant, il reste assez peu robuste comme indicateur et peu manquer de pertinence.

Intervalle interquantile : nommé aussi distance interquartile, il s'agit de l'écart absolu moyen des valeurs à la moyenne de la distribution

Variance : Elle permet de combiner toutes les valeurs à l'intérieur d'un ensemble de données afin d'obtenir la mesure de dispersion.

Soit une série de données numériques suivantes : 1, 2, 3

$$\text{Variance} = [\text{somme de l'écart au carré}] \div \text{nombre d'observations}$$

Ainsi

$$\text{Variance} = [(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0,667$$

Écart-type : Il s'agit de la racine carrée de la variance.

Si l'on prend l'exemple précédent : l'écart type est égale à racine carrée (0,667) =

La variance et l'écart-type sont les mesures de dispersion les plus couramment utilisées.

Généralement, plus les valeurs sont largement distribuées, plus l'écart-type est élevé.

Grâce à ces indicateurs, les statisticiens peuvent mettre en évidence des corrélations entre plusieurs variables ou plusieurs populations. Voici les différentes manières qui expliquent la corrélation :

La causalité : on observe qu'une variation de A entraîne systématiquement une variation de B. Il existe un vrai lien entre A et B.

Le hasard : une variation de A entraîne une variation de B, mais cela n'est arrivé qu'une fois. C'est uniquement dû au hasard

La réponse commune : une variation de C entraîne une variation de A et B.

La confusion : la variation de A et C entraîne la variation de B.

C'est ainsi grâce à ces techniques que l'on peut donc faire des prévisions et des estimations et ainsi donc établir des liens de causalité. Malheureusement, l'utilisation des statistiques dans ce but est limitée, car souvent on ne peut étudier simultanément un grand nombre de variables.

De nouveaux algorithmes ont donc été développés alliant à la fois recherche intelligente et statistique.

2. LES SCHÉMAS D'INFÉRENCE

Le Datamining met en œuvre des techniques de raisonnement, regroupées sous le nom de schémas d'inférence. Il en existe trois types : l'abduction, la déduction et l'induction.

a) *L'abduction*

Cette méthode de raisonnement permet par exemple de déterminer à partir d'une liste d'éléments qu'elle est la solution la plus probable en supprimant celle improbable.

Ex. : Tous les chats ont 4 pattes, Lilou a 4 pattes. Ainsi Lilou est un chat.

Dans cet exemple, comme l'on a peu de données descriptives, on pourrait obtenir un résultat aberrant si Lilou avait été une vache ou un chien.

C'est pour cela que c'est l'une des méthodes de raisonnement qui a besoin de beaucoup d'information afin que le risque d'erreur soit très minime.

C'est d'ailleurs une méthode régulièrement utilisée dans le milieu médical, afin de déterminer la pathologie la plus proche des symptômes du patient.

b) Déduction

Ce mode de raisonnement est le plus utilisé par l'homme et le plus familier, car il ne laisse pas de place au doute.

Ex. :

LiLou est un chat, or les chats ont 4 pattes. Ainsi Lilou a 4 pattes.

c) Induction

Ce dernier est le plus utilisé dans le Datamining. Contrairement à la déduction, l'induction génère la loi à partir des faits qu'elle étudie. Elle passe donc du particulier au général.

Le principe consiste à étudier une série de fait (cas particulier) afin d'en tirer une conclusion globale. Il n'y a aucune certitude concernant la véracité et la justesse de la conclusion. Mais plus l'échantillon sera grand, plus la certitude liée à l'exactitude de la conclusion se renforcera jusqu'à en faire une vérité.

Exemple :

Le cygne n°1 est blanc.

Le cygne n°2 est blanc.

Le cygne n°3 est blanc.

Et ainsi de suite...

Le cygne n°1000 est blanc.

Donc tous les cygnes sont blancs.

Dans ce cas, sur les 1000 premiers cygnes, nous n'avons rencontré aucune contradiction logique affirmant qu'il existe des cygnes d'une autre couleur que blanc. De plus, plus nous verrons de cygne blanc, plus nous en serons convaincus.

Ceci explique que la recherche se fait généralement par induction. La déduction est plutôt utilisée pour vérifier la cohérence des informations.

La faiblesse de ce raisonnement est qu'il nous suffira d'un seul cygne noir , pour que le raisonnement par induction s'écroule.

La théorie de la gravitation universelle est issue, à l'origine, d'un raisonnement inductif.

3. ARBRES DE DÉCISION

Il s'agit d'un outil d'aide à la décision.

Grâce à la représentation sous la forme d'un graphique d'un arbre, cet outil permet de représenter la situation étudiée en positionnant, à l'extrémité de chaque branche, les différents résultats possibles en fonction des décisions prises à chaque étape.

L'arbre de décision est un outil utilisé dans plusieurs domaines, grâce notamment à la lisibilité et la rapidité d'exécution qu'il offre pour des situations parfois très complexes. Il peut être appliqué dans des domaines très variés tels que la sécurité, la médecine, l'intelligence artificielle ou encore la fouille de données.

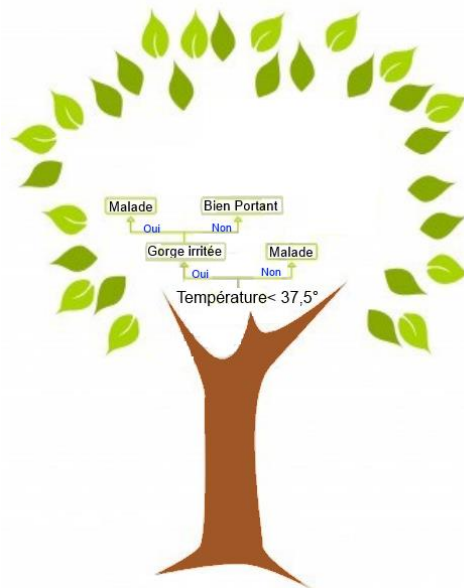


Illustration 6 : Arbre de décision

4. RÉSEAUX DE NEURONES

Un réseau de neurones artificiels est un modèle de calcul inspiré du mode de fonctionnement des neurones biologiques.

Il fait partie, d'une part, de la famille des applications statistiques et d'autre part de la famille des méthodes d'intelligence artificielle.

De manière générale, ce modèle de calcul permet l'apprentissage supervisé de nos machines : elles apprennent des faits grâce à l'aide des utilisateurs en mettant en œuvre le principe de l'induction, c'est-à-dire l'apprentissage par l'expérience.

C. Le Datamining : Étape de fonctionnement

Le Datamining n'est pas mis en place pour une simple exploration des données, mais bien pour mettre en évidence des corrélations qui échapperaient en temps normal à la vigilance des utilisateurs métiers.

Son fonctionnement passe par quatre étapes : l'identification du domaine d'étude, la préparation des données, l'action sur la base de données et enfin l'évaluation des actions.

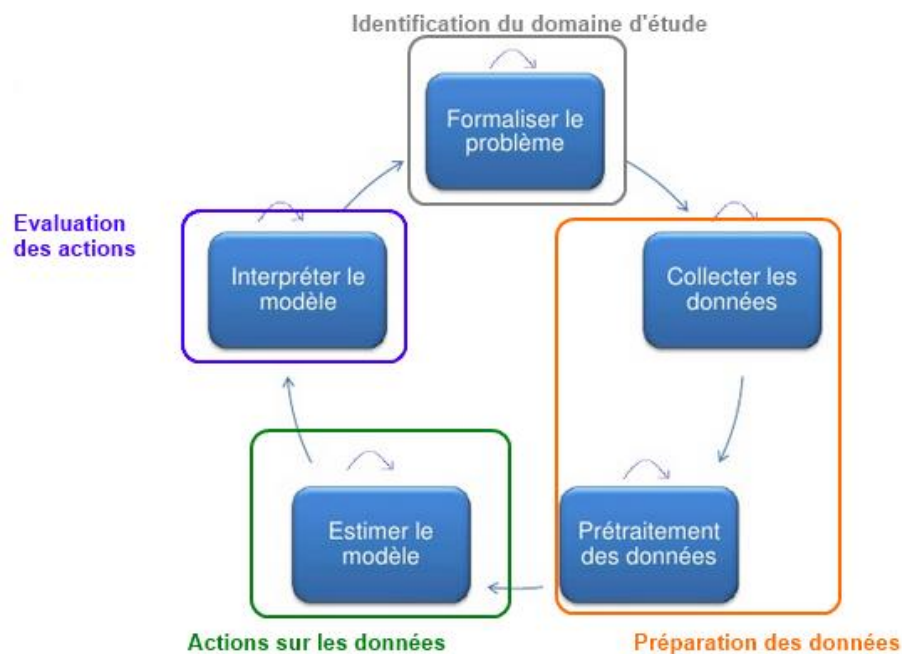


Illustration 7 : Fonctionnement du Datamining

1. IDENTIFICATION DU DOMAINE D'ÉTUDE

Cette étape primordiale permet de répondre aux questions suivantes : « de quoi parlons-nous ? et que souhaitons-nous faire ? » C'est donc ici que l'on définit un objectif général.

De ce fait, on peut aisément cibler l'espace de données qui va être exploré. Si ces données ne sont pas suffisantes, l'utilisateur reviendra en chercher par la suite. Mais en aucun cas, il ne prendra l'intégralité des données d'un coup, car outre le temps et les performances des machines nécessaires, cela risquerait de ne pas aboutir à des résultats pertinents en relation avec l'objectif.

Dans l'exemple de Wal-Mart, l'objectif était la recherche de corrélation entre les ventes de produits. C'est grâce à cela, que l'utilisateur métier a su sur quel élément il devait se focaliser à l'étude des résultats.

2. PRÉPARATION DES DONNÉES

Une fois les interrogations résolues et les objectifs clairement fixés, il faut se pencher sur les données. Les principales opérations de préparation sont les suivantes :

La sélection ou réduction: Il s'agit de la phase qui vise à déterminer les filtres qui permettent de ne sélectionner que les sous-ensembles intéressants. Utilisant les méthodes statistiques vues précédemment telles que l'échantillonnage et les indicateurs de tendance centrale et de dispersion, l'utilisateur peut ne garder par exemple que les attributs dont la moyenne est supérieure à un seuil ou ne conserver que les attributs qui ont un lien statistique significatif avec un attribut particulier.

Cette phase est d'ailleurs l'un des sujets majeurs de la recherche en Datamining.

Le nettoyage : Afin de ne pas gêner l'analyse, les valeurs manquantes ou aberrantes doivent être gérées correctement. Plusieurs méthodes existent pour cela. Pour les données manquantes, une des méthodes préconise leur remplacement par le mode de la distribution statistique (la valeur la plus fréquente). Quant aux données aberrantes, ils sont repérés grâce à une règle préétablie par l'utilisateur. Par exemple, toutes les données numériques dont la valeur sur un attribut donné s'écartent de la valeur moyenne plus deux fois l'écart-type, pourraient être considérées comme des données possiblement aberrantes.

Transformation : Cette étape permet de transformer un attribut A en un autre attribut A', qui sous certains aspects serait plus approprié pour l'étude. Plusieurs méthodes peuvent être utilisées telles que la discrétisation des variables continues ou la binérisation* des variables. La première consiste à construire une variable discrète à partir d'une variable continue, alors que la seconde est une transformation en binaire des variables nominales (tels les noms, prénoms, etc.) pour uniformiser le format de donnée et ainsi faciliter l'application de fonctions particulières.

Quel que soit le type de données, cette phase vise à les transformer afin qu'ils puissent subir les méthodes de Datamining. Le format de traitement le plus privilégié est le tableau numérique.

3. PHASE DE FOUILLE DE DONNÉES : ACTION SUR LA BASE DE DONNÉES

Les experts estiment que cette étape est le cœur du Datamining ou du processus* de découverte de données, puisque c'est ici que l'on découvre les nouvelles connaissances.

Parmi toutes les méthodes de fouilles de données, trois catégories se distinguent :

- Les méthodes de visualisation et de description
- Les méthodes de classification et de structuration
- Les méthodes d'explication et de prédiction

a) Les méthodes de visualisation et de description

Description

Comme son nom l'indique, il s'agit de décrire les données. Impliquant une exploitation supplémentaire et une consommation de ressources parfois importante selon la taille de la base de données, cette étape est néanmoins importante afin de « clarifier » les données et faciliter d'autres étapes telles que le groupement par similitudes ou la classification.

Exemple :

Les données décrivant d'autres données sont les métadonnées telles que : la date de création, de modification, l'opérateur de création, etc.

Optimisation

Une des techniques les plus appropriées de l'optimisation est le réseau de neurones. Son but est de maximiser ou minimiser la fonction d'évaluation à laquelle elle est associée.

b) Les méthodes de classification et de structuration

Classification

« La classification consiste à examiner des caractéristiques d'un élément ou objet afin de l'affecter à une classe d'un ensemble prédéfini. »

En informatique, cela va permettre de regrouper les données dans des catégories.

Cette fonction va donc permettre de créer des classes d'individus (dans un sens statistique) qui seront discrètes : hommes/femme, oui/non, rouge/gris, etc.

L'une des techniques les plus appropriées pour cela est l'utilisation arbres de décision.

Exemple :

- attribuer ou non un prêt à un client,
- établir un diagnostic,
- accepter ou refuser un retrait dans un distributeur,
- attribuer un sujet principal à un article de presse, etc.

Groupement par similitudes

Cette méthode consiste à regrouper ensemble tout ce qui présente des similitudes, c'est-à-dire ceux qui vont naturellement ensemble.

Exemple :

Sur une série de mots, on va pouvoir grouper ensemble les verbes, les adjectifs, les noms, etc.

Segmentation (ou clusterisation)

Le but de cette fonction est de segmenter une population hétérogène en sous-ensemble homogène. L'intérêt et la signification des groupes constitués doivent être fixés par un expert du domaine. Toutefois, cet intérêt dépend fortement des buts de la classification et d'estimation qui seront menées par la suite sur ces groupes. En effet, cette tâche est souvent effectuée avant les précédentes pour construire des groupes homogènes auxquels on appliquera les fonctions que l'on souhaite.

Exemple :

Soit un groupe d'homme et de femmes, il pourra être segmenté soit en fonction du sexe, soit en fonction de l'âge.

c) Les méthodes d'explication et de prédiction

Estimation

Si l'on souhaite estimer le revenu d'un ménage selon divers critères (professions, type et nombre de véhicules, etc.), nous aurons recours à l'estimation. Celle-ci consiste à estimer la valeur d'un champ à partir des caractéristiques d'un objet. Le champ à estimer est un champ à valeurs continues, ce qui permet par la suite de procéder à la classification selon un barème. C'est grâce à cela que l'on peut définir des tranches de revenus pour classer les individus par exemple.

Si elle présente plusieurs intérêts, deux sont des plus intéressants : le premier concerne le fait de pouvoir ordonner les résultats. De ce fait, si on le souhaite on peut ne récupérer que les n meilleurs valeurs. Ce procédé est ainsi souvent utilisé en marketing afin de proposer les meilleures offres aux consommateurs.

Le second intérêt est la possibilité de déterminer la position d'un élément dans sa classe. Si on souhaite ainsi exclure les cas limitrophes, cela en sera plus facile.

Pour cela la technique la plus appropriée est le réseau de neurones.

Exemple :

- noter un candidat à un prêt ; cette estimation peut être utilisée pour attribuer un prêt (classification), par exemple, en fixant un seuil d'attribution,
- estimer les revenus d'un client.

Prédiction

Tout comme les fonctions précédentes, la prédiction s'appuie sur des données passées et présentes afin de donner un résultat qui aura lieu dans le futur : la dimension temporelle est donc différente.

Les techniques les plus appropriés sont : les arbres de décision et les réseaux de neurones.

Exemple :

- prédire les valeurs futures d'actions,
- prédire au vu de leurs actions passées les départs de clients.

Règles d'association ou « analyse du panier de la ménagère » :

Cette fonction a pour but de déterminer sur une liste d'éléments, quels sont ceux qui sont associés. L'exemple type est l'analyse d'un ensemble de tickets de supermarché, afin de déterminer les articles qui se retrouvent souvent ensemble tel que (le poisson et le vin blanc ; la baguette et le camembert et le vin rouge, etc.). Grâce à cela des opportunités de ventes croisées peuvent être décelées et des regroupements attractifs de produits peuvent être envisageables.

Exemple :

L'analyse du supermarché Wal-Mart qui a mis en évidence la corrélation entre les ventes de bières et de couches le samedi après-midi.

D'autres fonctions existent, cependant beaucoup plus techniques, elles ne seront pas aborder dans ce mémoire.

d) Les méthodes d'apprentissage

Dans le cadre du Datamining, le système est capable de faire évoluer sa base de connaissance en inférant de nouvelles règles. On dit donc qu'il possède une capacité d'apprentissage.

Il en existe deux types : l'apprentissage supervisé et non supervisé.

Apprentissage supervisé

L'apprentissage supervisé désigne un ensemble de technique d'intelligence artificielle qui permet d'apprendre à la machine. À partir d'une base de fait déjà établi, on va mettre en place un traitement qui apprendra à l'automate à reconnaître les situations.

En d'autres termes, nous connaissons en tant qu'utilisateurs le résultat que l'on souhaite apparaitre. Ainsi nous entrainerons la machine à retrouver ce résultat.

Si elle fournit un résultat différent de celui attendu, on doit alors adapter le traitement. On répète, ainsi de suite l'opération jusqu'à ce que la machine fournisse le résultat escompté.

L'apprentissage est dit supervisé, car l'utilisateur interagit avec le système afin de faire évoluer la base de connaissance.

Le réseau de neurones est une technique permettant l'apprentissage supervisé.

Ce type de méthode est surtout utilisé pour la détection de fraude, ou encore le marketing téléphonique.

Apprentissage non supervisé

On parle d'apprentissage non supervisé, lorsque le système arrive à inférer de nouvelles connaissances seul, sans l'intervention d'un utilisateur.

Il utilise une grande base de connaissance, afin d'étudier et de trouver de nouvelles informations.

Le résultat des algorithmes de l'apprentissage non supervisé doit être analysée afin d'être retenu pour un usage ou tout simplement rejeté.

Ce type de méthode utilise par exemple les règles d'association ou de classification, etc.

Les domaines d'application sont l'identification de segments de marché et identification de document similaire.

Dans le cadre du Datamining, nous sommes dans le cas de l'apprentissage non supervisé. L'utilisateur métier n'intervient pas lors du processus d'apprentissage, mais une fois la connaissance découverte. Il la valide ou la rejette en bloc alors, avant de lancer un nouveau cycle de découverte de connaissance.

4. ÉVALUATION DES ACTIONS

Une fois les résultats obtenus, ils seront alors analysés. À ce stade, ce qui est évalué c'est non seulement la pertinence des résultats, mais aussi la performance du processus* de Datamining : les résultats sont-ils réalistes ? Sont-ils justes ou aberrants ?

S'il le faut, de nouvelles actions seront décidées. On considère alors que le premier cycle a abouti à l'expression d'objectifs nouveaux et plus affinés. Le processus* reprend alors à l'étape N°1.

C'est donc ici que l'utilisateur intervient pour valider ou rejeter la proposition du système. Selon le résultat, la proposition peut devenir une connaissance confirmée, ou une proposition à rejeter. Un exemple de résultat aberrant serait de voir un client lambda apparaître dans la liste des clients potentiellement intéressés par des ustensiles de cuisine, alors que celui-ci n'a jamais consulté ce type d'article et qu'il a toujours acheté ou consulté que des livres.

Quant au résultat réaliste, il s'agit en fait d'un résultat ambigu : il n'est ni totalement faux, ni totalement vrai. Tout dépend de l'utilisateur métier et de son appréciation de la situation.

L'actualité du 25 août 2013 est d'ailleurs marquée par cela : un internaute fan de roman d'histoire, s'est vu proposer par le réseau social Facebook, sur lequel il est inscrit, la lecture

du livre « Mein Kampf »¹. N'ayant aucun intérêt, ou n'ayant jamais cherché à consulter de document sur le troisième Reich, cet internaute a été très étonné. Cette histoire est un exemple parfait d'un résultat « ambigu » : il peut sembler choquant, mais reste néanmoins logique si ce livre a été désigné comme livre d'histoire. Dans ce genre de cas, seul l'utilisateur, ici l'internaute, peut accepter ou rejeter cette proposition.

D. Le Datamining : Domaines d'applications

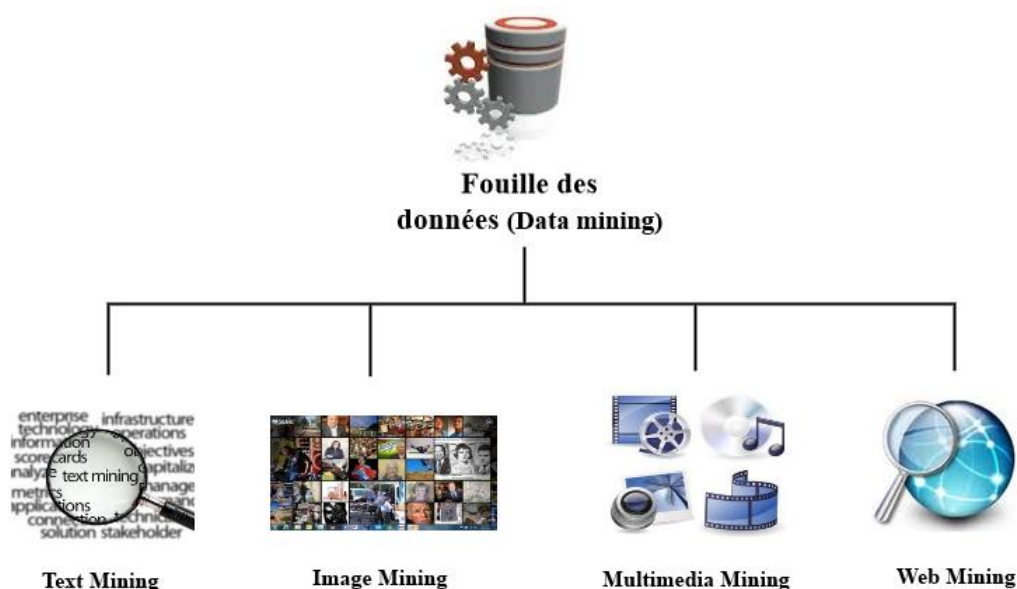


Illustration 8 : Domaines d'application du Datamining

1. LE TEXT MINING

Une base de données contient plusieurs types d'information dont une grande partie est en format « libre ». Ces données peuvent être des rapports, publications, mails, commentaires, etc. Tout aussi importants que des données numériques, ils recèlent aussi des informations et connaissances qui peuvent se révéler primordiales pour le processus* de prise de décisions. Il est donc important de les traiter aussi. Pour cela, les chercheurs utilisent le text mining, à savoir la fouille de données textuelle.

¹ Cet exemple n'est ni une plaidoirie pour ou contre le livre cité. Son but est de démontrer l'utilisation du datamining dans la vie de tous les jours et les éventuels résultats que ses techniques peuvent fournir. L'article est disponible sur le lien : <http://www.20minutes.fr/insolite/1213451-20130825-facebook-conseille-mein-kampk-a-fan-romans-historiques>

Celle –ci définit des stratégies d’exploitation des textes en format libre selon 2 niveaux de traitements : la recherche d’information et l’extraction de connaissances.

Le premier niveau repose sur des requêtes qui visent à chercher soit une lettre dans un texte soit des expressions complexes ou même des textes en exemples. Ces requêtes permettent donc un accès aux textes soit via leur contenu lexical ou sémantique.

Le second niveau, se focalisant sur l’extraction de connaissances à partir de texte, nécessite l’établissement de règles minimales. Dans la presse électronique, on trouve de tout. Des articles traitant de sujets sensibles comme ceux traitant de sujets banals. Un modérateur (personne veillant à la bonne utilisation des forums) souhaitera régulièrement vérifier que les propos laisser par les internautes ne sont ni insultants ni des incitations à la haine. Malheureusement, lors de ses recherches, le programme ne lui renverra pas forcément tous les commentaires illégaux, car certains ne sont pas assez explicites dans leur propos.

C’est ici qu’interviendront les algorithmes de text mining : les méthodes de Datamining vont aider l’usager à établir des règles minimales qui vont permettre de reconnaître ces textes.

Une fois validées par le modérateur, ces règles seront discriminantes pour ces commentaires qui seront alors bloqués.

2. L’IMAGE MINING

Tout comme les textes, les images peuvent aussi être traitées par les techniques de Datamining. Il s’agit alors de l’Image mining : son but est de permettre l’identification, la reconnaissance ou encore la classification automatique de grande base d’images.

Actuellement, pour pouvoir être traitée, chaque image est associée à un index qui décrit son contenu. Cette association se fait majoritairement manuellement, mais elle tend à être de plus en plus automatisée.

Comme pour des données standard, les images doivent être préparées pour les traitements : Après une transformation, un filtrage et une mise en forme visant à faire ressortir les caractéristiques importantes pour l’utilisateur, chaque image est transformée en un ensemble de vecteurs de nombres afin d’être incluse dans un tableau numérique. Dans ce dernier chaque ligne est une image et chaque colonne une caractéristique de cette image. C’est sur ce tableau que seront appliquées les méthodes du Datamining.

3. LE MULTIMEDIA MINING

Le multimedia mining correspond au traitement de contenu multimédia par les techniques du Datamining afin de trouver des relations statistiques ou des modèles. Obéissant aux mêmes principes que ceux de l'image et du texte mining, les données subissent aussi une préparation en vue de convertir les séquences vidéo en données numériques que l'on pourra inclure dans un tableau.

La différence dans ce cas, est qu'il faut gérer à la fois le son et la vidéo sans compter sur les données textuelles qui peuvent être présentes.

4. LE WEB MINING

Le web est une source d'information considérable pour les entreprises. Au même titre qu'il est intéressant pour les entreprises d'extraire la connaissance de leur base de données, il leur est important d'exploiter celle présente sur internet, d'autant plus que cette dernière est à la fois diverse, variée et mise à jour en temps réel. Pour cela, on utilise le web mining : il s'agit de l'ensemble des techniques visant à explorer, analyser et traiter les grands volumes de données issus d'une activité sur internet. L'exploration du web peut être divisée en trois parties :

- L'utilisation de la recherche sur internet : processus* concernant l'exploitation et l'analyse d'information stockée soit sur les fichiers logs* des serveurs ou celles fournies par les différents intervenants (site, forum, FAI, etc.)
- L'extraction de contenus d'internet : processus* se focalisant plus sur les informations contenues dans des documents stockés sur internet et leur extraction.
- La structure de l'exploration sur internet : processus* d'analyse des relations entre documents, ou pages stockés sur internet.

Cependant une problématique survient : comment reconnaître une donnée pertinente parmi le volume recueilli ? Pour cela, l'utilisateur peut définir une stratégie pour rapatrier les données depuis le web selon une profondeur de recherche, un type d'objets spécifique ou des sites à exclure, etc.

Du côté du propriétaire du site, ce qui est intéressant pour lui c'est d'exploiter les traces des visiteurs. En effet, en plus de l'heure et la date de la visite, chaque utilisateur laisse un

numéro de machine, l'identifiant du navigateur utilisé, l'ensemble des actions réalisées et des pages visités.

Ainsi le propriétaire du site, s'il exploite correctement ses données sera à même de les utiliser pour du web marketing. Nous aborderons d'ailleurs ce processus* ultérieurement.

E. Les outils informatiques existants : étude et comparatif

En tapant « software »(logiciel en anglais) et « Data Mining » dans un moteur de recherche internet, il faudrait plusieurs jours afin de dépouiller les résultats. En effet, ces dernières années, le domaine de la fouille de données a pris de plus en plus d'importance. Des entreprises ont en fait leur cœur de métier et des logiciels Open source* et propriétaire ont vu le jour. C'est devenu un domaine de concurrence féroce.

Nous allons, tout comme pour la gestion de projet aborder, un logiciel Open source*, RAPID MINER et un logiciel propriétaire IBM SPSS

1. OPEN SOURCE* : RAPIDMINER

Considéré comme leader mondial open source* pour la fouille de données, RapidMiner se décline soit en application autonome pour l'analyse de donnée, soit en module à intégrer dans ses autres produits pour l'extraction des données.

Utilisé dans plus de 40 pays, il inclut tout un ensemble de fonctionnalité permettant l'application des méthodes de Datamining de la préparation de données à l'évaluation des résultats.

Multiplateforme, il présente les résultats sous différentes formes : Graphique, nuage de points histogramme, camembert, etc. Cet outil s'adapte à l'utilisateur afin de rendre les résultats explicite et compréhensible.

Bien qu'open source*, ce logiciel se décline aussi en version payante pour les entreprises. Les prix n'étant pas clairement affichés sur le site, il faut prendre contact avec l'entreprise afin de les obtenir.

2. PROPRIÉTAIRE : IBM SPSS

Précédemment, nous avons abordé l'implication d'IBM dans la société américaine. En fournissant des logiciels d'analyse prédictive à la police, notamment de la ville de Memphis, cette société a su se mettre en avant et attirer l'attention sur ces produits.

La suite logicielle qui a permis de baisser le taux de délinquance et de vol à main armée est la suite IBM SPSS. C'est celle que nous présenterons, en tant que solution informatique propriétaire.

Logiciel d'extraction de données et d'analyse, il dispose d'une interface visuelle permettant aux utilisateurs d'exploiter rapidement tout le potentiel des algorithmes sans aucune programmation préalable.

Bien qu'il s'est illustré dans un domaine social, via la lutte contre le crime, le logiciel IBM SPSS est utilisé dans des domaines très différents notamment : La relation client (CRM *), la détection de fraude et de prévention, la gestion des risques, la gestion et l'amélioration de la qualité des soins, etc.

Multiplateforme, cette solution se distingue en deux éditions : SPSS Modeler Professionnel et SPSS Modeler Premium. La seconde plus complète que la première, offre des modules d'analyse de texte, d'entité et de réseaux sociaux.

Selon le type de licence souhaité, les prix du logiciel avec un support de 12 mois varient entre 21 000 euros et 75 000 euros avec un pic possible à 120 000 euros.

3. LES TECHNOLOGIES COMMUNES

Les grandes entreprises du secteur du Datamining, ont toutes des solutions performantes à proposer à leur client.

Microsoft propose donc SQL server 2012, alors qu'Oracle propose Oracle Data Mining.

La différence entre ces solutions est le choix de la technologie. Selon le but recherché, le choix va se porter sur une technologie ou un ensemble de méthodes bien particulières. Les plus répandus sont : les réseaux artificiels de neurones, les arbres de décision et la méthode du plus proche voisin. Chacune de ces techniques analysant les données de différentes façons.

Les réseaux de neurones : modèles non linéaires, permettant la prédiction, ils permettent l'apprentissage par le biais de la formation. Bien que les techniques de modélisation prédictive soient puissantes, elles sont souvent difficiles à mettre en place et à déployer. Le domaine d'application le plus courant est la lutte contre les fraudes. En raison de leur complexité, ces techniques sont plus efficaces dans des situations où elles peuvent être

appliquées plusieurs fois de suite, telles que lors des examens des transactions par carte de crédit qui ont lieu tous les mois pour rechercher les anomalies.

Les arbres de décisions sont des structures en forme d'arbre qui représentent des ensembles de décision. Ces décisions produisent des règles, qui sont ensuite utilisées pour classer les données. Les arbres de décision sont la technique privilégiée pour construire des modèles compréhensibles. Elles peuvent être utilisées pour évaluer, par exemple, si l'organisation a une stratégie de marketing appropriée au vu des profils des clients.

La méthode du plus proche voisin utilise la classification : elle étudie les enregistrements de données et les classes par ensemble de données semblables. Les commissaires aux comptes peuvent utiliser cette approche pour définir par exemple ce qu'est un document intéressant pour eux et ainsi demander au système de rechercher des éléments semblables.

Chacun de ces modèles présente des avantages et des inconvénients qui doivent être pris en considération avant de les utiliser. Les réseaux de neurones, par exemple, nécessitent de renseigner toutes les situations possibles et leur résultat pour être représenté numériquement. La technique d'arbre de décision est la plus utilisée, car simple à mettre en œuvre et directe. Enfin, la méthode du plus proche voisin s'appuie plus sur les liens entre les objets similaires et, par conséquent, fonctionne mieux pour l'extrapolation plutôt que des demandes de renseignements prédictifs.

Ce domaine est en constante évolution, car il est lui-même à la jonction de domaines qui évoluent sans cesse. Il est donc important pour les clients de réaliser une réelle étude de marché et bien connaître leur besoin avant de réaliser un quelconque investissement.

F. Le Datamining : Exemple d'utilisation dans un projet

Les exemples d'utilisation, les plus courants pour le Datamining sont l'email marketing et le web analytic.

1. E-MAIL MARKETING

Le but de ce type de marketing est de fournir un avantage concurrentiel à l'entreprise qui déploie cette solution.

Grâce au Datamining, l'entreprise arrive à détecter des besoins et des opportunités commerciales. En effet, l'Email marketing permet la mise en place d'une stratégie de segmentation des bases de contact. Ainsi ce ciblage des campagnes de publicité est plus intelligent, ce qui permet d'augmenter le ROI de l'entreprise.

Les techniques de Datamining interviennent au niveau des données afin d'identifier celles pertinentes pour le commercial : qui a acheté quoi ? Quel type d'article a-t-il consulté avant l'achat ? Combien de fois est-il revenu sur le site avant l'achat ?

Identifier les groupes d'individus ayant des comportements similaires est facilité et ainsi le ciblage publicitaire aussi.

Grâce aux techniques de Datamining, ce travail laborieux qui incombait aux marketeurs est non seulement plus rapide, mais aussi plus précis. De ce fait, les prises de décisions sont améliorées puisque, d'une part, le risque de se tromper dans la classification d'un individu est minimisé et d'autre part, les décideurs ne se basent plus sur des intuitions, mais sur des faits concrets.

L'email marketing permet donc d'identifier des profils types de contacts :

- les clients et prospects à forte valeur ajoutée (consommateur avéré),
- les clients mono commande (un article à la fois) ou multi commandes (ceux achètent toujours plusieurs articles pour rentabiliser les frais d'envois)
- les clients VIP, ceux qui prennent toujours les abonnements premium
- etc.

Cependant, ce n'est pas sa seule utilité. En effet, appliquer les techniques de Datamining aux données d'un site web permet aussi d'anticiper des phénomènes qui en temps normal n'aurait pas pu être prédit. Il s'agit par exemple, d'un désabonnement, ou d'un départ chez la concurrence ou encore l'achat de produit bien spécifique (non habituel).

Si l'on considère que fidéliser un client revient moins cher que d'en acquérir des nouveaux, toutes les entreprises ont un intérêt vital à détecter et comprendre pourquoi ses clients s'en vont.

Le Datamining le permet. En établissant une carte des accès des utilisateurs passés s'étant désabonnés par exemple, il arrive à « comprendre » et trouver les « raisons » (phase

d'extraction des connaissances) qui ont abouti à de tels actes. Ainsi lorsqu'un client répète le même type de comportement, il sera classé parmi les « futurs désabonnés ». Grâce à cela, les marqueteurs peuvent intervenir rapidement en envoyant une publicité ciblée ou une autre action marketing et ainsi « reconquérir » le client.

Du côté du client, ce dernier ne se sent plus harcelé ou saturé de mails commerciaux sans intérêt. Il ne reçoit, à la fréquence adéquate, que des mails en relation avec ses intérêts. Ainsi il reviendra plus facilement, vers l'entreprise où il se « sent compris ».

Augmentation du ROI, amélioration de la relation client, opération de ventes croisées et additionnelles plus efficaces et amélioration de l'E-réputation (réputation numérique) de l'entreprise et de la valeur économique du client sont, entre autres, autant de bénéfices que peuvent attendre les entreprises utilisant cette alliance de technologie.

Ainsi par la définition et la réalisation d'opérations ciblées automatisée et envoyée au moment opportun, l'email marketing allié au Datamining, permet, la mise en place de scénario intelligent qui aboutit au développement d'une relation durable avec le client.

Grâce à ses techniques, le Datamining permet aux entreprises de transformer une communication directe digitale en une relation client forte, tout en étant au plus près des nouveaux comportements et attentes des internautes.

Retour immédiat et ajustement rapide des actions directes font donc du marketing un processus* en temps réel.

2. WEB ANALYTIC

Derrière cette notion complexe se regroupe l'ensemble des mesures de collecte et d'analyse des données provenant d'internet dans le but d'améliorer son utilisation.

Allié au Datamining, le Web analytic permet de suivre le ROI des projets digitaux. Il fait alors partie du web mining que l'on a vu précédemment.

Pour toute entreprise, ayant un ou plusieurs projets sur internet, le datamining s'avère nécessaire. En effet, face à des parcours numériques d'achats de plus en plus complexes, au fait que l'on a souvent de multiples consultations avant l'achat, les techniques de la fouille de données permettent de déterminer précisément quelle visite a été plus déterminante qu'une autre dans l'acte d'achat.

Dans l'hôtellerie et le transport, cela est un enjeu crucial. En comprenant, le cheminement qui va amener les utilisateurs à réserver leur place, ces sociétés peuvent anticiper et ainsi prévoir les ventes. Ceci permet, par la suite, une optimisation de la politique de prix et donc un meilleur ROI.

Il est donc important pour une entreprise, comme pour l'Email marketing, de déterminer les différents types de profil de ses clients afin de fournir une offre adaptée à chaque profil.

Pour cela, l'entreprise va récupérer sur internet, toutes les informations transitant en relation avec ses services. Elle pourra ainsi recueillir les informations sur les internautes qui se sont intéressés à ses offres. Soit par les traces laissées à la consultation de produits, soit par une FAQ*, soit par un achat, ou un avis laissé sur un site tiers, l'entreprise va se retrouver avec un amas de volume considérable. En appliquant des techniques de Datamining, et notamment un algorithme de typologie*, la société de services va avoir deux possibilités :

Soit éditer des scores par types de visiteur, soit déduire des types de comportements de consommation.

Dans le premier cas, le scoring* permet de créer des catégories de type « Internaute avec haut potentiel de consommation », « moyen potentiel » et « faible potentiel ».

En réalisant cette classification par offre, l'entreprise peut alors estimer le nombre de clients futurs issus du web et prédire un chiffre d'affaires sur la saison estivale par exemple.

Dans le second cas, la classification est aussi de mise. Mais celle-ci est d'un tout autre genre. En effet, si la première a pour but de déterminer le nombre de clients potentiels avec une marge d'erreur minime, cette seconde classification va porter sur les services en général.

Elle aboutira, par exemple, aux deux catégories suivantes : celle de « Internautes à la recherche d'information avant achat » et une autre « Internaute souhaitant des offres à prix réduit ».

Si la proportion d'internautes dans la première catégorie est très importante, cela implique que les supports actuels de l'entreprise de service (site web, plaquettes publicitaires, flyers, etc.) ne contiennent pas assez d'informations, ou du moins que celles-ci ne sont pas suffisantes et assez explicites. La société fera alors le choix d'ajuster sa communication afin

que les internautes puissent facilement trouver les informations et que ces dernières soient assez déterminantes pour aboutir à un achat.

Face à la crise, les entreprises ont dû mettre en place des offres de plus en plus adaptées à la baisse du pouvoir d'achat des consommateurs. Cependant, qu'est qu'une offre « pas chère », ou une offre « à un bon prix ». Ces notions sont totalement subjectives et dépendent de la perception de chaque individu. Le web mining permet de répondre à ce type d'interrogation. En fouillant le web et grâce aux techniques de Datamining, il en ressortira les souhaits et attentes des internautes. L'entreprise pourra alors comparer ses offres avec les besoins des consommateurs et les ajuster ou diversifier.

Le Datamining appliqué à l'analyse du web se révèle donc un outil décisionnel important permettant de conforter une entreprise dans sa stratégie ou de l'accompagner dans l'élaboration de cette dernière.

Nous avons ainsi abordé le datamining dans son ensemble. Après une présentation de ce domaine, et de son utilité, nous avons vu son impact et son importance grandissante pour les entreprises désirant exploiter leurs données pour augmenter leur ROI, découvrir des connaissances et surtout contrôler le risque lors de la prise de décision.

Maintenant que nous avons éclairci la notion de datamining et tout ce qu'elle implique, nous allons présenter une solution informatique de fouille de données appliquée à la gestion de projet afin de nous rendre compte techniquement des limites et des avantages de ce type de solution. Nous répondrons ainsi à notre problématique de départ.

III. Mise en place de l'architecture informatique : Interconnexion des outils de GP et de Datamining

Nous avons mis en évidence les différences entre un système de gestion de projet traditionnel et un système décisionnel incluant le datamining. Nous allons maintenant apporter une solution open source* afin d'allier les deux ensembles.

Précédemment, nous avons abordé la gestion de projet appliquée au domaine informatique, puis à celui du bâtiment à titre d'exemple.

Dans cette partie, nous nous focaliserons sur la gestion de projet informatique. L'accès à un grand volume de données et aux outils open source* étant plus facile, il s'est avéré plus simple d'apporter une solution dans ce domaine.

Une fois cela exposé, nous tenterons d'étendre nos observations au monde de la gestion de projet dans le Bâtiment afin de mettre en exergue les différences et similitudes ainsi que les contraintes rencontrées.

A. Interconnexion d'outils de gestion de projet et de datamining

Il existe deux solutions possibles si l'on souhaite allier logiciel de gestion de projet et datamining. La première consiste à intervenir directement dans la phase de création du logiciel de gestion de projet afin d'y inclure un module de datamining.

L'avantage de cette solution est l'obtention d'un seul logiciel complet et donc un seul référent si l'on est client et que l'on a des problèmes. Le premier inconvénient majeur est que l'application ne prendra en charge qu'un seul format de données, ceux de l'application. Le second inconvénient serait le prix d'une telle solution : en effet pour que celle-ci réponde aux besoins du client et de l'ensemble de ses services, il faut une solution spécifique. Or celle-ci peut avoir un coût astronomique, sans compter les coûts (en temps et en argent) d'une migration des anciens outils vers le nouveau.

Du point de vue du prestataire informatique, ce dernier devra s'assurer que son équipe de développement possède les notions nécessaires afin de maîtriser et réaliser le module décisionnel dans son intégralité.

La seconde solution consiste, quant à elle, à utiliser les deux applications côte à côte. Là encore deux solutions :

- Soit les deux applications sont connectées à la même base de données.
- Soit l'application de gestion de projet envoie les données sous format Excel, ou autre, à l'application de datamining qui va alors les importer dans sa propre base.

Cette seconde solution permet une gestion indépendante de chaque application. Si l'un subit un dysfonctionnement, cela ne rend pas pour autant l'ensemble non fonctionnel. L'inconvénient majeur est le fait de devoir gérer deux applications au lieu d'une.

Pour plus de malléabilité, nous choisirons la seconde solution citée ci-dessus: 2 environnements distincts, mais s'échangeant les données. De cette manière, il sera plus aisé d'apprécier les résultats fournis par le logiciel de gestion de projet avec et sans les indicateurs révélés par le datamining.

B. Informations et données

Après avoir abordé les manières possibles d'intégrer du datamining, nous allons soulever le problème de la récolte de données. En effet, quel type de données ou d'informations allons-nous récolter ? Et d'où viennent-elles ?

À vrai dire, toutes les informations issues du logiciel de GP sont importantes. Comme nous l'avons vu précédemment, le datamining permet de trouver des corrélations sur un grand volume entre des éléments à priori sans relation. Ainsi nous serons intéressées par toutes les données liées aux métiers du développement informatique.

Dans cette partie, nous allons ainsi présenter le premier environnement que nous souhaitons interfacer : celui de la production de logiciel. Celui-ci, totalement open source, est un exemple type de ce qui se trouve dans la majorité des entreprises.

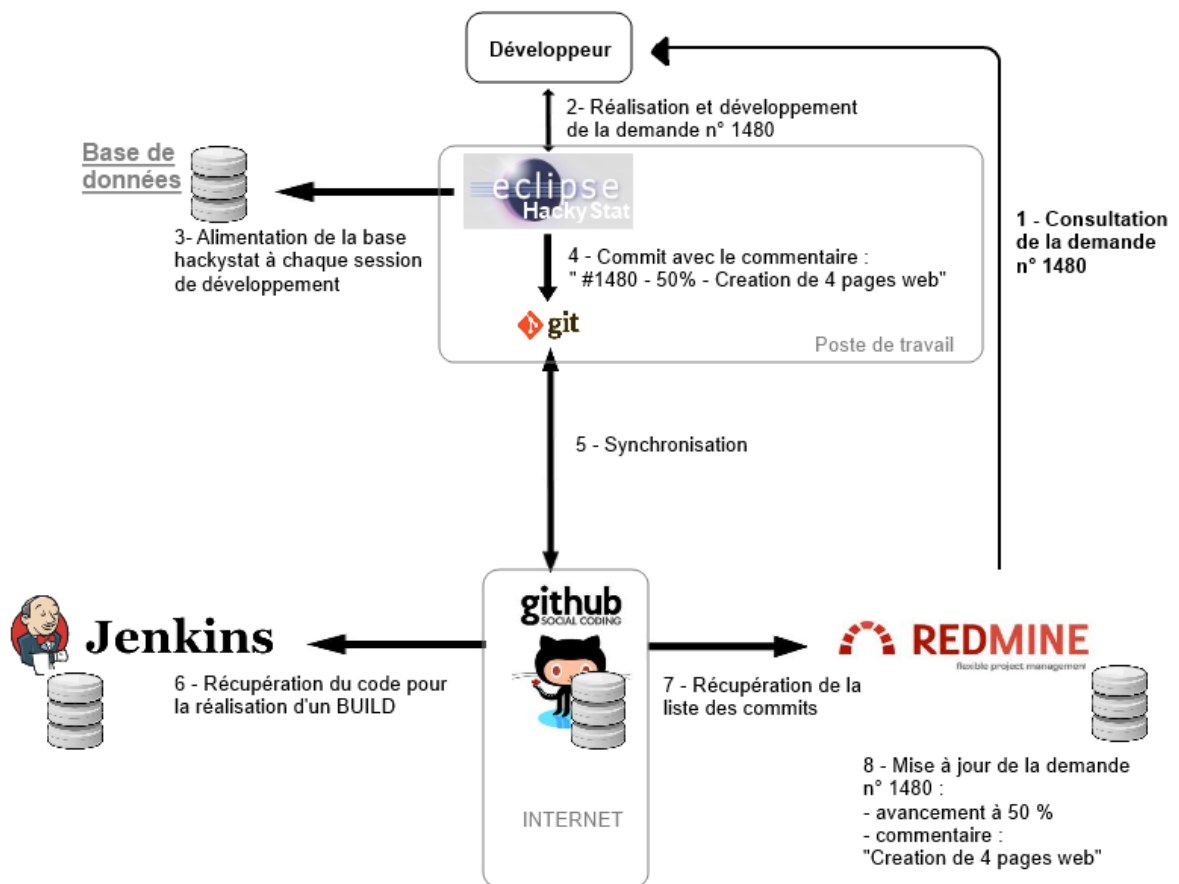


Illustration 9 : Environnement de développement

1. GITHUB : GESTION DES VERSIONS

GITHUB est une plateforme d'hébergement et de gestion des versions des projets en cours de développement utilisant la technologie GIT. Proposant des services gratuits et payants, elle est l'une des références dans le monde du développement Open source*. En effet, les projets stockés sont accessibles à tous et surtout en libre téléchargement pour tous ceux souhaitant s'impliquer dedans.

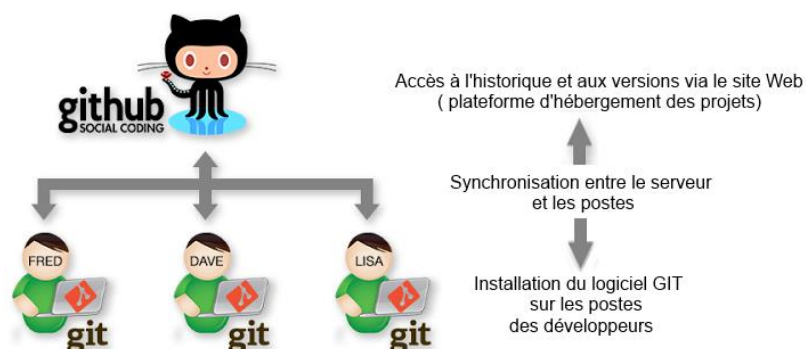


Illustration 10 : Fonctionnement de GIT et GITHUB

Ce site web propose donc un hébergement de projet en utilisant le programme Git. Ce dernier est un logiciel de gestion de version décentralisé : c'est-à-dire que chaque personne participant sur le projet en possède une version complète sur son poste de travail en local. Elle peut donc intervenir dessus sans avoir besoin d'une connexion internet. Celle-ci n'est utile que pour récupérer la dernière version du projet avec les modifications des autres développeurs, ou pour mettre à jour le projet sur le site avec ses propres modifications. GIT permet ainsi de gérer les versions d'un programme, d'alimenter l'historique et de garder une traçabilité des modifications : nous pouvons ainsi savoir qui a fait quoi à quel instant. En effet, il est possible de suivre l'évolution de chaque fichier d'un projet, de tracer toutes les modifications subites et surtout de restaurer n'importe quelle version depuis sa création. Cette gestion des versions est un atout majeur pour les projets nécessitant l'intervention de plusieurs collaborateurs.

GITHUB, quant à lui, en plus d'être le service d'hébergement des répertoires GIT, apporte plusieurs fonctionnalités intéressantes tel que celles liées aux réseaux sociaux. Les développeurs peuvent donc interfacer leur projet avec un compte Twitter ou Facebook et ainsi partager et diffuser encore plus leur réalisation et l'avancée de celle-ci. Indispensable et complémentaires à GIT, il permet l'accès à la base de données relative à l'ensemble des projets. De plus, sa capacité à s'interfacer avec des outils d'intégration continue* comme Jenkins, permet de contrôler l'entière fonctionnalité de l'application à chaque modification. À chaque fois qu'un intervenant « commit », c'est-à-dire qu'il modifie un fichier, l'enregistre et l'intègre dans la plateforme web, l'application est contrôlée dans son intégralité afin de savoir si la nouvelle version du fichier s'intègre bien avec le reste du projet. Si un dysfonctionnement surgit, l'ensemble des membres de l'équipe sont avertis afin que cela soit rectifié rapidement. Nous aborderons JENKINS plus en détail par la suite. À chaque commit, des données sont sauvegardées. Ce sont elles qui nous intéressent. Stockées sur le serveur web, elles constituent un historique et peuvent se révéler riches en informations concernant la vie d'un projet.

Voici un commit réalisé par moi-même sur un projet nommé CineBook :

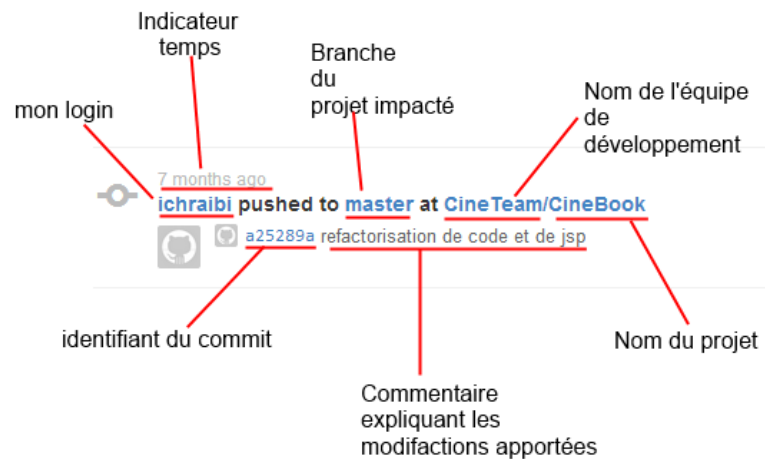


Illustration 11 : Données relatives à un commit

Comme nous pouvons le constater, à chaque commit, GIT sauvegarde : le login du développeur, la date (puisque'il est capable de calculer le temps écoulé), la branche du projet impacté, le nom de l'équipe et le nom du projet. Le développeur a aussi la possibilité de commenter son commit afin d'informer les lecteurs par la suite des modifications apportées. Si nous consultons le commit en lui-même, voici les informations auxquels nous accédons :

refactorisation de code et de jsp

master Cinebook-Iteration8 CineBook-Iteration7

Identifiant complet du commit

Ikram Chraibi authored 7 months ago 1 parent ffc85db commit a25289a7b435b2ed09b157f62666c473ae220b88

Showing 8 changed files with 19 additions and 100 deletions. Nombre de modifications total survenues dans le projet

src/main/java/com/cineteam/cinebook/web/film/RechercherFilmAction.java

```
@@ -22,7 +22,7 @@ public String execute(HttpServletRequest request) {
22 22     String film_recherche = (String) request.getParameter("recherche");
23 23     if(!StringUtils.isEmpty(film_recherche)){
24 24         List<Film> filmsParMotCle = provider.rechercherFilmParMotCle(film_recherche);
25 - request.setAttribute("filmsParMotCle", filmsParMotCle);
26 + request.setAttribute("films", filmsParMotCle);
26 26     }
27 27     return "listeFilms.jsp";
28 28 }
```

src/main/java/com/cineteam/cinebook/web/film/RecupererFilmsVusAction.java

```
@@ -32,9 +32,9 @@ public String execute(HttpServletRequest request) {
32 32     filmsVus = provider.getFilmsParIds(filmsVusParIds);
33 33 }
```

Illustration 12 : Données relatives aux fichiers à chaque commit

Grâce à l'illustration n°12 , nous constatons que GITHUB sauvegarde des données détaillées au niveau d'un fichier.

Ainsi les données qui nous intéresseront seront réparties au niveau des projets, mais aussi au niveau des commits.

Au niveau du projet	Au niveau de chaque commit
<ul style="list-style-type: none"> • Membre de l'équipe • Nombre de commit • Nombre de branche • Nombre de release (livrable*open source* réalisé) • Taux de participation de chaque membre • Pourcentage d'ajout du code* de chaque membre • Pourcentage de suppression de code* de chaque membre 	<ul style="list-style-type: none"> • Responsable du commit • Nombre d'ajout et de suppression • Date du commit • Branche impactée • Liste des fichiers modifiés et liste des modifications sur chaque fichier

Tableau 2: Comparatif entre les données d'un projet et celles d'un commit sur GITHUB

Grâce à ces données, nous aurons donc l'ensemble des éléments relatifs aux projets pour appliquer les données de datamining.

2. JENKINS : INTÉGRATION CONTINUE

Connu à l'origine sous le nom de HUDSON, JENKINS est un outil d'intégration continue conçu et réalisé en langage de programmation* JAVA. Reconnu pour sa simplicité d'utilisation et son efficacité, il s'interface avec tous types de projet, quel que soit le langage de programmation* et le système d'exploitation*. Avant de présenter plus en détail l'outil, nous allons d'abord définir la notion d'intégration continue.

a) Intégration continue (IC)

L'intégration continue est l'un des piliers du développement informatique moderne. Mis en place au sein d'une équipe, elle permet d'améliorer la qualité de travail et de diminuer considérablement le pourcentage d'erreur livré chez le client.

En effet, cette méthode permet de fluidifier le processus* de développement, aide à détecter et à corriger les dysfonctionnements plus rapidement. S'adressant aux développeurs et aux non-développeurs, il fournit un écran de contrôle très utile explicitant les étapes qui se sont bien déroulées et celles plus problématiques.

Concrètement, cela implique que l'IC, dans sa forme la plus simple, se compose d'un outil qui va contrôler les modifications réalisées dans le code* à chaque intégration. Dès qu'un changement est détecté, l'application dans son intégralité sera compilée. L'ensemble des modules et fonctions du projet sera compilé afin de savoir si tout fonctionne : ce contrôle porte le nom de BUILD. Dans le cas de JENKINS, à chaque commit (ou à une fréquence paramétrée), il va télécharger depuis GITHUB l'ensemble du code* afin de le compiler.

Si le BUILD échoue, l'ensemble de l'équipe de développement reçoit un mail précisant le module et l'emplacement du code* problématique.

Ainsi même si la personne responsable n'y fait pas attention ou est dans l'incapacité d'intervenir, le reste de l'équipe sait où intervenir afin de rectifier les choses.

L'équipe possède donc toujours un livrable* valide pouvant être exécuté pour le client.





b) Fonctions de jenkins

Jenkins est un outil offrant de multiples fonctions : en plus, de permettre l'intégration continue pour les projets informatiques, ils offrent aussi une gestion de l'historique des BUILDS, ainsi qu'un graphe traçant l'évolution d'une activité du projet sur une période de l'année.

Au niveau des données, JENKINS est aussi précis que GITHUB. En plus de stocker des informations relatives au projet et à chaque BUILD réalisé, il stocke aussi des données concernant les fichiers modifiés.

Reprenons l'exemple du projet CineBook réalisé par l'équipe CineTeam : il a été configuré pour une synchronisation entre les commits sur GITHUB et les BUILDS sur JENKINS. À chaque commit, un build automatique est lancé afin de vérifier la viabilité des modifications.

Sur le tableau de bord* du projet, voici un exemple de ce que l'on peut observer :

Identifiant auto-généré de chaque build			
CineTeam » CineBook #151	7 mo. 12 j	stable	
CineTeam » CineBook #150	7 mo. 13 j	back to normal	
CineTeam » CineBook #149	7 mo. 13 j	broken since this build	
CineTeam » CineBook #148	7 mo. 13 j	stable	




Période écoulée depuis le build

Etat du projet (livrable)

Illustration 13: Vue globale des builds du projet CineTeam

Cette vue permet de savoir à quel moment, ou durant quelle période, l'application a été ou non fonctionnelle.

Dans un même esprit, mais plus détaillé, voici une autre vue des BUILDS du projet :

#151 (17 déc. 2012 15:31:13) — Date et heure du build	
1. ajout menu_membre en JS+nettoyage imports des jsp (commit: 1099aa74eb049d0b5e34d5413e4312f267b73bd) — Alpha / detail	
2. ajout import javascript  la page d'erreur (commit: f426ec9bc72691bd1878219dc93453d46c8f5f1e) — Alpha / detail	
#150 (17 déc. 2012 10:14:09) — Login du développeur	
1. modif apr  s build faild (commit: 1fd14860ede7cadf074c825e97eed10b654927ec) — Alpha / detail	
#149 (17 déc. 2012 09:29:25)	
1. refactoring nettoyage tests actions (commit: 4d954c6e7e77387d0ffc942e23f6102d1851108e) — Alpha / detail	
2. refactoring tests actions (commit: 8b64a6adacf6d6acca022b1828f9e41f37f49ca0) — Alpha / detail	
3. tests d  comment  s (commit: 6d71a7508e6712a2d7c8d808f6b469fba02ed68e) — Alpha / detail	
#148 (16 déc. 2012 20:20:33) — Commentaire du commit associé	
1. refactoring nettoyage recupererfilmsVus (commit: fa6d7821b9b276d6a4422e7cb0470b40e2638bf0) — Alpha / detail	

Identifiant du commit associé

Illustration 14 : Données relatives à un build d'un projet

Nous constatons ainsi que JENKINS permet de sauvegarder les données suivantes :

Au niveau du projet	Au niveau de chaque commit
<ul style="list-style-type: none"> Nom de l'équipe Dernier BUILD réussi et sa date Dernier BUILD échoué et sa date Nom de la branche du projet qui est sur JENKINS 	<ul style="list-style-type: none"> Responsable du BUILD Identifiant du commit associé Date et heure du BUILD Durée du BUILD Commentaires du commit (repris pour le BUILD)

Tableau 3 : Comparatif entre les données d'un projet et celles d'un build sur JENKINS

Ce sont celles-ci qui nous intéresseront pour l'extraction par la suite.

3. REDMINE : GESTION DES PROJETS

Comme nous l'avons présenté précédemment REDMINE est un logiciel de gestion de projet basé sur la création de demandes ou de tickets qui sont ensuite attribués à une personne.

La demande peut être de différents types : la correction d'un dysfonctionnement (bug), une demande d'évolution (ajout d'une nouvelle fonctionnalité) ou encore un ajustement à apporter. Selon le but recherché, il peut être un outil de suivi des problèmes ou « bug tracker » et un outil de suivi des évolutions ou encore un outil d'évaluation de la productivité des programmeurs.

Réalisé avec le langage de programmation* RUBY, il permet de s'interfacer avec d'autres outils afin de faciliter la gestion et le suivi des projets. Dans notre cas de gestion de projet informatique, c'est l'alliance GITHUB et REDMINE qui sera intéressante.

Expliquons cela : lorsqu'un développeur voit une demande lui être attribuée, il va alors la réaliser. Une fois le développement réalisé, il va alors notifier cela sur l'application et ainsi de suite.

Dans cette configuration-là, l'environnement de développement et celui de gestion du projet sont totalement dissociés. Les données sont donc dispatchées et non synchronisées. De plus, rien ne garantit que le développeur renseigne correctement REDMINE avec ce qu'il a réalisé ou qu'il en aura le temps tout simplement.

Dans notre solution, nous préconiserons un interfaçage entre l'outil de gestion de projet et celui de gestion de version. Nous espérons ainsi mettre en place le scénario suivant :

Lorsqu'un développeur recevra une demande, il la réalisera normalement. Les choses changent au moment où il intègre (commit) ses modifications sur GITHUB : en effet, au lieu de renseigner un commentaire quelconque, il devra suivre une norme prédéfinie de la forme suivante :

« Identifiant de la demande – Option 1 – Option 2 – commentaires »

Grâce à cette norme, il sera possible de paramétrer* REDMINE afin qu'il récupère la liste des commits réalisés. Il pourra ainsi mettre à jour les demandes qui ont été travaillées ainsi.

Les option1 et 2 sont des mots-clés qui permettront par exemple de clôturer, d'annuler ou de modifier l'avancement d'une demande (à 50 % par exemple).

L'avancement sera donc mis à jour presque en temps réel : la synchronisation se fait à intervalle de temps régulier qu'il faudra choisir au préalable.

Grâce à cela, le volume de données est important, synchronisé (le numéro du commit est renseigné dans la demande) et surtout mis à jour. Les techniques de datamining seront plus performantes. Les données seront récupérées selon deux axes : celui des projets et celui des demandes associées à chaque projet.

Au niveau du projet	Au niveau de chaque commit
<ul style="list-style-type: none"> • Nom du projet • Nom de l'administrateur • Nom des membres de l'équipe • Liste des annonces (communiqués) • Liste des commentaires de chaque annonce (intervenant, date et heure, etc.) • Liste des demandes • Historique de chaque demande • Code* du projet 	<ul style="list-style-type: none"> • Identifiant • Libellé • Statut • Priorité • Responsable • Catégorie • Date de début • Échéance • Pourcentage d'avancement • Description • Liste des commits • Lien vers le code* du projet en relation avec la demande

Tableau 4 : Comparatif entre les données d'un projet et celles d'un build sur JENKINS

Une série d'illustrations sont présentes en Annexe 3 afin d'explicitier l'outil REDMINE et les données qu'il fournit.

4. IDE, INTEGRATED DEVELOPMENT ENVIRONMENT : LES OUTILS DE DÉVELOPPEMENTS

Pour les entreprises commercialisant des logiciels, la programmation informatique - soit la production de code* - est la base même de leur activité. Plus celle-ci est maîtrisée,

mieux c'est. La productivité des programmeurs est donc un élément clé, cependant ce n'est pas le seul. En effet, la qualité du code* (respect des normes, absence de bug) est aussi déterminante pour l'équipe que le client. Pour cela, le choix de l'outil de développement est crucial.

Nous allons, ainsi, aborder les IDEs, principal outil de programmation. Ensuite, nous verrons à travers un exemple, quelles sont les données que l'on peut récupérer lors de la réalisation du logiciel et qui pourrait servir pour une gestion de projet informatique beaucoup plus sûr.

a) Définition

IDE signifie *Integrated Development Environment*, à savoir Environnement de développement intégré. Il s'agit de l'outil principal qu'utilisent les développeurs lors de la création de logiciels.

En réalité, un IDE est un regroupement d'une multitude d'outils aidant les développeurs dans leur tâche quotidienne afin d'augmenter leur productivité. Ces outils fonctionnent ensemble et permettent une automatisation de tâches complexes qui facilitent les opérations lors du développement. Payant ou gratuit, spécifique à un langage de programmation* ou très généraliste, il existe de nombreux outils de développements sur le marché. Le choix d'un IDE pour une entreprise se fait soit en raison d'une stratégie particulière, soit en raison d'un langage de programmation* de prédilection ou d'un projet, soit en raison des fonctionnalités intéressantes que l'entreprise souhaite utiliser.

Globalement, ils comportent tous les fonctionnalités suivantes :

- un éditeur de texte : Il permet la saisie du code*, squelette de l'application, et utilise souvent un système de mise en couleur du texte afin de mettre en avant des éléments tel que le nom de procédures* ou fonctions*. Le but étant toujours de rendre les choses plus faciles et accessibles pour un développeur.
- Un compilateur est un traducteur : il traite le code* source rédigé dans un langage de programmation* afin de le transformer en un autre langage, celui de la machine. Cette dernière peut alors comprendre le code* et ainsi exécuter le logiciel. C'est grâce à cet outil que les développeurs peuvent vérifier la bonne syntaxe de leur code* et donc l'existence ou non d'erreur.

- Un débogueur : Analysant le code* source, il permet d'exécuter le programme pas à pas afin d'aider le développeur à localiser les bugs. En affichant explicitement les données transitant dans le programme, il permet de vérifier leur valeur et donc la logique établie.

Les IDE permettent de créer des fichiers comportant parfois des centaines de lignes, et des logiciels composés à leur tour de centaines de fichiers. Il est donc important d'avoir des outils d'aides à la programmation performants.

La plupart des IDE acceptent l'ajout d'options supplémentaires grâce notamment des plugins. Il s'agit d'un petit logiciel ou paquet qui vient se greffer à l'IDE principal afin d'apporter des fonctionnalités supplémentaires. Ne pouvant fonctionner seuls, la présence d'un logiciel hôte est nécessaire à leur mise en fonction.

GIT, par exemple, existe soit au format standard d'une application à installer, soit au format plugin. Il peut alors être intégré à un IDE et en faire ainsi partie à part entière.

Dans le cadre de notre architecture totalement OPEN SOURCE*, nous nous baserons sur l'IDE Eclipse pour la suite.

b) ECLIPSE, un exemple d'IDE



Illustration 15 : Logo de l'IDE Eclipse

Totalement Open source*, Eclipse, à l'origine initié par IBM, est un projet de la fondation ECLIPSE. Pensé et créé pour être extensible, universel et polyvalent, cet IDE s'appuie principalement sur le langage de programmation* JAVA, même s'il en supporte beaucoup d'autres.

Il est à la fois un IDE permettant des activités de programmation, mais c'est aussi un atelier de génie logiciel* et un Framework*. Cela signifie, respectivement, qu'il permet la production de programmes de manière quasi industrielle et qu'il fournit un ensemble de composant logiciel permettant de bâtir l'architecture (les fondations) d'un logiciel tels les librairies et les API*.

Décliné et organisé en un ensemble de sous-projets de développements logiciels, il est basé sur la notion de plugin : toutes ses fonctionnalités ont été créées en tant que plugin.

Grâce à cela, les développeurs choisissent d'inclure ce qui les intéresse, et ne sont pas contraints en terme de volume ou limité en terme d'options. Il permet même l'ajout de plugins « étrangers », soit développés par des personnes extérieures à la fondation.

Apprécié des développeurs aussi bien débutants, qu'expérimentés, il est utilisé dans le monde académique et professionnel. Aujourd'hui, environnement de développement libre, il est l'une des réussites du monde de l'Open source*.

c) PSP : Processus de développement personnel ou Personnel Software Process*

Le PSP a été initialement pensé afin d'aider les ingénieurs en développement à comprendre leurs lacunes et donc améliorer leurs performances. En utilisant, une procédure rigoureuse pilotée par les données récoltées pendant les phases de programmation, les ingénieurs peuvent alors prendre conscience de leur évolution et mettre en place des axes de progression. Ils peuvent ainsi perfectionner leur estimation de temps et donc leur planification, améliorer la qualité de leur projet et enfin réduire le nombre de défauts présent dans l'application.

Le but du PSP est donc de permettre aux développeurs à produire des logiciels de qualité absolue* (zéro défaut) dans les temps impartis en améliorant leur propre processus* de programmation.

Conçu initialement pour un accompagnement individuel, le PSP se décline aussi pour un groupe d'individu : c'est le TSP pour Team Software Process ou Processus* de développement en équipe, visant à améliorer le développement en équipe.

Comme cité précédemment, « la programmation informatique - soit la production de code* - est la base [l'] activité. Plus celle-ci est maîtrisée, mieux c'est. ». Limiter le risque lors de la gestion de projet informatique est en relation directe avec l'activité de l'équipe qui s'occupe du projet. Récolter des données sur les actions de chaque développeur dans l'équipe, permettra de mieux connaître les capacités de chacun, son évolution et ainsi mieux répartir les tâches et intégrer les personnes sur les projets.

Le PSP est donc au centre de la gestion du risque dans gestion de projet informatique.

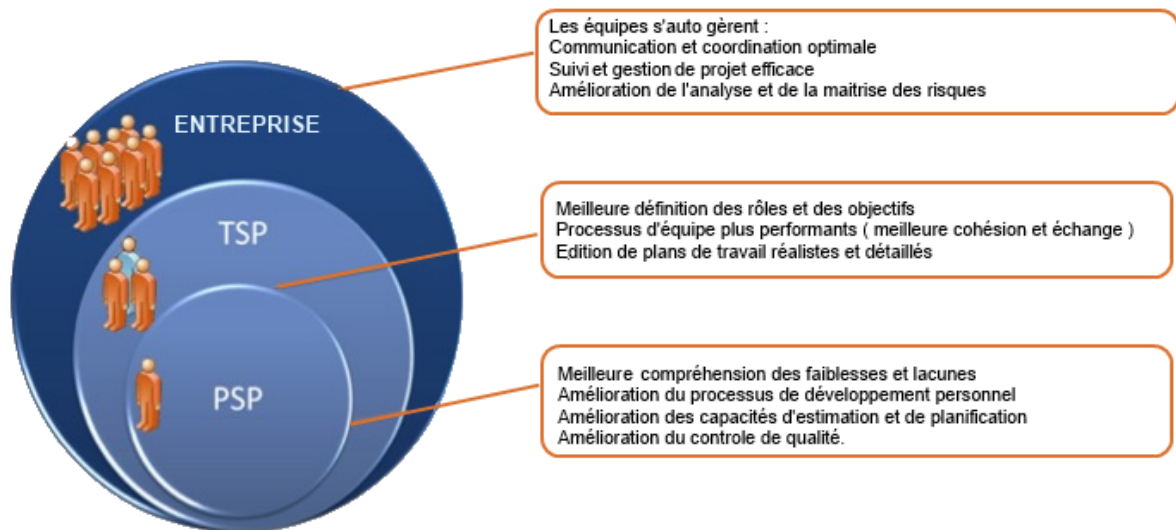


Illustration 16 : le PSP au centre de l'amélioration de l'entreprise

Basé sur un système de mesure simple et efficace, le PSP vise à récolter les données lors de la phase de réalisation du logiciel. Cela se fait dans un journal daté, où sont retranscrits les défauts, les erreurs de compilation, la fréquence d'exécution du programme et l'ensemble des activités du développeur. Les données sont traitées, stockées puis analysées afin d'en faire ressortir des tendances.

La récolte des données se fait directement sur l'IDE en question. C'est cette partie que nous allons aborder ensuite.

d) Les données : LOGGING et plugging Hackystach

Pour un programmeur, il existe deux types de données qui sont intéressantes : les premières sont celles liées à l'exécution du programme développé que ce soit en cours de réalisation ou une fois en production. Les secondes sont celles liées à la conception du logiciel et donc aux méthodes du programmeur.

Le premier type de données concernent la vérification de la logique et du code* : cela permet de savoir si l'application a bien été pensée, d'où viennent les bugs et comment les résoudre. Bien qu'intéressantes, elles sont plus en relation avec le produit fini qu'avec le programmeur. À cause de cela, nous nous focaliserons plus sur les données liées au PSP du

développeur. Car c'est en améliorant ce paramètre que l'entreprise gagne un collaborateur compétent et efficace sur du long terme.

La récolte de données liées à cela peut se faire de deux manières : Le LOGGING ou l'utilisation de plugin. Nous allons ainsi aborder ces deux solutions avant de déterminer laquelle sera mise en place dans notre architecture finale.

e) **LOGGING**

Initialement le logging, consiste à ajouter des traitements dans une application afin de permettre l'émission et le stockage d'informations suite à des événements. Ce stockage se fait alors dans un fichier log avec une extension « .log ». Il s'agit de fichier texte simple consultable avec n'importe quel éditeur de texte, et dans lesquels sont enregistrés chronologiquement des événements détaillés. Utile et nécessaire en informatique, il permet de déboguer* plus rapidement, de stocker les traces d'exécution (information, avertissement) et surtout de comprendre et vérifier le flux des traitements exécutés.

Son importance est proportionnelle à la taille du projet et au nombre d'intervenants.

L'idée, dans le cadre de notre architecture, serait de mettre en place une gestion des versions des logs*. En définissant des normes pour les informations enregistrées dans les fichiers logs*, on peut déterminer l'action, l'heure, la cible de l'action, le nombre d'erreurs de compilation et tout un tas d'informations.

Ainsi on pourrait obtenir un historique des fichiers logs* et déterminer l'évolution du développeur en étudiant les indicateurs qui en ressortiront.

Grâce à ces fichiers et leur historique, il sera possible d'établir des statistiques tel que : quel est l'erreur de compilation qui survient le plus souvent ? Déterminer les actions que fait fréquemment le développeur et dans quel ordre. Ces fichiers pourront apporter aux programmeurs un regard objectif à leur processus* de développement et voir même alimenter des indicateurs tels que ceux du succès en programmation, ceux du respect des conventions de nommage, de la bonne utilisation des capacités de l'IDE et celui lié à la fréquence d'exécution (la compilation étant continu sous Eclipse).

Bien qu'intéressante, l'utilisation du logging est contraignante : il faut dans un premier temps paramétrer* Eclipse pour récupérer les informations que l'on souhaite et si cela n'est pas possible, le développer nous-mêmes.

De plus, les messages à ajouter dans les fichiers doivent être définis précisément. Ensuite, il faudrait traiter l'ensemble des logs* et leur historique afin de transcrire les données dans une base de données. Laborieuse, la mise en place de cette solution prendrait beaucoup de temps surtout qu'elle peut dégrader les performances de l'IDE, de l'application et donc les conditions de travail.

Nous allons présenter une autre méthode qui, reprenant l'idée du LOGGING, s'avèrera moins contraignante.

f) PLUGING : HACKYSTAT

Hackystat est un Framework* open source* pour la collecte, l'analyse, la virtualisation et l'interprétation de données liées au processus* de développement personnel.

Une fois « attaché » à un IDE, il va discrètement (sans ralentir le développement) recueillir et envoyer des données brutes. Celles-ci, très détaillées, couvrent un ensemble d'éléments : le temps des sessions de travail, le temps passé par tâche, les échecs d'exécution, leur fréquence, la liste des événements réalisés, le respect des normes de conventions, etc.

Les données peuvent être envoyées à un service web appelé le HackyStat SensorBase, pour un stockage en ligne des données, ou être directement stockées sur un serveur de récolte et stockage de donnée.

Dans le premier cas, le répertoire de SensorBase peut être interrogé par d'autres services web pour former des abstractions* de haut niveau. Ces dernières pourront être ensuite intégrées à d'autres services pour être communiquées ou incluses dans des mécanismes d'analyses et de génération de reporting*.

Dans le second cas, les données stockées pourront directement être traitées par un algorithme de fouille de données.

Un des buts de cet outil est de faciliter « l'intelligence collective » dans le domaine du développement logiciel en permettant la collection, l'annotation et la diffusion d'informations.

Grâce à cela, les analyses pertinentes et les connaissances utiles sont accessibles à tous et chacun peut profiter de l'expérience d'autrui.

Les services hackystat sont conçus pour coexister et compléter d'autres composants et services d'un système d'information dédié au génie logiciel*.

En effet, existant sous la forme de plugin, il peut être interfacé avec ECLIPSE, ce que nous allons exploiter pour l'environnement que nous souhaitons mettre en place.

Il est important de souligner, cependant, que ceci ne justifie pas entièrement notre choix de hackystat.

Des études et expériences menées par le professeur Gregory Dyke de l'École des mines de Saint-Étienne, montre que l'on peut pousser la récolte de données assez loin grâce à cet outil. En effet, étant donné qu'il est open source*, il offre la possibilité à ceux qui le souhaitent de le modifier afin de collecter des informations beaucoup plus précises. Il a ainsi déjà été modifié pour inclure le traçage de l'activité de la souris, l'activité du clavier, les données liés à l'exécution de l'application, et celles liées à l'utilisation même des fonctions propres à l'IDE.

Grâce à cela, il est donc possible de savoir tout ce que le développeur fait pour le projet sur l'IDE.

Dans le cadre d'un travail en binôme, comme ce fut le cas pour l'étude menée, cela rend possible de savoir quel sont les binômes synchronisés, ceux disparates, et le type de problèmes rencontrés de manière fréquente. Les raisons d'échec peuvent être plus facilement cernées.

Au vu des possibilités qu'offrent cet outil, nous allons choisir cette solution-là plutôt que celle du LOGGING. La récolte de grands volumes de données à partir de l'IDE Eclipse sera ainsi possible sans altérer les performances de la machine.

Nous venons de voir le premier environnement dit de développement : nous avons ainsi mis en évidence l'importance de chaque outil dans le processus* global et surtout la vision et les données qu'il apporte.

GITHUB et Jenkins apportent une vision instantanée du projet à un instant t. Telle une photo, ils permettent d'établir un état des lieux global, mais détaillé du projet : Avancement, fonctionnalités, mais aussi intervenants, lignes de code* modifié, etc.

REDMINE, quant à lui, apporte une vision discrète de l'avancement de l'application. Pour chaque tâche, le développeur peut intervenir plusieurs fois dessus et ainsi à chaque fois, il renseignera sa demande Redmine pour donner son avancement. On sait ainsi, combien de temps, la tâche a réellement pris, qui est intervenu, les contraintes, etc.

Enfin, l'IDE Eclipse, apporte une vision continue et dynamique du projet. A chaque session de développement, on recueille continuellement les actions et événements réalisés.

Toutes les données générées, amas volumineux et conséquent, doivent ainsi nous permettre de mettre en évidence, via des techniques de datamining, des indicateurs clés pertinents pour la gestion du risque dans les projets informatiques.

OUTILS	VISION APPOREE
GITHUB et JENKINS	Vision instantanée à un instant t Comme une capture d'écran
REDMINE	Données discrètes
IDE	Vision dynamique Aspect continu

Tableau 5 : Vision apportée par les composants de l'environnement de développement.

C. Solutions d'extraction et de stockage des données

Nous venons de voir, les sources des données que nous allons traiter. Maintenant, nous allons nous pencher sur leur stockage et gestion. Précédemment, nous avons affirmé envisager d'utiliser une solution séparant les bases de données opérationnelles de celles liées au décisionnel. Nous allons ainsi aborder l'outil clé pour la gestion d'un grand volume de données : Le DATAWAREHOUSE.

1. DATAWAREHOUSE

Grâce à GITHUB, nous allons pouvoir récupérer un grand volume de données diversifié.

La problématique qui se pose alors est la suivante : où allons-nous stocker ces informations ? En effet, le datamining nécessite d'accéder aux données, qu'il va ensuite transformer avant de les traiter. Les données récupérées depuis GITHUB doivent être stockées afin que l'accès

soit facilité. De plus, lorsque le processus* de datamining va générer de nouvelles connaissances, il va les inclure dans la base de données. Or nous ne pouvons « écrire » nos nouvelles données dans la base de données GITHUB puisque nous n'avons pas les droits. Il nous faut donc stocker les données de telle sorte à ce que l'on est les droits dessus. Nous verrons par la suite que ce n'est pas le seul avantage d'un DATAWAREHOUSE (DW).

a) Définition

Dans une entreprise, toutes les données issues des activités de l'entreprise sont stockées dans une base de données. Celle-ci est dite opérationnelle, car elle est en relation directe avec les opérations réalisées par l'entreprise (achat, vente, développement, construction, etc.)

Lors d'un processus* de prise de décision, les décideurs ont besoin, en plus de données mises à jour et complètes, de données synthétiques se présentant sous une forme intelligible (tableaux, graphiques, courbes, indicateurs, etc.). L'obtention de tel résultat nécessite la réalisation de requêtes assez gourmandes en ressources et en temps. Celles-ci exécutées sur la base de données opérationnelle de l'entreprise, ralentirait, voir bloquerait le travail des collaborateurs. De plus, rien ne garantit que les données soient complètes à cet instant-là. Ainsi, afin d'éviter cela, la plupart des entreprises ayant un fort besoin d'outil de décisionnel ont fait le choix d'avoir un Datawarehouse.

Dédié à l'informatique décisionnelle, il est aussi appelé Entrepôt de données. C'est en réalité une base de données « orientées sujet » regroupant une partie ou l'ensemble des données fonctionnelles d'une entreprise. Techniquement, il permet de délester la base de données opérationnelle (celle des systèmes de production) de l'entreprise de requêtes qui pourrait nuire à sa performance.

Un DW regroupe des données orientées sujets, intégrées, historisées et non volatile. Il est organisé et pensé pour un accès rapide et synthétique à une information stratégique pour la prise de décision.

Orientées Sujet :

Dans une base de données opérationnelle dédiée aux systèmes de production, les données sont regroupées par processus* ou système d'information (SI) fonctionnel voir par thème : celles liées aux contrats sont ensemble, celles liées à l'achat aussi, etc.

Le problème d'une telle vision est qu'elle n'est pas du tout transversale. De plus, les données peuvent être redondantes entre les différents outils de l'entreprise et surtout déphasées les unes par rapport aux autres.

L'intégration des données dans un DW implique leur unicité. En effet, afin que le processus* de décision puisse se faire clairement, les données relatives à un même sujet ne sont intégrées qu'une seule fois. Les données sont donc uniques et surtout orientées sujet au lieu d'être orienté opérations. Ainsi la base de données est construite selon les thèmes qui touchent aux métiers de l'entreprise (clients, produits, risques, rentabilité,...).

Données intégrées

Les données intégrées dans un DW proviennent, selon la taille de l'entreprise, soit d'une seule base de données opérationnelle, soit de plusieurs SI. De plus, elles peuvent provenir de plusieurs supports possibles et exister sous toutes formes possibles (fichier Excel, document Word, mail, Base de données locale, etc.)

Il faut donc les homogénéiser afin qu'elles soient compréhensibles par tous les utilisateurs : un rassemblement et une mise en forme sont donc nécessaires avant l'intégration finale.

Cette première phase d'uniformisation permet entre autres de garantir que les données sont uniques. Grâce à des filtres et des vérifications, l'information est intégrée une seule fois : l'unicité est donc garantie.

Cette phase nécessite la définition d'une norme et aussi celle d'un référentiel unique.

Chaque donnée ajoutée aura un identifiant unique et une description. Grâce à cela, à chaque nouvel ajout, on saura si une donnée a déjà été ajoutée et on évitera les doublons.

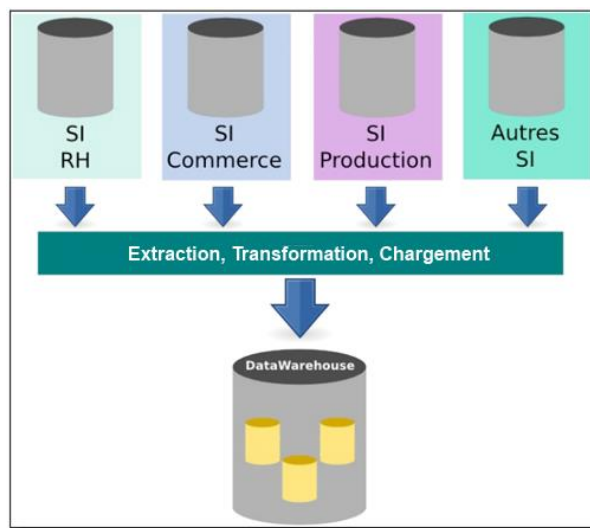


Illustration 17: Concentration de l'information de l'entreprise dans un Datawarehouse

Données historisées

Une donnée ajoutée au DW est unique, mais aussi figée dans le temps. Alors que dans un système de production, les données sont régulièrement mise à jour, dans un DW, il en est de tout autre. À son intégration dans le DW, un marqueur temps est ajouté à la description de la donnée. Les ajouts de valeur supplémentaires de cette même donnée font l'objet de nouvelles insertions et non pas de mise à jour. Grâce à cela, il est possible de tracer l'évolution d'une donnée dans le temps. Par exemple, l'évolution du nombre de BUG traités par un développeur au cours des derniers mois, ou le chiffre d'affaires d'un point de vente. Représentant parfois 60% à 80% de la charge totale d'un projet de mise en place d'un DW dans une entreprise, la phase d'intégration est l'une des plus complexes.

Données non volatiles

Une mémoire non volatile est une mémoire qui conservera ses données tout le temps à partir du moment où celles-ci ont été intégrées et cela même s'il y a eu une coupure de courant. Ex. : le disque dur d'un ordinateur.

Dans le cas d'un DW, des données non volatiles, sont des données qui seront conservées à jamais, jusqu'à ce qu'elles soient supprimées sciemment (mais ce n'est pas le but).

Cette caractéristique est une conséquence du besoin de traçabilité des données et de leur évolution. Ainsi, une donnée à un même instant t aura toujours la même valeur et cela, quel que soit le temps écoulé.

b) Structure

Ainsi, le DW en intégrant différentes bases de données opérationnelles, permet un accès rapide aux informations historisées. De plus, il fournit des outils d'analyse de ces données qui rendent leur utilisation pour les processus* de prise de décision plus facile.

Par conséquent, le DW permet de « résumer » les données et d'uniformiser et rectifier celles incohérentes.

Conçu pour contenir les données en adéquation avec les besoins de l'entreprise et répondre de manière centralisée aux requêtes des utilisateurs, le DW n'est pas régi par une règle unique de stockage ou de modélisation.

En effet, on distingue quatre classes de données organisées selon un axe historique et un axe synthétique.

L'ensemble peut être présenté comme ce qui suit :



Illustration 18 : Répartitions des données dans un Datawarehouse

Les données détaillées

Appelées aussi « données élémentaires », elles proviennent de la base de données opérationnelle et représentent les événements les plus récents. Malgré leurs gros volumes,

elles présentent des avantages évidents. En effet, elles se caractérisent par de la profondeur et un grand niveau de détail. On peut leur appliquer différents axes d'analyses et surtout, elles contiennent une information relative à un moment passé. L'inconvénient majeur est la nécessité d'avoir du matériel performant afin de pouvoir gérer le grand volume qu'elles représentent.

Les données agrégées

Ce sont le résultat de données élémentaires que l'on a agrégé selon les axes d'analyse souhaités. L'agrégation de données correspond, globalement, au groupement (calcul et traitement) d'un lot de données afin d'obtenir un résultat synthétique. De ce fait, les données agrégées représentent un volume moins important que pour les données détaillées.

De plus, elles offrent une grande facilité d'analyse et un accès rapide.

Soulignons cependant que l'agrégation des données revient à les figer selon les axes d'analyses que l'on a prédéfinis. En effet, il est impossible une fois le processus* d'agrégation réaliser de retrouver le niveau de détail ou la profondeur des données de départ.

Par exemple : si l'on a fait une agrégation du chiffre d'affaires d'une entreprise par mois, il sera impossible de le faire par jour par la suite.

Une moyenne, une somme, ou un dénombrement du nombre d'enregistrements sont des exemples d'agrégations de données.

Si le modèle relationnel est privilégié pour le stockage de données élémentaires, ce des modèles dits en Étoile ou en Flocon qui sont privilégiés pour le stockage de données agrégées.

Ces deux modèles sont décrits en ANNEXE 4.

Les métadonnées

Définies comme données décrivant les données, les métadonnées décrivent les règles ou processus* attachés aux données du système. Permettant de faciliter les recherches de données, il s'agit d'une véritable aide décrivant l'information contenue dans le DW.

Ainsi, elles permettent notamment de connaître, tout d'abord, les données entreposées, leur format, leur signification et leurs degrés d'exactitude. Ensuite, elles informent aussi sur les processus* de récupération et d'extraction dans les bases sources. Enfin, elles renseignent les utilisateurs sur la date du dernier chargement du DW et sur l'historique des données sources et celles du DW.

Elles sont destinées aux utilisateurs, aux équipes responsables des processus* de transformations des données, aux équipes de création des données agrégées, aux équipes d'administration de base de données, etc.

Idéalement, ces métadonnées sont conservées dans un référentiel pour le DW.

Les données historisées

Chaque nouvelle insertion dans le DW, fait l'objet d'une insertion. Ainsi chaque nouvelle valeur d'une donnée, aboutit à la création d'une nouvelle occurrence de cette donnée avec une nouvelle valeur et un nouveau marqueur temps.

Grâce à cela, un historique des données est créé.

c) Architectures

Architecture réelle

C'est l'architecture la plus répandue et utilisée pour les systèmes décisionnels. Elle consiste à stocker les données destinées au Datawarehouse séparément du système de production. L'entreprise se dote alors de deux systèmes de stockage volumineux : l'un opérationnel, l'autre décisionnel.

Ce seront, par la suite, des extractions périodiques du premier qui viendront alimenter le second en ayant subi entre temps des traitements de nettoyage et de transformation des données.

L'avantage principal est de disposer de deux environnements différents chacun dédié à un domaine précis et surtout d'avoir des données préparées prêtes à l'emploi pour le décisionnel. Les inconvénients majeurs sont le coût d'une telle architecture en terme de matériel, de maintenance et de personnel pour la gestion, et surtout le manque d'accès en temps réel à des données préparées issues de l'opérationnel (dû aux extractions périodiques).

L'illustration n°17 intitulée « Concentration de l'information de l'entreprise dans un Datawarehouse » est un exemple d'architecture réelle.

Architecture virtuelle

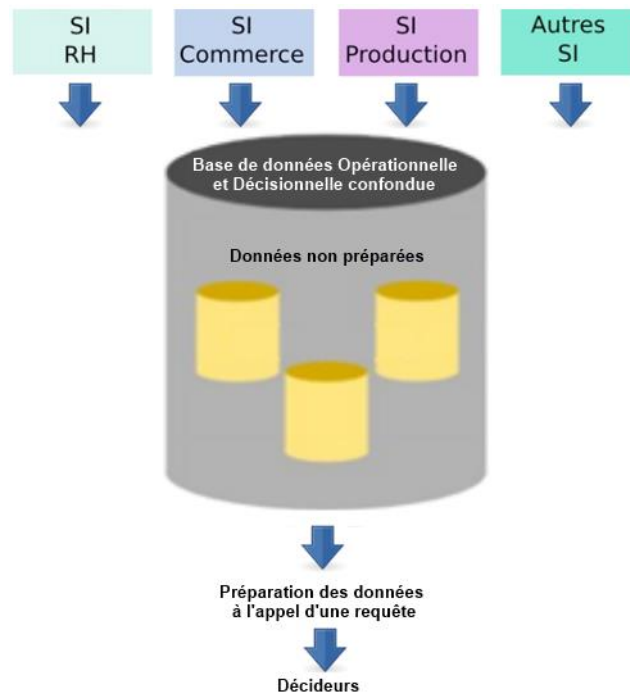


Illustration 19 : Architecture virtuelle d'un Datawarehouse

Cette architecture préconise un seul lieu de stockage des données. Ainsi les données opérationnelles et celles dédiées au décisionnel sont stockées sur le même système. À la différence près que ce sont les mêmes. En effet, ici les données ne sont pas préparées puis stockées. Elles seront préparées lorsque les décideurs en auront besoin (appel à une requête).

L'avantage est donc une économie de coûts et un accès en temps réel à l'ensemble des données mises à jour. L'inconvénient étant le fait que les données ne sont pas préalablement préparées. Le temps de traitement est donc plus long.

Cette architecture est peu répandue chez les professionnels.

Architecture remote

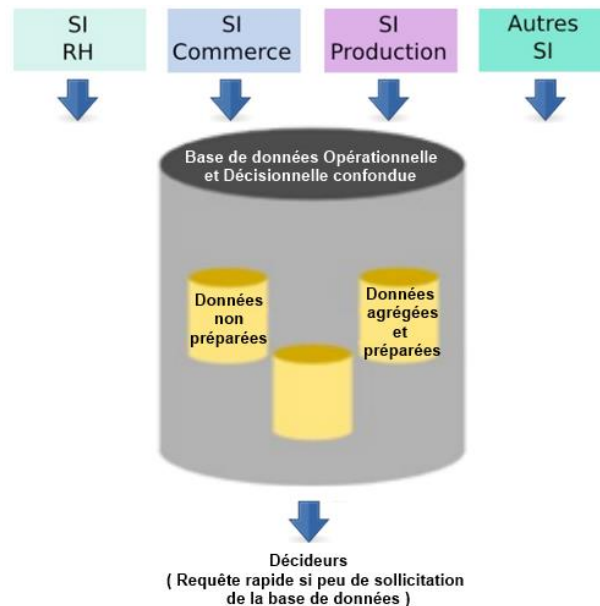


Illustration 20 : Architecture remote d'un Datawarehouse

Combinaison des deux architectures citées précédemment, elle est rarement utilisée dans le monde des entreprises. Son objectif est de faciliter l'accès aux données tout en gardant le niveau de détail dans la base de données opérationnelle. Pour cela, cette architecture préconise l'implémentation physique des niveaux agrégés : les données sont alors préparées, agrégées et stockées dans la base de données opérationnelle.

L'avantage est de n'avoir qu'un seul système de stockage. Cependant, l'inconvénient est que la base de données va vite être remplie vu qu'elle contiendra à la fois les données opérationnelles et celles dédiées au décisionnel. De plus, plus elle sera sollicitée, plus le temps de réponse aux différentes requêtes sera important.

Dans notre solution, nous préconiserons l'architecture réelle. Car d'une part, nous ne pouvons pas écrire dans la base GITHUB et d'autre part, le temps de récupération et de traitement des données serait trop conséquent pour permettre une quelconque exploitation. Stocker les informations issues du site web et des différents autres serveurs JENKINS et REDMINE, après une extraction, nous permettra de réellement observer l'évolution de la base de données et des agrégats. De plus, nous ne risquons aucune interruption de traitement ou de ralentissements dus à une trop grande sollicitation des bases de données.

d) DATAMART

Connu aussi sous le nom de « Magasin de données », un Datamart est une vue partielle d'un DW, mais orienté métier. Il s'agit d'une base de données orientée décisionnelle, mais destinée à un petit nombre d'utilisateurs. Elle se rapporte souvent à un secteur d'activité particulier d'une entreprise où un seul métier y est exercé. Alimentée par le Datawarehouse, cette base permet ainsi de réduire le nombre de traitements sur une base de données opérationnelle tout en fournissant des services spécifiques aux utilisateurs et aux métiers qu'ils ont. Du fait qu'il soit plus petit qu'un DW, un Datamart, permet aussi un accès aux informations plus rapide. En effet, alors que dans l'entrepôt de données, les faits détaillés sont conservés, dans le datamart, un certain niveau d'agrégation a déjà été effectué. On perd donc en détail afin de gagner en temps de traitement.

Si le découpage par métier est souvent privilégié (un Datamart pour le service des ventes, des ressources humaines, etc.) il est aussi possible d'adopter une répartition par sous-ensemble organisationnel (un datamart par succursale).

	Data Warehouse	Data Mart
Cible utilisateur	Toute l'entreprise	Département
Implication du service informatique	Élevée	Faible ou moyen
Base de données d'entreprise	SQL type serveur (relationnel)	SQL milieu de gamme, bases multidimensionnelles
Modèles de données	À l'échelle de l'entreprise	Département
Champ applicatif	Multi sujets, neutres	Quelques sujets, spécifiques
Sources de données	Bases de données opérationnelles	Le data warehouse
Stockage	Base de données	Plusieurs bases distribuées
Taille	Centaine de To et plus	Une à 2 dizaines de GO
Temps de mise en place	9 à 18 mois pour les 3 étapes	6 à 12 mois (installation en plusieurs étapes)

Tableau 6 : Comparatif entre les DataWarehouses et les Datamarts

Dans notre solution, nous n'aurons pas à nous préoccuper de cet aspect. En effet, il n'est intéressant que lors de la phase opérationnelle, soit après la mise en place du Datawarehouse, pour une utilisation quotidienne des données et sous réserve d'avoir plusieurs métiers nécessitant du décisionnel.

Cependant, il est important de connaître cet élément afin de comprendre les enjeux accompagnant la mise en place d'un datawarehouse.

Dans notre cas, nous nous intéresserons directement aux données du Datawarehouse pour une exploitation en datamining.

2. ETL : EXTRACT, TRANSFORM, LOAD

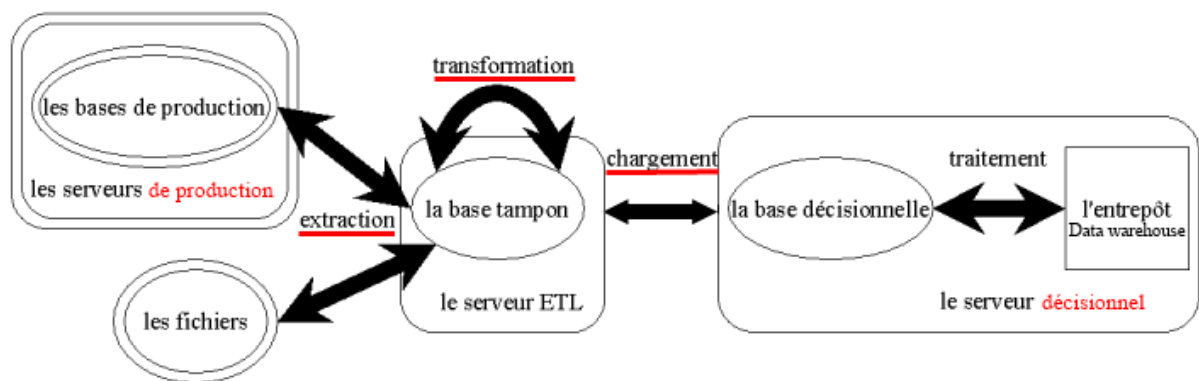


Illustration 21 : Fonctionnement d'un ETL doté d'une base tampon

Intervenant en amont du DW, une solution ETL (pour Extract- Transform- Load) est chargée de la récolte des informations, de leur transformation dans un format adéquat, puis de leur chargement ou intégration dans le Datawarehouse.

D'une part, cette solution nécessite, au niveau technique, un interfaçage de l'outil avec les différentes sources d'informations. En effet, il se peut que les différents services d'une entreprise n'utilisent pas tous les mêmes outils informatiques et donc aient chacun leur propre base. Il faut donc que l'ETL mis en place s'interface correctement avec les bases de chacun afin de récupérer correctement le flux de données. Il doit être capable de gérer les informations de plusieurs supports notamment des documents Word ou Excel.

Il garantit la qualité des données non seulement dans le Datawarehouse, mais aussi dans toute la chaîne de systèmes qu'il alimente.

D'autres parts, au niveau fonctionnel, l'outil a besoin de l'établissement de normes surtout pour la transformation de donnée. Impliquant de nouveaux enjeux, la normalisation concerne les données tant sur le fond que la forme : nom des champs, suppression des doublons, taille, type de champs, etc. Cette étape de définition de la norme à suivre, impacte tout le processus* d'harmonisation et donc le processus* de prise de décision par la suite.

En résumé, le but d'un ETL est de transformer des données élémentaires issues des systèmes de production en informations stockées dans le Datawarehouse. Comme son nom l'indique, cela se fait en 3 étapes : Extraction, Transformation et Chargement pour Load.

a) Extraction

Nommée aussi phase d'identification et d'épuration, elle consiste à définir et identifier les données les plus pertinentes en fonction de chaque source.

Dans tous les cas, il y a toujours un choix à faire : extraire l'ensemble des données en une seule fois ou par lot de manière incrémentale.

La première manière est la méthode la plus rapide, mais cependant aussi la plus risquée, car la moindre erreur peut être fatale au processus*.

La seconde manière, plus lente, permet cependant de limiter les effets d'erreurs ponctuelles. De plus, si elle se fait selon une référence temps (par exemple par mois, ou par semaine) elle entrainera moins de données à contrôler et à traiter, par la suite.

Le choix de la fréquence de l'extraction des données se fait selon plusieurs critères : le temps d'exécution des procédures*, le volume de données à récupérer, la disponibilité, etc. Ce sont autant de facteurs qui peuvent impacter la fréquence.

Ainsi, afin d'éviter que les bases de productions ne soient saturées, la plupart des entreprises planifient cette tâche lors de moments de faible sollicitation des bases opérationnelles. Soit généralement la nuit.

Cependant, ce moment est aussi celui privilégié pour les sauvegardes. Celles-ci aussi importantes que l'extraction elle-même doivent être réalisées sans encombre. Il est donc évident que l'extraction et les transformations de données ne peuvent se faire sur les serveurs opérationnels.

La solution privilégiée est alors la base Tampon.

b) Transformation

Cette étape correspond à la mise au format nécessaire des données extraites. Cela passe par des calculs, fusion, regroupement ou éclatement de données si nécessaire. Cet ensemble d'opérations vise à vérifier et transformer l'information selon des critères et normes préétablis. Tout cela se fait directement dans la base tampon.

La phase d'extraction permet de récupérer les informations depuis les tables de bases opérationnelles des différents SI. Les données sont alors structurées comme en production, constituant ainsi des tables temporaires. Ces dernières, dans la base tampon, vont subir les procédures* de transformation afin de « préparer les données » à être injectées dans la base décisionnelle. L'illustration n°22 présente, globalement, la modification de structure subite par les données, des bases de production vers la base décisionnelle.

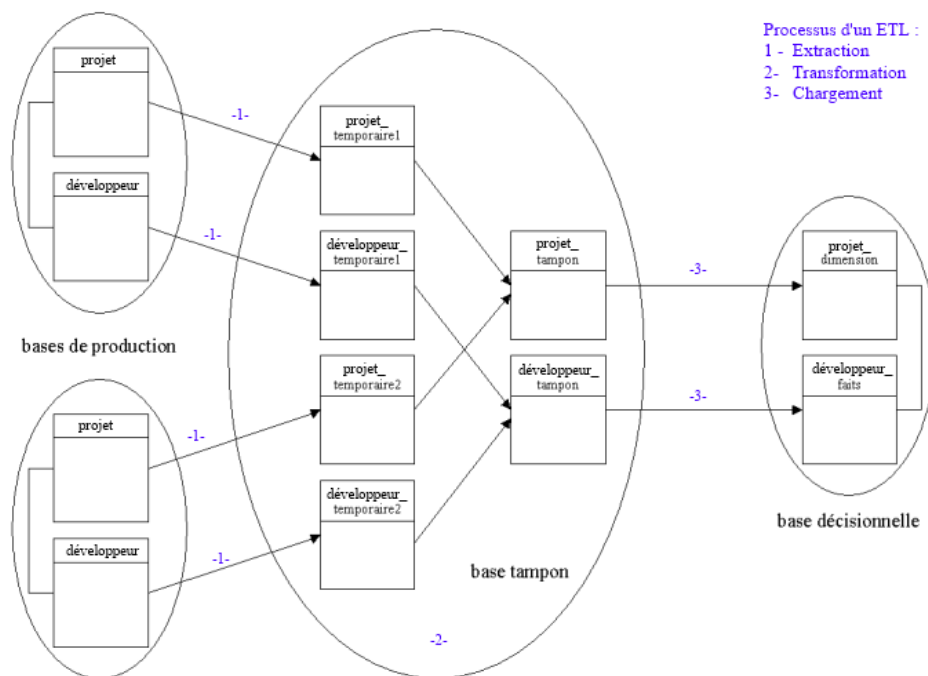


Illustration 22: Format de données durant les phases d'un ETL

C'est une fois dans la base tampon, que les données sont uniformisées. Elles sont réparées, complétées et formatées, si cela est possible.

Dans le cas de codes postaux erronés ou invalides, ils peuvent par exemple être corrigés en utilisant un annuaire des codes postaux.

Un exemple de complétion de données serait de déduire la région où est domicilié un propriétaire à partir du numéro d'immatriculation de son véhicule.

Enfin, les SI d'une même entreprise n'utilisent pas forcément la même horloge, il faut donc traiter les données temporelles (date, heure) afin de les synchroniser dans les tables temporaires, avant leur transfert dans les tables tampon. Le format de date et de chaîne peut aussi être retravaillé afin de correspondre aux normes de la base de données décisionnelle.

Un autre exemple de transformation est l'homogénéisation : selon leur provenance, une même information peut être présente sous des formes différentes. Par exemple, les noms des intervenants sur un projet peuvent arriver sous la forme de deux colonnes (nom et prénom) ou une seule colonne (Nom+ prénom). Les procédures* de traitement veilleront à homogénéiser le format de données afin qu'il soit le plus pratique. Un format prévaudra par rapport à l'autre en fonction de sa cohésion avec l'objectif final et la base de données décisionnelle.

Dédiée seulement à la transformation des données, la base tampon n'est sollicitée que pour cette étape. Elle n'est ni sollicitée par la production, ni par le décisionnel, elle n'existe que lors des traitements de l'ETL.

c) Chargement

C'est l'alimentation du Datawarehouse avec les données préparées. Cette alimentation ne se fait pas en une seule fois. Il convient de la faire en plusieurs temps.

La base de données d'un Datawarehouse est un modèle relationnel. Cela implique qu'il existe un lien entre une ou plusieurs tables de données. Par exemple, le login d'un développeur est présent à la fois dans la table des développeurs, mais aussi celle des demandes. C'est grâce à ce lien que l'on peut savoir qui a traité quelle demande, ou que l'on peut connaître le nom prénom du développeur à partir de son login présent dans l'enregistrement de la demande.

À cause de ces liens et relations dans la base décisionnelle, il convient de charger d'abord les tables qui ne contiennent aucun lien (les intervenants, les projets, etc.), pour ensuite charger celles qui contiennent des liens vers celles déjà chargées (les demandes, les bugs, etc.)

Ensuite, pour chaque table, il faudra charger les nouvelles données, puis les nouvelles versions des données déjà présentes dans l'entrepôt.

En théorie, un ETL n'est pas toujours indispensable en amont d'un Datawarehouse. Si les bases

de données opérationnelle et décisionnelle sont confondues alors l'ETL n'a pas d'utilité (architecture remote et virtuelle). Si elles sont distinctes, mais similaires et simples, alors il n'est pas utile non plus, car l'ETL n'apporte aucune valeur ajoutée dans le cadre de calculs simples comme pour un traitement SQL simple. Son atout majeur est son excellent niveau de performance lors de phases de calculs complexes et cela, peu importe le volume de données. Les prérequis matériels sont, par contre, importants : la mémoire vive* de la machine sur lequel l'ETL fonctionne doit être importante afin de supporter « le poids » des requêtes de transformations et des traitements de la base tampon.

3. MARCHE DU DÉCISIONNEL : DATAWAREHOUSE ET ETL

La Datawarehouse et son ETL, dans le cas de notre infrastructure, constituent notre environnement d'extraction et de stockage de données, que l'on doit interfacer. Élément clé pour la mise en place d'un processus* décisionnel, le datawarehouse est aussi l'élément le plus couteux dans notre infrastructure.

En effet, si les outils de développement sont tous open source* et gratuit, la mise en place du datawarehouse, est couteuse, en terme de temps, d'argent et de ressources humaines et matérielles.

Il faut donc allouer un serveur pour le décisionnel et des personnes compétentes, pour la gestion et la maintenance. Selon la solution choisie, open source* ou propriétaire, les budgets varient beaucoup, car il y a toujours le coup humain qui est important.

En effet, dans un système décisionnel, l'entrepôt est un élément majeur. Sa qualité, son administration et sa maintenance doivent être assurées. Il faut donc allouer des ressources à cela, ce qui engendre des coûts.

Parmi les solutions, il y a celle open source* nommé InfoBright (pour les bases de données Mysql) et celles OLAP, dédié à de l'analyse multidimensionnelle, sous le nom de JEDOX PALO OLAP SERVER et MONDRIAN PENTAHO. Connu et réputé, elles ne font cependant pas le poids face à une solution propriétaire en particulier, celle de l'entreprise TERADATA.

Très présente sur le marché, cette société a su être reconnue plusieurs années de suite, comme le leader incontestable. Elle a même réussi à s'imposer chez des entreprises bien particulières telles que Lidl, la chaîne de hard discount. En effet, ce dernier a récemment investi dans cette solution performante et chère afin de parfaire sa plateforme analytique. Ce fait en dit long sur l'attrait du monde professionnel pour la solution Terradata.

Les prix peuvent varier entre de 240 000 € à 900 000 € en fonction de la puissance serveur délivrée pour ce prestataire. Il s'agit certes d'un investissement conséquent, mais si l'on considère l'importance de la fiabilité des données dans un processus* de décision, il s'avère alors non seulement nécessaire, mais surtout vite amorti au vu de la sécurité que ce processus* apporte alors.

Au niveau des ETLs, il en existe de nombreux sur le marché : propriétaires et gratuits. Cependant, la plupart des entreprises préfèrent créer leurs propres outils, car d'une part, ils seront alors en parfaite adéquation avec leur besoin, et d'autres parts, elles n'auront ni à communiquer leur format de données à un prestataire extérieur ni à déboursier une somme importante.

En effet, ces dernières années le marché de l'ETL propriétaire a subi plusieurs bouleversements. De nombreux rapprochements entre les entreprises leader ont eu lieu dans le secteur, ce qui aboutit à une forte augmentation des tarifs.

Ainsi, IBM propose son offre InfoSphere DATASTAGE (la plus utilisée en France) à partir de 120 000 euros. Informatica, autre entreprise leader, propose une offre nommée Informatica PowerCenter à partir de 180 000 euros. C'est d'ailleurs le leader mondial du marché ETL.

Enfin, Oracle, propose une offre de 3000 euros à 9000 euros par processeurs*. Microsoft a revu l'an passé la gestion de ses licences pour sa solution SQL server 2012, passant d'une licence par processeurs* à une licence par cœur* (un processeur pouvant être composé d'un ou plusieurs cœurs).

Au niveau open source*, les ETLs sont très performants. Contrairement aux datawarehouses, dans le monde de l'ETL les solutions open source* représentent des alternatives très sérieuses à celles payantes. Les leaders dans ce créneau-là sont Pentaho Data Integration, et Talend Open Studio.

Complets, ils permettent une implémentation complète à moindres coûts et surtout bénéficient des contributions de communautés très actives. Le partage des avancées de chacun permet de créer une expérience commune et un échange des solutions et astuces à l'échelle mondiale. Ces outils totalement gratuits, peuvent être déployés sur n'importe quelle machine, peut importer son nombre de processeurs ou le nombre de serveur. Pour les petites et moyennes entreprises, cela représente des avantages indéniables.

Dans notre cas par exemple, nous ne ferons pas appel à un ETL propriétaire pour récupérer le flux de données, mais nous proposerons notre propre solution. Cela nous permettra d'explorer plus en profondeur le fonctionnement d'un ETL et surtout nous permettra de garder notre indépendance vis-à-vis d'un quelconque éditeur de solution.

D. Mise en place de l'ensemble

Dans les deux parties précédentes, nous avons abordé d'une part, les informations et données que nous souhaitons exploiter et d'autre part, les solutions d'extraction et de stockage que nous souhaitons mettre en place.

Dans cette partie, nous allons interfacer ces deux environnements : ainsi nous apporterons une solution à notre problématique de départ.

1. PARTIE 1 : CONSTRUIRE UN DATAWAREHOUSE

Cette étape est l'une des plus cruciales dans la mise en place de notre infrastructure, car, quelle que soit la manière dont elle est menée, elle impacte directement les décisions stratégiques des décideurs.

Rappelons que le Datawarehouse doit permettre de réaliser des requêtes purement consultative, mais néanmoins basée sur des analyses puissantes et pertinentes, et surtout exhaustives. Toutes les données doivent être prises en compte, tout en affichant des temps de réponses faibles.

Ses données orientés sujet, intégrées, historisées et non volatiles lui permette de fournir des informations historiques et transversales de l'entreprise. En regroupant l'ensemble des données de celle-ci, il permet à la fois une centralisation et une meilleure gestion du système informatique décisionnel.

Nous allons ainsi aborder ici, la mise en place d'un Datawarehouse : conception, construction et administration sont les 3 étapes clés.

a) Conception

Cette première étape permet de définir la finalité du Datawarehouse. En général, les analystes se demandent qu'elle est l'activité que l'on souhaite piloter grâce à cela.

Les données à collecter sont recensées et déterminées : quels sont les données à analyser et pourquoi ? Les aspects techniques de la réalisation sont aussi abordés : sous quel format les données doivent-elles être stockées ? Quelle base privilégier ?

Cette phase correspond donc à l'analyse du projet : de ses contraintes et objectifs. Elle sert à éclaircir toutes les zones d'ombre et à s'assurer que rien n'a été oublié au risque d'engendrer un processus* de mise en place infaisable ou de le retarder tout simplement. Les modèles de données et les démarches d'alimentation sont décrits précisément dans l'équivalent d'un cahier des charges*. Ainsi pour bien construire un entrepôt de données dans une entreprise, il faut bien connaître les métiers utilisateurs.

Dans le cadre de notre infrastructure, nous avons déjà décrit les données lorsque l'on a abordé les différents composants de l'environnement de développement. (Tableau n°2 à n°4) Notre but étant de réaliser des techniques de datamining sur des données issues de la gestion de projet informatique nous avons donc proposé de récupérer toutes les données possibles.

Pour l'ensemble des projets, il nous faudra donc :

- L'ensemble de données liées à la gestion de version grâce à GITHUB
- L'ensemble de données liées à la compilation des projets avec JENKINS
- L'ensemble de données liées au processus* personnel de développement de chaque développeur de chaque projet avec ECLIPSE et le plugin HACKYSTAT
- Et enfin, l'ensemble des données et documents liés à la gestion de projet avec REDMINE (documents électroniques, diagramme de gant, planning, demandes et réalisation, etc.)

Le projet est ainsi considéré dans son ensemble, avec tous les intervenants quel que soit le statut hiérarchique : développeur, chef de projet, décideur, etc.

b) Construction

Cette étape correspond au travail technique. C'est la phase d'extraction des données des différentes bases de données de production, qu'elles soient internes ou externes. Elles sont ensuite nettoyées et homogénéisées selon des règles stockées sous format de métadonnées.

Les données sont ensuite injectées dans le Datawarehouse. Soit cela se fait directement, soit périodiquement grâce à des applications d'interface entre les sources de données et le DW, ou grâce à des serveurs de réplication. C'est lors de cette phase que l'ETL intervient s'il y en a besoin.

Comme cité précédemment, pour les besoins de notre architecture, nous préconisons notre propre solution. Élément Important, nous l'aborderons dans la partie suivante qui lui est dédiée.

c) Administration

Une fois le Datawarehouse, mis en place et alimenté, il faut l'administrer.

Comme tout composant informatique d'un système d'information, il doit être configuré et administré afin d'obéir aux règles de confidentialité et sécurité de l'entreprise.

L'administration d'un Datawarehouse se compose de plusieurs tâches dont les buts sont :

- Assurer la qualité et la pérennité des données aux différents applicatifs
- Assurer la maintenance
- La gestion de la configuration
- Les mises à jour du matériel
- L'organisation et l'optimisation du système
- La mise en sécurité

Les deux derniers points sont les plus importants pour administrer un Datawarehouse.

L'optimisation est un métier à part entière, car il y a constamment des problématiques liées aux performances et à la volumétrie. Il faut savoir se poser les bonnes questions au bon moment. Si l'on prend le cas des données agrégées, celle-ci, représentant une synthèse d'information, est sensée faciliter l'accès à l'information pour les utilisateurs. Il faut donc que le choix des données à agréger et des axes soit pertinent.

Quant à la sécurité, en matière de décisionnel, elle doit être encore plus renforcée à cause du stockage massif de données confidentielles.

2. PARTIE 2 : DÉVELOPPER UNE APPLICATION ETL

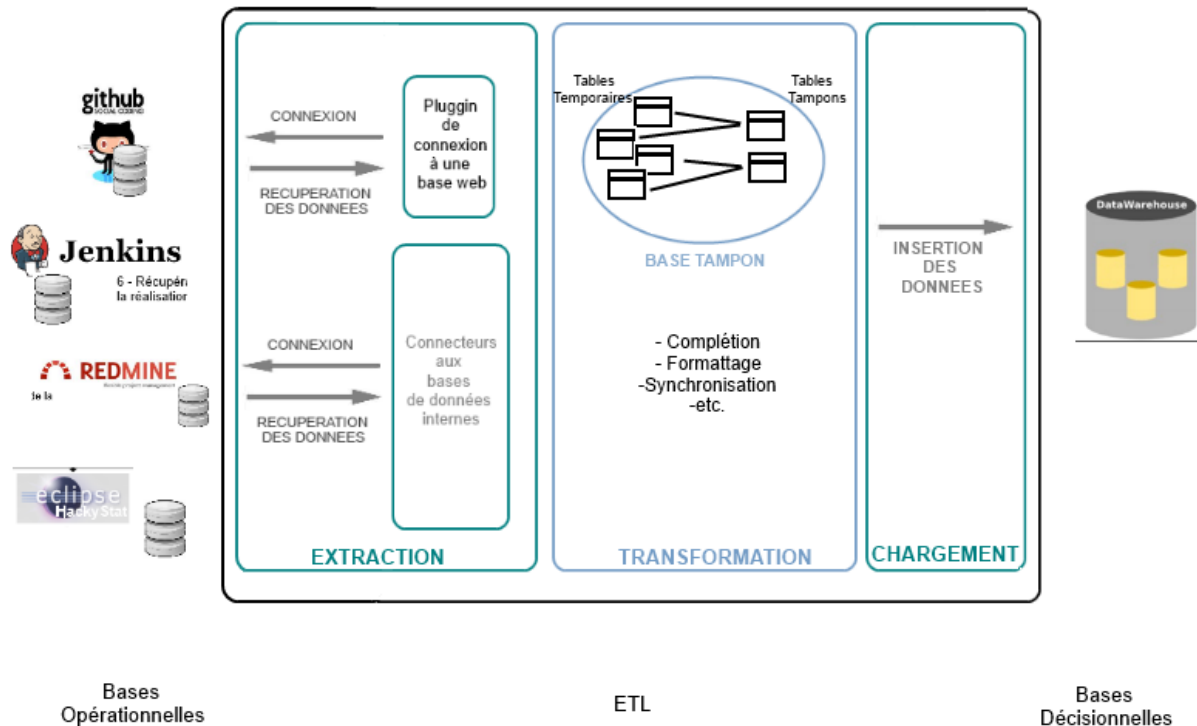


Illustration 23 : Fonctionnement de l'ETL dans l'architecture proposée

Dans le cas de notre solution, nous avons différentes sources d'informations : GITHUB, JENKINS, ECLIPSE et REDMINE. Il va nous falloir récupérer les données dites de production afin de les inclure dans la base de données du Datawarehouse.

Parmi tous nos outils de développement, GITHUB est une exception. Ses données étant sur internet, il nous oblige à avoir une gestion particulière. Il faudra donc récupérer ses données via une interface spécifique avant de les soumettre aux procédures* de traitement.

À part cette exception, toutes les données traitées par l'ETL proviendront de bases de données opérationnelles internes.

Ainsi notre serveur ou machine ETL aura trois interfaces distinctes, chacune reliée à un module différent :

- Une interface web : s'interconnectant avec GITHUB, elle sera reliée au module qui va permettre de récupérer tous les projets GITHUB qui nous sont rattachés.
- Une interface production : permettant l'extraction des données de la base de production, elle sera reliée au module de traitement des données.
- Une interface Datawarehouse : celle-ci sera reliée à l'entrepôt des données dans lequel elle chargera les données.

a) Extraction

Le premier traitement des données se fera lors de l'extraction.

L'idéal dans les bases opérationnelles serait d'avoir pour chaque enregistrement une date de création et une date de mise à jour.

Ainsi l'ETL, pourra extraire les données les plus récentes ou celles qui ont été modifiées depuis la dernière extraction. Non seulement le gain de temps est important, mais cela permet aussi de ne pas saturer le serveur avec des contrôles de données inutiles (déjà extraites).

Grâce à cela, l'extraction peut être incrémentale (par période) au lieu d'être en une seule fois pour l'ensemble des données (volumineuse et lente).

b) Transformation

Lors de la saisie des données dans une base de données opérationnelle, des erreurs peuvent subvenir. Il est donc important que les données soient vérifiées.

Bien que les données proviennent de différentes sources, elles sauront soumises à des contrôles communs qui interviendront directement dans les tables tampons.

Nous mettrons, par exemple, en place des vérifications afin que seuls les enregistrements complets soient récupérés, ainsi que ceux dont on peut compléter les données manquantes.

Les autres enregistrements qui ne peuvent être complétés seront mis de côté.

Imaginons, en effet, un enregistrement qui ne possède ni date de création ni date de mise à jour. Ce dernier devra être exclu du traitement, car il ne peut être situé dans le temps.

De ce fait, lors des calculs, il peut très bien induire un résultat aberrant et ainsi faussé les reporting*.

Par ailleurs, comme nous l'avons vu précédemment, c'est lors de cette étape aussi que les données sont complétées, formatées et synchronisées.

L'ETL devra donc obéir à des normes préfixées. Ses modules de transformations permettront d'avoir des données un format adapté au Datawarehouse, et surtout au besoin des décideurs.

c) Chargement

Les tables tampon finales dans notre base tampon sont le reflet de bases décisionnelles.

Une fois les données traitées et prêtes, elles seront injectées de manière incrémentale dans la base de données.

Il s'agit tout simplement d'insertions de données qui selon la présence de clé et de contraintes d'intégrité, devront se faire dans un ordre bien précis.

C'est donc cette dernière phase de l'ETL qui va permettre d'alimenter notre Datawarehouse.

La réalisation du Datawarehouse avant celle de l'ETL n'est pas un choix dû au hasard. Se focaliser d'abord sur le Datawarehouse permet de se concentrer sur le fonctionnel et l'objectif d'un tel projet. Les choix techniques se font alors ensuite. En effet, une fois que le but du Datawarehouse est fixé, les modèles de données sont ensuite établis.

C'est ensuite en se basant sur cela, que l'ETL saura comment agencer les informations afin de créer les tables tampons correspondantes à celle présente dans le Datawarehouse.

L'ETL représente 70% du temps de la mise en place d'un entrepôt de données, car il est l'élément qui concentre le plus de contraintes techniques, présente à la fois en amont et en aval.

3. PARTIE 3 : INTERCONNECTER LE TOUT

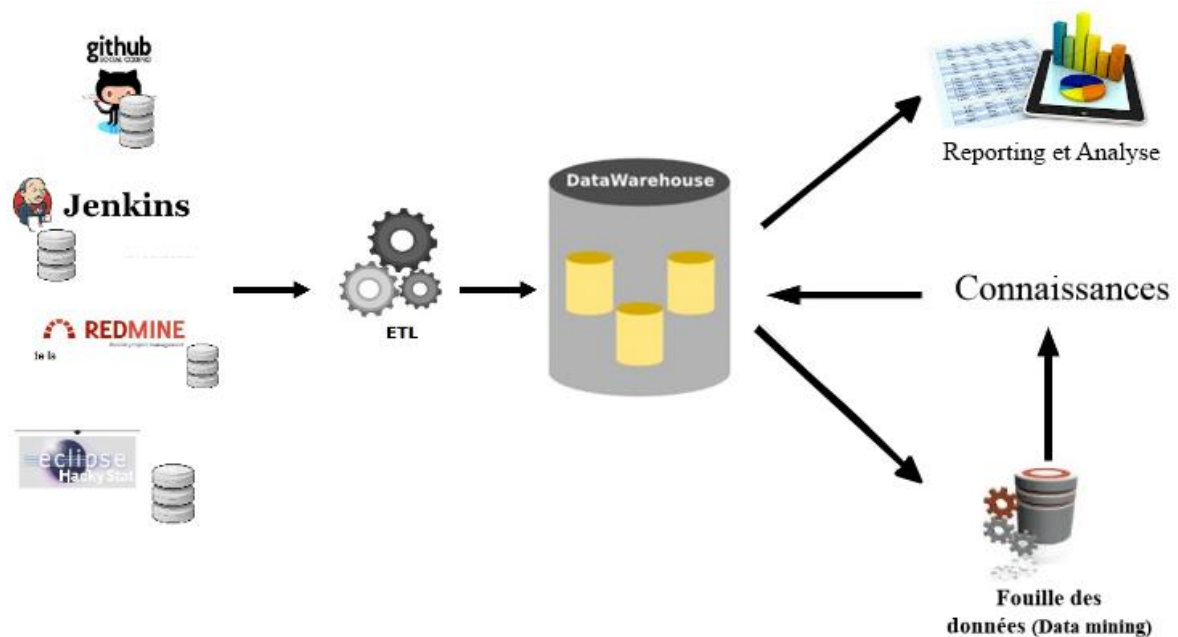


Illustration 24 : Fonctionnement global de notre architecture

Une fois le Datawarehouse et l'ETL mis en place, on peut alors établir les connexions avec les bases de production.

L'ETL va établir la connexion avec les sources opérationnelles. Grâce à des connecteurs, il pourra extraire les données directement.

Celles-ci seront ainsi transformées selon des normes prédéfinies afin d'être injectées dans le Datawarehouse.

C'est le chargement du Datawarehouse. Ce dernier pourra ainsi servir de base décisionnelle pour des reporting*, des analyses, la réalisation de cube de données* et surtout pour le datamining.

En effet, même si notre infrastructure doit permettre la mise en place de techniques de datamining, soulignons aussi qu'elle sert avant tout à mettre en place une solution décisionnelle pour les décideurs. Les outils et techniques standard sont aussi bien possibles que les techniques d'exploration et donc de prédictions.

Grâce aux grands volumes de données stockés dans l'entrepôt, la fouille sera possible.

Les nouvelles corrélations seront ainsi mises en évidence et une fois validée par les analystes, elles pourront alors être intégrées dans l'entrepôt directement.

Le reporting* sera alors alimenté d'une part par des données provenant de la production, mais aussi de l'exploration des données. Il pourra alors être de type explicatif, ou prédictif, ou encore les deux à la fois.

Au niveau financier, pour une entreprise informatique, le bilan est le suivant :

- L'environnement de développement et de gestion des projets est gratuit : totalement open source*, il permet aux développeurs d'intervenir dessus afin de le configurer à leur besoin.
- L'environnement ETL est aussi gratuit : un des avantages de le concevoir soit même pour une société, en plus de la confidentialité préservée et de la parfaite adéquation entre le produit et le besoin
- Selon la base de données souhaitée, le Datawarehouse peut s'avérer très couteux ou non. Si le choix se porte sur une solution d'un éditeur de logiciel, alors choisir TERRADATA est conseillé, car en tant que leader, ses services et experts sont très réputés pour la qualité de service.
- Les outils de datamining open source* conviendront parfaitement à l'infrastructure. Les solutions Rapid Miner et Weka, sont toutes deux performantes. De plus, la caractéristique open source* garantit la possibilité de faire évoluer l'outil au fur et à mesure des besoins (ou innovation) dans le domaine sans engendrer de coût important.

Pour une société informatique, nous préconisons donc des solutions open sources* pour la gestion et l'exploitation des données (développement et datamining). L'entreprise pourra alors bénéficier des avantages d'une communauté d'échange ouverte, diversifiée et compétente en cas de difficulté, et cela à moindre coût.

Pour la solution de stockage, il vaut mieux se doter de la solution la plus optimale, surtout en prévision d'une croissance de l'entreprise, du nombre de projets, de leur taille et donc de celle des équipes de développements.

E. Manipulation, Résultats et Travaux connexes

La mise en place de l'environnement de développement a été pensée de telle sorte qu'il soit accessible à un développeur seul, une TPE, une PME ou une SSII.

Ainsi cette première partie, responsable de la génération de données, n'a donc pas été une difficulté pour ma part.

La difficulté majeure rencontrée a été d'une part le volume de données à générer. Et d'autre part le Datawarehouse. En effet, je ne dispose pas du volume de données nécessaire pour l'exécution de techniques de datamining, ni d'un entrepôt de données assez conséquent pour cela.

Il aurait fallu créer des volumes de données de l'ordre du Gigaoctet ou du Téraoctet, soit plusieurs projets avec plusieurs intervenants, afin de réellement tester l'infrastructure. Ainsi, il aurait été possible de recueillir des résultats de datamining et étudier l'apport de ce domaine à la gestion de projet informatique.

Ainsi à défaut de pouvoir proposer mes propres résultats, je vais présenter ceux d'études similaires ayant inspiré en partie ma réflexion. Nous aurons ainsi une vision des possibilités qu'offre la mise en place de solutions de datamining lors de la gestion de projet.

L'une des études les plus complètes dans le domaine est celle de Adem Karahoca, Professeur et Docteur en Genie Logiciel. Ce dernier dans son livre intitulé *"Data Mining Applications in Engineering and Medicine"*, publié en Aout 2012, se penche sur l'utilisation et l'apport du datamining dans le monde de l'ingénierie et de la médecine.

Dans le domaine de la gestion de projet informatique, il a réalisé une étude sur plus de 4106 projets informatiques de différents pays, différentes entreprises, et pour des secteurs différents (assurance, banques, etc.)

En se basant sur des données liées à la qualité (nombre de défauts, type de bug), l'environnement et l'architecture (langage de programmation*, système d'exploitation*, méthodologie, et développement), il a mis en évidence la compatibilité des techniques d'exploration de données avec les méthodes de gestion de projet, depuis l'initialisation au bilan. L'étude a permis de développer une méthode facilitant et améliorant les processus* d'estimation de temps et des coûts, ainsi que l'estimation de la qualité du produit final.

En limitant les risques potentiels liés aux défauts des logiciels et à la mauvaise gestion, la qualité du produit s'améliore nettement. La maîtrise des estimations et du processus* qualité rendent non seulement les projets plus performants, mais aussi les équipes qui s'en chargent.

Une autre étude, menée à l'université de Technologie de BRISBANE QUEENSLAND en Australie, met en évidence l'apport bénéfique des solutions de datamining dans le cycle de vie d'un projet informatique. Celles-ci permettent de déterminer et prédire les problèmes. Ainsi au lieu d'avoir un processus* de résolution long et coûteux, celui-ci est court et rapide : l'équipe est aidée et guidée dans la recherche de la source des bugs. Les techniques d'exploration des données apportent une assistance aux développeurs, ce qui facilite grandement leur travail et la qualité de celui-ci.

Les données étudiées ont été recueillies à partir dans un environnement de production réel de logiciels informatiques. Elles ont permis de montrer l'amélioration de la qualité et de l'efficacité du processus* de développement logiciel. Le risque diminue grandement au fur et à mesure que les techniques d'estimation et d'analyse des problèmes sont maîtrisés par les équipes.

D'autres études existent dans ce domaine. Très intéressantes, elles soulignent toutes, sous des angles différents, les avantages du datamining - et de ses différentes techniques - appliqué à la gestion de projet : meilleure estimation de temps, résolution de bug plus rapide, diminution du nombre de bug et amélioration notable des équipes et des produits, etc.

Prônant une insertion du datamining dans le processus* de gestion de projet dès les premières étapes, ces études apportent toutes une nouvelle vision du génie logiciel*. Dans un environnement dynamique complexe alliant humains et techniques, le datamining apporte plus de sécurité et donc plus de sérénité pour les décideurs.

F. Et si on ajoutait des informations extérieures ?

Le Datamining est l'exploration des données. Celles-ci sont non exhaustives. On peut donc ajouter de nouvelles données, sans craindre de devoir altérer les procédures*.

Au contraire, l'ajout de sources de données supplémentaires peut être bénéfique et ainsi mettre en évidence de nouvelles corrélations.

Durant notre étude, nous nous sommes focalisés sur la gestion et la réalisation d'un projet informatique. Nous avons donc ciblé les sources de données les plus logiques afin justement d'explorer de nouveaux aspects de la gestion de projets. Nous espérons ainsi permettre aux décideurs d'une entreprise d'avoir plus d'éléments lors de la prise de décisions, à l'initialisation des projets par exemples ou pour la décision de budgets liés aux projets et aux développeurs.

Une suite possible de cette étude serait l'élargissement du périmètre des données à l'entreprise tout entière. Il serait alors intéressant de rajouter les données des clients pour lesquels les projets sont réalisés et les interactions avec eux (Marketing, prospection, relance, etc.). Les formations et le vécu des intervenants durant un projet (chef de projet, assistants, analystes, etc.) pourraient aussi faire partie des paramètres à considérer.

L'ensemble des données citées n'interagit pas directement sur la gestion de projets, mais elles l'impactent peut-être d'une manière ou d'une autre ? Ce genre de relation peut ainsi être mise en évidence grâce au datamining.

De même, à l'heure où les réseaux sociaux sont incontournables dans la vie numérique d'un individu, il serait intéressant d'étudier l'impact de tels outils sur les projets d'un individu ou entreprise et la gestion adoptée. Aujourd'hui sommes-nous plus performants, car nous avons plus d'information à disposition, ou plus fainéants ? La qualité des processus* de développement personnel sont-ils impactés par les réseaux sociaux ? Comment ? Pourquoi ? Quel type de réseaux sociaux ?

En résumé, toutes les informations intervenant dans la vie d'une entreprise, même les plus anodines, peuvent impacter les projets. Ce sont ces impacts qu'il serait intéressant d'étudier par la suite.

G. Généralisation à d'autres domaines

Notre étude et solution d'architecture porte sur la gestion de projet informatique. Cependant, il est possible de l'élargir à d'autres domaines. Notamment celui du Bâtiment que nous avons pris en exemple précédemment.

Il est vrai que dans ce milieu, il n'existe pas d'outils de gestion de version, de compilation ou de développement. Cependant, il en existe d'autres, performant dans leur domaine, générant de grands volumes de données.

Outre les données liées au projet lui-même, issu des esquisses, des études –projets, il existe des données importantes à récolter lors de la réalisation des projets des bâtiments.

Il faut savoir qu'aujourd'hui sur les chantiers, un responsable est chargé chaque jour (si possible) de tenir un journal des événements. Ainsi, que ce soit sur un ordinateur portable, une tablette ou directement sur son Smartphone, le responsable saisit des informations diverses et variées : météo, incident s'il y en a eu, travaux effectués, retard prix, etc.

Cette utilisation nomade permet un suivi des travaux et donc la création d'un historique.

Allier ces données à celle d'un logiciel de gestion de portefeuille d'affaire, permettrait de récolter des données bien précises.

Ce type de logiciel de gestion gère un ensemble de projet ou d'affaire. Il permet de tracer l'ensemble du projet depuis le devis à la facturation avec les reporting* adéquats et tout un ensemble d'options supplémentaires (gestion des collaborateurs, des plannings, des commandes, etc.).

On pourrait ainsi envisager, d'une part, de recueillir les données sur l'ensemble des projets de l'entreprise et d'autre part ceux liés à l'avancement sur les projets (déroulement des chantiers, météo, matériaux, etc.)

Comme proposé précédemment pour la gestion de projet informatique, ici encore, l'inclusion de données supplémentaires est aussi possible (relation clients, documentation ou mail, etc.).

Ces données seraient stockées dans des bases de productions, puis grâce à un ETL elles seraient incluses dans un Datawarehouse. Les techniques de datamining seraient exécutées pour générer de nouvelles connaissances.

Il semblerait ainsi que la généralisation de notre architecture à d'autres domaines soit possible, en théorie. En effet, le seul prérequis nécessaire est d'avoir la possibilité de récolter un volume hétérogène et conséquent de données lié à l'activité ou le métier de l'entreprise.

À partir de l'ETL, le reste de l'architecture sera la même que celle proposée, à un élément prêt : le Datawarehouse est construit - ou choisi - selon les besoins de l'entreprise cible (et à son métier), il faut donc qu'il soit adapté à celle-ci.

Conclusion

« Être toujours plus performant en toujours moins de temps » tel est le défi des hommes d'aujourd'hui ainsi que des entreprises. En partant de ce postulat, nous avons mis en évidence le besoin grandissant des entreprises actuelles pour des outils de plus en plus complets. La concurrence féroce, allié à un contexte technico-économique dynamique et complexe, rend la course à l'innovation et à la performance plus que nécessaire, voire obligatoire.

Aujourd'hui, toute entreprise désirant être compétitive se doit de maîtriser sa production ainsi que sa gestion. À côté de cela, les décideurs doivent pouvoir se projeter et anticiper l'avenir avec un risque minimisé.

Par conséquent, il est nécessaire de leur fournir des solutions informatiques qui offrent à la fois une gestion optimale du présent et une préparation maîtrisée du futur : c'est l'alliance de la gestion de projet et du datamining.

Nous avons ainsi abordé, en premier, la gestion de projet, premier pilier d'une entreprise. Après avoir mis en évidence les 4 étapes clés dont elle se compose - Conception, Planification, Réalisation, Terminaison - nous avons pris pour exemple deux domaines bien distincts : le bâtiment et l'informatique. Bien que différents, ils possèdent deux points en commun : le besoin de contrôle de l'information et l'insuffisance des indicateurs actuels préfixés.

Que ce soit un projet de restauration de bâtiment ou de création d'un logiciel, les décideurs ont besoin de savoir quelles sont les conditions pour lesquelles le projet est rentable, ou non. Il leur faut pour cela, prendre en compte les données centralisées des projets passés et actuels afin de pouvoir prédire les situations futures. Or entre le « turn-over »* des employés, la multiplicité des outils informatiques et des interlocuteurs, il n'est pas toujours évident d'avoir la bonne information au bon moment.

Le second point commun, concernant l'insuffisance des indicateurs actuels, souligne le manque de référence et d'outils pour les décideurs. Les indicateurs se basent sur des données pertinentes, mais passées. Ils fournissent donc un état des lieux passé ou présent, mais pas futur. Jusqu'à présent, les dirigeants ne se basaient que sur leurs expériences et leur intuition pour anticiper l'avenir et ainsi faire des pronostics.

Rien de factuel, que de l'intuitif, ce qui n'est pas sans risque. Enfin, le caractère préfixé des indicateurs peut être problématique, puisqu'il masque les paramètres environnementaux. En se focalisant sur un indicateur en particulier, les responsables peuvent négliger d'autres informations.

C'est donc en raison d'un besoin d'une vision transverse, actualisée et maîtrisée de l'ensemble de l'organisation que les dirigeants cherchent à se doter des solutions les plus pointues.

Parmi celles-ci, le datamining semble avoir beaucoup de succès. Second point de notre mémoire, la fouille ou l'exploration de données semble avoir été la science phare de ces dernières années. Face à l'explosion des volumes de données générées chaque jour, via les multiples supports existants, il a fallu trouver des techniques efficaces et rapides pour l'analyse, la compréhension et la détection d'opportunités.

À la jonction de l'intelligence artificielle, des statistiques et des traitements de base de données, le datamining est le processus* de découverte de connaissances, soit l'existence de corrélations entre des éléments à priori sans rapports. C'est ainsi qu'il génère de nouvelles connaissances. Une fois validées, celles-ci peuvent être ajoutées dans le lot de départ, afin de créer un nouveau lot à analyser.

Il s'agit donc d'un processus* itératif et interactif. Composé de 4 phases - l'identification du domaine d'étude, la préparation des données, l'action sur celles-ci puis l'étude des résultats - chaque itération nécessite une intervention humaine. En effet, les résultats sont soumis à l'appréciation de personnes. À la différence que ces derniers ne sont pas des experts, mais de simple utilisateur possédant une bonne connaissance du métier ou du domaine d'étude. Le datamining démystifie donc l'analyse de données en l'ouvrant à tous ceux du métier.

Se déclinant en plusieurs sous domaines, le datamining s'applique sur tous les formats de données possibles. Text Mining, Image Mining, Multimedia mining représentent respectivement l'exploration de textes, d'images, de contenus multimédia et web. Ce dernier est très utilisé dans les domaines de l'email marketing et du web analytics. Dans le premier cas, l'exploration du web permet de faire du marketing un avantage concurrentiel et un processus* efficace en temps réel. Dans le second cas, elle permet de suivre le ROI des projets digitaux et de consolider des stratégies.

Une fois les notions théoriques de gestion de projet et de datamining éclaircies, nous avons abordé la solution à notre problématique.

Voulue accessible à tous, celle-ci est une architecture qui se compose de deux environnements. Le premier est lié à l'activité de l'entreprise, le second, au stockage et à la gestion des données.

Le premier contexte, point de départ de notre architecture, dépend de l'entreprise et de ses souhaits. Quels que soient les choix et métiers des sociétés, il faut que les données générées soient volumineuses, détaillées et diverses. Avoir des visions différentes d'un même projet à un même instant sera d'autant plus intéressants.

L'ensemble de ces données de production servira de bases pour les techniques de fouilles de données.

La seconde moitié de l'architecture est l'élément clé d'une solution décisionnelle : il s'agit de l'ETL et de l'entrepôt de données. Si l'environnement de développement était libre de toute contrainte hormis celle de générer un grand volume de données, celui du stockage est tout le contraire. En informatique décisionnelle, les requêtes et analyses complexes, portant sur plusieurs Gigaoctets de données, doivent avoir des temps de réponse faibles. Pour cela un Datawarehouse est nécessaire. La structure même de cet outil a été pensée pour permettre l'accès rapide à des informations complexes. Pour l'alimenter, il faut alors avoir recours à un ETL. Responsable de l'extraction des données depuis les bases opérationnelles, il est aussi chargé de la transformation des données, selon des normes prédéfinies, avant de les insérer dans l'entrepôt. Ce second outil est l'intermédiaire entre l'environnement de production et celui décisionnel : il est l'interprète rendant possible notre architecture.

Une fois l'ensemble, mis en place, l'exploration des données est possible.

De nombreuses études, menées à travers le monde, ont prouvé que l'alliance datamining – gestion de projet informatique est un avantage concurrentiel indéniable : en améliorant le processus* de développement personnel de chaque programmeur, l'équipe entière et donc les projets s'amélioreraient nettement. La découverte de bug était anticipée par le processus* de datamining, ce qui facilitait leur correction.

Les processus* d'estimation de temps, de budgets et donc les prédictions se faisaient avec plus de sécurité : la qualité du projet s'en faisait alors ressentir.

En parvenant à détecter les binômes disparates ou ceux qui se correspondent, une réorganisation des équipes est possible afin d'améliorer la cohésion des groupes.

Tous ces avantages, démontrés scientifiquement, soulignent l'apport bénéfique du datamining à la gestion de projet informatique.

Cependant, il nous faut nuancer. Car cette architecture, bien que simple en apparence, possède quelques contraintes : le volume de données et le prix du Datawarehouse.

Bien que de nombreuses entreprises fassent des développements informatiques, elles ne génèrent pas toutes de grands volumes de données nécessitant l'utilisation de datamining. Le risque est maîtrisable et la prédiction à leur échelle est possible sans nécessiter de grands investissements humains et matériels.

De plus, bien que la plupart aient un besoin en décisionnel, toutes ne peuvent pas investir des milliers d'euros pour un Datawarehouse, surtout que ce dernier nécessite une maintenance et donc des ressources qui lui sont dédiées à part entière.

Ainsi, les avantages apportés par le datamining sont indéniables dans le cadre de la gestion de projet, cependant, les investissements nécessaires, doivent être faits en fonction de la taille et des réels besoins de l'entreprise. Longue et coûteuse en ressources, le choix de la mise en place d'une telle architecture doit se faire en toute connaissance de cause, et seulement si le ROI est au rendez-vous.

Le datamining est une méthode parmi tant d'autres permettant de faire du décisionnel. En effet, il existe des méthodes engendrant moins de coûts et d'investissement, mais tout aussi performantes dans leur domaine. L'une d'entre elles est l'OLAP et une autre les statistiques.

La méthode OLAP, pour Online Analytical Processing, correspond au traitement analytique en ligne de données selon différents axes, afin d'obtenir des rapports de synthèses spécifiques. Destinées à fournir une vue transversale de l'entreprise aux dirigeants, les applications de type OLAP se basent sur le calcul de cube* ou hypercube*.

Il s'agit tout simplement d'analyses portant sur trois dimensions ou plus, par exemple : le calcul du chiffre d'affaires en fonction, du temps, de la région et du produit.

Ces applications restent explicatives. Toutefois, elles peuvent être couplées à du datamining afin de prédire et simuler des tendances.

Encore plus accessibles, les statistiques alliées aux requêtes de base de données sont considérées comme outil décisionnel puissant. En effet, certains professionnels du décisionnel estiment que le datamining n'a rien inventé, puisque tout était déjà présent. Préconisant l'utilisation de requêtes d'interrogation de base de données, ils affirment travailler et utiliser cela plutôt que des systèmes coûteux et onéreux des grandes entreprises. Il existe donc un débat parmi les partisans du décisionnel concernant les méthodes et solutions à privilégier. Cependant au fur et à mesure des entretiens et recherches menés, un consensus implicite a émergé, soulignant l'importance pour une entreprise de réaliser des choix selon ses réalités. Se doter de la dernière solution de décisionnelle est certes bien, mais se doter de celle que l'on maîtrise le plus est encore meilleur.

L'explosion du volume de données, a donc apporté ces dernières années, tout un lot de problématique à résoudre et de débat à suivre. Il y a ainsi ceux en rapport avec les nouveautés réelles du datamining, et ceux de la gestion et du stockage que nous avons partiellement abordé dans ce mémoire. Mais ce n'est pas tout, il y a aussi les problématiques liées à l'optimisation des requêtes et traitements, celles de la sécurisation des gros volumes, de la visualisation, etc. C'est l'ensemble de ces préoccupations qui a abouti à la naissance du BIG DATA.

Nouvelle notion, elle ne possède pas de définition propre. Cependant, il existe un consensus parmi les experts décisionnels pour décrire le BIG DATA comme un terme qui recouvre l'ensemble des problématiques liées aux très grands volumes de données, notamment leur stockage et leur analyse.

Avec l'expansion des tablettes, des Smartphones et autres objets connectés, l'amas de données généré est passé de 1,2 zétaoctet (10^{21} octets) en 2010 à 2,8 zétaoctets en 2012. Et bien que les technologies se sont adaptées pour cela, il s'avère, aujourd'hui que ce n'est plus suffisant. Les entreprises ayant besoin de stocker les données dans des entrepôts (parfois

onéreux) ne peuvent tout stocker pour des raisons évidentes de coûts. Cela implique donc qu'elles se doivent de faire un choix et filtrer les données. Ainsi, aujourd'hui les entreprises sont incapables d'exploiter toutes les données à leur disposition. Il faut donc redéfinir des aspects entiers de la gestion des données dans le monde de l'informatique décisionnel, afin de lever cette impossibilité. Cela passe par l'adoption du BIGDATA.

Il existe, pour définir cette notion précisément 3 caractéristiques, les 3 enjeux majeurs, nommées les 3V : Volume, Variété et Vélocité.

Le terme Volume désigne une volumétrie en pleine expansion. Que ce soit les réseaux sociaux, ou les installations scientifiques, on dépasse désormais le téraoctet de données générées en une seule journée. Rien qu'en une minute sur internet, on a 30h de vidéo uploadée, 204 millions d'emails échangés et plus de 100 000 de tweets* publiés.

Le terme Variété illustre le nouveau challenge qui s'offre aux entrepôts de données. Ces derniers contiennent des informations structurées dues aux bases de données relationnelles. Or celles-ci ne représentent que 20% de l'amas de données global d'une entreprise. Il reste 80% de données non structurées à exploiter. Différentes et brutes, tant dans le fond que la forme, ces données nécessitent de nouvelles procédures* d'extraction et de transformation pour une intégration dans les outils traditionnels. Or cela n'est pas prévu dans la majorité des cas.

Enfin le terme Vélocité, mets en exergue le besoin de rapidité des outils traditionnels pour traiter les données en temps réel. Représentant la fréquence à laquelle les données sont générées, capturées et partagées, la vélocité impose aux systèmes décisionnels de faire de la fouille de flux de données soit dû Data Stream Mining.

Ainsi, engendrant de nouvelles pratiques, de nouvelles technologies et de nouvelles préoccupations, le BIG DATA est donc l'ensemble des challenges de la gestion de l'information avec des outils qui n'existaient pas jusqu'à présent. Nouveauté depuis peu dans le paysage informatique, elle tend à remplacer progressivement les solutions standards existantes afin de devenir elle-même le nouveau standard à suivre.

Ainsi, le datamining, jeune « science », va-t-il subir ce sort?

En fait, non. Il est vrai que ces dernières années, le terme « BIG DATA » a pris de l'ampleur au point d'accaparer une grande partie de la scène médiatique informatique, cependant, les

deux concepts se complètent parfaitement : Le BIG DATA recouvre les problématiques liées au stockage et à l'analyse des données, alors que le datamining est le moyen de les traiter, de les analyser pour découvrir des corrélations, des tendances, et donc de valoriser les données en informations pertinentes. Plus le volume de données est grand, plus la technologie de datamining doit être performante. Il est vrai que les besoins d'analyse de données induits par le BIG DATA vont au-delà des notions originelles du datamining. Cependant, elles gardent tout leur sens.

Si c'est dans l'informatique décisionnelle que l'on entend le plus parler de données, ce n'est pas le seul domaine où elles sont importantes. Aujourd'hui, tout est régi par cela. Selon l'information qui en résulte, nos actions sont différentes. Il existe un domaine en particulier où les données ont un impact non négligeable, c'est celui de la médecine et de l'industrie pharmaceutique.

Que ce soit lors de recherches concernant un traitement potentiel, ou le suivi de patient, il s'avère que le datamining est utilisé depuis des années dans ce domaine-là. En effet, la prise de conscience croissante relative à la valeur des connaissances cachées incite les analystes de ces métiers à se doter de solutions adéquates pour expliquer ou prédire des situations complexes. L'exploration de données permet ainsi d'identifier les réelles relations entre les dosages d'un médicament et son effet réel, ou encore de découvrir les typologies* de patients, en passant par l'établissement de diagnostics et de pronostics.

Concrètement, cela signifie que désormais, il existe des machines capables de fournir des consultations médicales lorsque le spécialiste n'est pas disponible.

Dans l'analyse du génome humain, le datamining est utilisé à la fois pour expliquer le fonctionnement de certains gènes, mais aussi pour simuler des situations où un gène viendrait à manquer. C'est grâce à ces études que les chercheurs peuvent établir, avec un degré d'incertitude, la fonction d'un gène.

Aujourd'hui, nous pouvons estimer que le datamining, soit l'exploitation des données en général en est à ses prémices. L'explosion des volumes de ces dernières années, allant continuer sur sa lancée, va donc générer de nouvelles problématiques.

Les technologies existantes vont donc devoir évoluer aussi afin de s'adapter aux nouveautés, au risque de devenir obsolète. Une caractéristique du datamining est son alliance avec des méthodes liées à l'intelligence artificielle. Cette dernière est une science très dynamique et innovante qui s'est insérée petit à petit dans notre quotidien, grâce notamment aux assistants mobiles.

Science dont le but est de simuler, artificiellement, le raisonnement humain, elle est aussi à ses prémisses. De nombreux laboratoires et entreprises visent à essayer, chaque jour, de trouver des méthodes reproduisant les réflexes et la compréhension humaines. Alors certes, ce domaine en est à ces prémisses aussi, mais l'évolution des machines alliée à celle des méthodes d'analyses et de compréhension des hommes assure à cette science, et donc au datamining, un avenir prometteur.

Ainsi s'il est possible à l'avenir de créer des machines qui comprennent, apprennent et analysent beaucoup plus vite et tel des humains, alors, il est sûr que quasiment rien n'a été fait, et que tout reste à faire, à l'avenir, dans le Datamining.

Glossaire

ABSTRACTION : Identifie et regroupe des caractéristiques et traitements communs applicables à des entités ou concepts variés ; une représentation abstraite commune de tels objets permet d'en simplifier la manipulation

ALGORITHME DE TYPOLOGIE : Algorithme permettant de découper un échantillon en plusieurs sous ensemble aussi différent que possible les uns des autres, et avec des individus aussi semblables uns des autres au sein de chaque sous-ensemble. Ce découpage se fait selon plusieurs axes (âge, poids, sexe, etc.) et non pas un seul comme pour les algorithmes de segmentation

API (Application Programming interface) : Interface de programmation servant de façade clairement délimitée par laquelle un logiciel offre des services à d'autres logiciels. Ces derniers accèdent à des fonctionnalités du logiciel partageant, sans accéder les détails de la mise en œuvre.

ASYNCHRONE : Terme désignant un type d'échange de données entre deux machines où les données échangées sont émises et analysées selon une référence de temps différente et un rythme variable.

Son contraire est synchrone.

BASE DE DONNÉES PARALLÈLES : En informatique, le parallélisme consiste à implémenter des architectures permettant de traiter des informations de manière simultanée, ainsi que les algorithmes spécialisés pour celles-ci. Ces techniques ont pour but de réaliser le plus grand nombre d'opérations en un temps le plus petit possible.

Ainsi, des bases de données parallèles sont des bases qui contiennent une information dupliquée. Lorsqu'une requête complexe est lancée, elle est exécutée sur l'ensemble des bases de données. Le temps de réponse est ainsi plus petit puisque le travail à effectuer est divisé.

BINERISATION : Transformation d'une valeur en une suite binaire (01001)

CAHIER DES CHARGES : Document contenant la liste des besoins, des exigences et des contraintes qu'il faut respecter lors de la réalisation d'un projet. Lors de la gestion de projet, il est la référence. S'il est signé par le prestataire et le client, il possède alors une valeur contractuelle.

CODE : Texte représentant des instructions qui doivent être exécutées par la machine. Il est écrit dans un langage de programmation donné, permettant ainsi une meilleure compréhension par des humains. Il est la base des logiciels et applications.

Coûts ENGAGÉS : Correspondent aux coûts engagés par l'entreprise afin de réaliser un projet. Ils prennent en compte l'ensemble des investissements relatif au projet (recrutement, nouveau matériel, etc.)

CRM (Customer Relationship Management) : Signifiant Gestion de la relation client, l'acronyme désigne l'ensemble des outils et techniques destinés à capter, traiter, analyser les informations relatives aux clients et aux prospects dans le but de les fidéliser.

CUBE DE DONNÉES : Nom donné à une analyse lorsque celle-ci porte sur 3 axes ou plus.

Un tableau, par exemple, ne comporte traditionnellement, que 2 axes. Si un troisième axe est ajouté, alors on obtient la forme d'un cube.

DÉBOGUER : Action pour un développeur d'analyser un bug (dysfonctionnement informatique d'un logiciel) afin de comprendre son origine, ses conditions d'apparition et ensuite de le corriger

DIAGRAMME DE GANTT : outil utilisé en ordonnancement et en gestion de projet (souvent en complément d'un modèle de PERT*). Il permet de visualiser dans le temps les diverses

tâches composant un projet sous la forme d'un graphe connexe, valué et orienté. Il est souvent l'outil de référence afin de juger l'avancement du projet.

FICHER LOGS : Terme générique pour désigner les fichiers dont l'extension est « .LOG ». Ce sont souvent des fichiers de texte qui servent à enregistrer une succession d'événements survenue dans une application ou un ensemble d'applications.

FAQ (Frequently Asked Questions) : Présente sur les listes web, il s'agit d'une liste récapitulant les questions les plus fréquemment posées ainsi que leur réponse. Son but est donc de fournir aux internautes une réponse rapide, pour les questions les plus récurrentes et ainsi éviter, par la même occasion la saturation des services en charge de la relation client.

FONCTION (DEVELOPPEMENT INFORMATIQUE) : Portion de code représentant un sous-programme qui effectue une tâche ou un calcul relativement indépendant du reste du programme et qui renvoie une valeur.

FRAMEWORK : Ensemble d'outils et de composants logiciels organisés formant un squelette de programme et dont le but est d'aider les programmeurs dans leur travail.

GÉNIE LOGICIEL : Science de génie industriel qui étudie les méthodes de travail et les bonnes pratiques des ingénieurs en développement logiciel. Le génie logiciel s'intéresse en particulier aux procédures systématiques qui permettent d'arriver à ce que des logiciels de grande taille correspondent aux attentes du client, soient fiables, aient un coût d'entretien réduit et de bonnes performances tout en respectant les délais et les coûts de construction.

HYPERCUBE : Il s'agit d'un n -cube*, n désignant le nombre d'axes d'analyses souhaités.

INTERFACE GRAPHIQUE : Dispositif de dialogue homme-machine dans lequel les objets à manipuler sont dessinés sous forme de pictogrammes à l'écran. Exemple : l'écran d'ordinateur est une interface graphique.

Langage de programmation :

LICENCE PAR CŒUR : Système de licence utilisé lorsqu'une application tourne sur plusieurs cœurs d'un ou plusieurs processeurs. Plus le nombre de ces derniers sera croissant, plus l'application sera rapide et performante. Il faut donc acheter un nombre de licences suffisant pour avoir des temps de réponse adaptés

LICENCE PAR PROCESSEUR : système de licence utilisé lorsqu'une application tourne sur plusieurs processeurs. Plus le nombre de ces derniers sera croissant, plus l'application sera rapide et performante. Il faut donc acheter un nombre de licences suffisant pour avoir des temps de réponse adaptés, selon les traitements réalisés.

LIVRABLE : Produit destiné à la livraison. En informatique, il s'agit souvent d'une version d'un logiciel, accompagné de la documentation et autre élément nécessaire à la bonne utilisation du produit.

MÉMOIRE VIVE : Appelé RAM aussi, est la mémoire informatique dans laquelle un ordinateur place les données leur de leur traitement. Elle se caractérise par sa rapidité d'accès - importante pour fournir rapidement les données au processeur - et sa volatilité. Cette dernière caractéristique implique que toutes les données sont perdues dès lors que la machine est éteinte ou cesse d'être alimentée.

MODELE DE PERT (Program ou Project Evaluation and Review Technique) : Modèle utilisé en gestion de projet pour représenter et analyser de manière logique les tâches à réaliser dans un projet. Il se présente sous la forme d'un graphe de dépendances composé d'unité (carré) qui désigne les tâches. À chacune d'entre elles sont associées une date de début et une date de fin. Le modèle ou diagramme de PERT permet de déterminer le chemin critique* qui conditionne la durée minimale du projet.

Le but d'un tel modèle est de trouver la meilleure organisation possible pour qu'un projet soit terminé dans les meilleurs délais, et d'identifier les tâches critiques, c'est-à-dire les tâches qui ne doivent souffrir d'aucun retard sous peine de retarder l'ensemble du projet.

MODÈLE DU CHEMIN CRITIQUE : utilisé en gestion de projet, il désigne la liste ordonnée de tâches (ou opérations) à effectuer afin de réaliser le projet dans les délais. Les tâches faisant partie du chemin critique sont les tâches élémentaires à maîtriser afin de ne pas entraîner de retard sur le projet.

ODT : Extension des fichiers textes créés avec la suite bureautique OpenOffice.org

OPEN SOURCE : Désignation s'appliquant aux logiciels dont la licence respecte des critères établis par l'Open Source Initiative (organisation dédiée à la promotion des logiciels open source). Cela implique : une possibilité de libre redistribution, d'accès au code source et de créer des travaux dérivés.

PARAMÉTRAGE : Le paramétrage d'un produit informatique permet d'adapter ce dernier au mieux aux besoins d'un client donné. Cela implique de configurer l'application ou l'outil en question afin qu'il soit en parfaite adéquation avec l'environnement technique et l'utilisation que l'on souhaite en faire.

PAREMENT : En construction, surtout en maçonnerie, il s'agit de la face visible de la construction. Exemple : pavés, pierre, briques, etc.

PROCÉDURE (DEVELOPPEMENT INFORMATIQUE) : Sous-programme contenant du code et réalisant une action. Contrairement à la fonction, elle ne renvoie pas de valeur.

PROCESSUS (INFORMATIQUE) : Programme ou application en cours d'exécution par un ordinateur.

PROCESSUS (GESTION DE PROJET) : Selon la norme ISO 9001 pour la gestion de la qualité, il s'agit d'un système organisé d'activités qui utilise des ressources (personnel, équipement, matériels et machines, matière première et informations) pour transformer des éléments entrants en éléments de sortie dont le résultat final attendu est un produit ou un service. Le processus à un propriétaire qui est garant de la bonne fin et du bon fonctionnement de celui-ci.

PVC (POLYCHLORURE DE VINYLE) : Polymère thermoplastique de grande consommation. Le PVC est la seule matière plastique d'usage courant utiliser dans l'industrie des vêtements et des tapisseries et autres produits commerciaux.

QUALITÉ ABSOLUE : désigné aussi de qualité totale, il s'agit d'une démarche de la gestion de la qualité dont l'objectif est l'obtention d'une très large mobilisation et implication de toute l'entreprise pour parvenir à la qualité parfaite en réduisant au maximum les gaspillages et en améliorant en permanence les éléments de sortie.

REDISTRIBUABLE(LOGICIEL) : Cela désigne le fait de pouvoir librement installer autant de fois que l'on souhaite sur autant de machines que l'on souhaite un logiciel.

REPORTING : Désigne l'ensemble des présentations et bilans analytiques d'une entreprise pour une ou l'ensemble de ses activités. En plus d'être un moyen de faire un point régulier sur la stratégie de l'entreprise, c'est aussi un outil de communication auprès des actionnaires et du public.

Outil décisionnel, il permet de comparer la maturité de différentes entreprises ou de différente entité d'une même entreprise au regard d'indicateur de développement soutenable (environnementaux, sociaux, éthique, santé ou économique)

RÉSEAU ETHERNET : Protocol de réseau local permettant la communication entre machines informatiques reliées par un câble.

SCORING : Technique de hiérarchisation des données qui dans le cadre d'une campagne de marketing direct permet d'évaluer par une note ou un score la probabilité qu'un individu réponde à une sollicitation ou appartienne à la cible recherchée.

Le score est obtenu à partir des données quantitatives et qualitatives disponibles sur l'individu (données socio démo, comportement d'achat, réponses précédentes, ...) auxquelles est appliqué un modèle de scoring.

Le scoring permet d'optimiser les résultats de campagnes en concentrant les envois ou contacts sur les individus ayant la plus forte probabilité de réponse ou appétence pour le produit ou l'offre.

C'est aussi une technique utilisée en intelligence artificielle.

SYSTÈME D'EXPLOITATION : ensemble de programmes qui dirige l'utilisation des capacités d'un ordinateur. C'est le premier logiciel lancé au démarrage de l'ordinateur et dont la principale mission est d'assurer la liaison entre les ressources matérielles, l'utilisateur et les applications. Exemple : Windows Vista, Windows 7, Windows 8, Ubuntu, Mac OS, etc.

TABLEAU DE BORD : Outil d'évaluation, de l'organisation d'une entreprise ou d'une institution, constitué de plusieurs indicateurs. Dédié au décisionnel, cet outil permet d'évaluer une situation ou un élément (chiffre d'affaire par exemple) selon des critères définis sur une période donnée.

TRIANGLE FOU DU CHEF DE PROJET : Expression résumant les trois contraintes à respecter pour un chef de projet lors de la gestion de projet informatique : les délais, les coûts et la qualité du produit. Ce sont les trois axes à constamment surveiller afin de mener le projet à bien.

TRYPTIQUE : Expression désignant les 3 axes importants pour le chef de projet : les coûts, les délais, et la qualité.

TURN OVER : Terme désignant le renouvellement de l'emploi ou des employés. Une entreprise ayant un Turn over important implique qu'elle possède un rythme de renouvellement des effectifs important.

TWEETS : Il s'agit d'un message écrit par une personne sur le réseau social Twitter.

TPOLOGIES (DES PATIENTS) : La typologie démarche méthodique consistant à définir ou étudier un ensemble de types, afin de faciliter l'analyse, la classification et l'étude de réalités complexes. Ainsi la typologie des patients est l'étude des classes de patients selon des critères prédéfinis

URSSAF (Unions de recouvrement des cotisations de sécurité sociale et d'allocations familiales) : Organismes privés délégataires d'un service public. Leur principale mission est la collecte des cotisations salariales et patronales destinées à financer le régime général de la Sécurité sociale, ainsi que d'autres organismes ou institutions.

VEILLE INFORMATIONNELLE : Processus de suivi d'une entreprise de l'actualité et de toute information susceptible d'influencer ses activités ou son devenir. Cette veille se produit vis-à-vis de l'environnement extérieur (concurrent, nouveauté, législation, etc.), mais aussi de l'environnement intérieur (syndicat, conflit, etc.).

WIKI : Désigne un site web rédigé par plusieurs personnes et dont le but est d'expliquer et définir un ensemble d'éléments (fonctionnalité). Les pages du site peuvent être modifiées rapidement et librement par n'importe quel internaute souhaitant ajouter sa contribution. Wikipédia en est l'exemple le plus populaire.

Dans la production de logiciel, ce terme est souvent repris pour désigner la documentation qui accompagne un logiciel afin d'en expliciter le fonctionnement et les fonctionnalités.

Liste des tableaux, schémas et illustrations

A. Liste des illustrations

Illustration 1 : Opération d'achat de cartouche d'imprimante.....	12
Illustration 2 : Etapes de gestion de projet.....	15
Illustration 3 :les Phases d'un projet	21
Illustration 4: Datamining à la jonction de l'IA, les statistiques et les bases de données	40
Illustration 5 : le Datamining, un processus itératif et interactif	41
Illustration 6 : Arbre de décision.....	46
Illustration 7 : Fonctionnement du Datamining	47
Illustration 8 : Domaines d'application du Datamining	54
Illustration 9 : Environnement de développement	66
Illustration 10 : Fonctionnement de GIT et GITHUB.....	66
Illustration 11 : Données relatives à un commit.....	68
Illustration 12 : Données relatives aux fichiers à chaque commit.....	68
Illustration 13: Vue globale des builds du projet CineTeam	71
Illustration 14 : Données relative à un build d'un projet.....	71
Illustration 15 : Logo de l'IDE Eclipse	75
Illustration 16 : le PSP au centre de l'amélioration de l'entreprise	77
Illustration 17: Concentration de l'information de l'entreprise dans un Datawarehouse	84
Illustration 18 : Répartitions des données dans un Datawarehouse.....	85
Illustration 19 : Architecture virtuelle d'un Datawarehouse	88
Illustration 20 : Architecture remote d'un Datawarehouse	89
Illustration 21 : Fonctionnement d'un ETL doté d'une base tampon.....	91
Illustration 22: Format de données durant les phases d'un ETL	93
Illustration 23 : Fonctionnement de l'ETL dans l'architecture proposée.....	100
Illustration 24 : Fonctionnement global de notre architecture	103

B. Liste des tableaux

Tableau 1: Comparatif d'outil de gestion de projet	33
Tableau 2: Comparatif entre les données d'un projet et celles d'un commit sur GITHUB.....	69
Tableau 3 : Comparatif entre les données d'un projet et celles d'un build sur JENKINS	71
Tableau 4 : Comparatif entre les données d'un projet et celles d'un build sur JENKINS	73
Tableau 5 : Vision apportée par les composants de l'environnement de développement.....	81
Tableau 6 : Comparatif entre les DataWarehouses et les Datamarts.....	90

Table des matières

Sommaire.....	1
Introduction	3
I. Les Prérequis en gestion de projet et limites	12
A. Prérequis en gestion de projet	12
1. Rappel de la définition.....	12
2. Buts de la gestion de projet : Importance et enjeux	13
3. Les acteurs principaux	14
4. Fonctionnement : Données et indicateurs	15
a) CONCEPTION.....	15
b) PLANIFICATION et RÉALISATION	16
c) TERMINAISON.....	16
d) BILAN	17
5. INDICATEURS DE SUIVI DE PROJET : KPI ou ICP	17
a) DÉFINITION	17
b) CHOIX et ENJEUX DES KPI	18
c) EXEMPLE : le ROI, Return On Investment ou Retour Sur Investissement RSI.....	19
d) EXEMPLE : KPI de COÛT.....	19
e) EXEMPLE : KPI de DELAI.....	20
f) EXEMPLE : KPI de RESSOURCES.....	20
B. Exemple des gestions de projets	21
1. Gestion de projet informatique	21
a) FONCTIONNEMENT	21
b) PROBLÉMATIQUE PROPRE À LA GPI	25
c) INDICATEURS.....	25

2.	Gestion de projet BTP	26
a)	FONCTIONNEMENT	26
b)	PROBLÉMATIQUE PROPRE À LA GPB	29
c)	INDICATEURS.....	30
3.	GP et KPI : Identification des forces potentielles.....	31
C.	Les outils informatiques.....	32
1.	OPEN SOURCE : REDMINE	33
2.	PROPRIÉTAIRE : MS-PROJET	34
D.	Les Limites : causes d'échec.....	34
1.	CAUSE ÉCHEC AU NIVEAU HUMAIN ET ENTREPRISE	35
2.	OUTIL INFORMATIQUE NON ADAPTE : EXPLICATIF ET NON PAS PRÉDICTIF.....	36
3.	INDICATEURS.....	36
II.	Les éléments de solution : DATAMINING.....	39
A.	Le Datamining : Théorie et Origine.....	39
B.	Le Datamining : Notions de base nécessaire à la compréhension	42
1.	LES STATISTIQUES.....	42
a)	Indicateur de tendance centrale.....	42
b)	Indicateur de dispersion	43
2.	LES SCHÉMAS D'INFÉRENCE	44
a)	L'abduction.....	44
b)	Déduction	45
c)	Induction.....	45
3.	ARBRES DE DÉCISION.....	46
4.	RÉSEAUX DE NEURONES	46
C.	Le Datamining : Étape de fonctionnement.....	47

1.	IDENTIFICATION DU DOMAINE D'ÉTUDE.....	47
2.	PRÉPARATION DES DONNÉES.....	48
3.	PHASE DE FOUILLE DE DONNÉES : ACTION SUR LA BASE DE DONNÉES.....	49
a)	Les méthodes de visualisation et de description	49
	Description.....	49
	Optimisation	49
b)	Les méthodes de classification et de structuration	49
	Classification	49
	Groupement par similitudes.....	50
	Segmentation (ou clusterisation)	50
c)	Les méthodes d'explication et de prédiction.....	50
	Estimation	50
	Prédiction.....	51
	Règles d'association ou « analyse du panier de la ménagère » :	51
d)	Les méthodes d'apprentissage	52
	Apprentissage supervisé.....	52
	Apprentissage non supervisé.....	52
4.	ÉVALUATION DES ACTIONS	53
D.	Le Datamining : Domaines d'applications.....	54
1.	LE TEXT MINING.....	54
2.	L'IMAGE MINING.....	55
3.	LE MULTIMEDIA MINING	56
4.	LE WEB MINING	56
E.	Les outils informatiques existants : étude et comparatif	57
1.	OPEN SOURCE* : RAPIDMINER.....	57

2.	PROPRIÉTAIRE : IBM SPSS	57
3.	LES TECHNOLOGIES COMMUNES	58
F.	Le Datamining : Exemple d'utilisation dans un projet	59
1.	E-MAIL MARKETING	59
2.	WEB ANALYTIC	61
III.	Mise en place de l'architecture informatique : Interconnexion des outils de GP et de Datamining.....	64
A.	Interconnexion d'outils de gestion de projet et de datamining	64
B.	Informations et données	65
1.	GITHUB : GESTION DES VERSIONS	66
2.	JENKINS : INTÉGRATION CONTINUE	69
a)	Intégration continue (IC).....	69
b)	Fonctions de jenkins	70
3.	REDMINE : GESTION DES PROJETS.....	72
4.	IDE, INTEGRATED DEVELOPMENT ENVIRONMENT : LES OUTILS DE DÉVELOPPEMENTS.....	73
a)	Définition	74
b)	ECLIPSE, un exemple d'IDE	75
c)	PSP : Processus* de développement personnel ou Personnel Software Process	76
d)	Les données : LOGGING et plugging Hackystach.....	77
e)	LOGGING	78
f)	PLUGING : HACKYSTAT	79
C.	Solutions d'extraction et de stockage des données	81
1.	DATAWAREHOUSE.....	81

a)	Définition	82
b)	Structure	85
c)	Architectures.....	87
d)	DATAMART	90
2.	ETL : EXTRACT, TRANSFORM, LOAD	91
a)	Extraction	92
b)	Transformation	93
c)	Chargement	94
3.	MARCHE DU DÉCISIONNEL : DATAWAREHOUSE ET ETL	95
D.	Mise en place de l'ensemble	97
1.	PARTIE 1 : CONSTRUIRE UN DATAWAREHOUSE	97
a)	Conception.....	98
b)	Construction	99
c)	Administration	99
2.	PARTIE 2 : DÉVELOPPER UNE APPLICATION ETL	100
a)	Extraction.....	101
b)	Transformation	101
c)	Chargement	102
3.	PARTIE 3 : INTERCONNECTER LE TOUT	103
E.	Manipulation, Résultats et Travaux connexes	105
F.	Et si on ajoutait des informations extérieures ?	106
G.	Généralisation à d'autres domaines	107
	Conclusion	110
	Glossaire	118
	Liste des tableaux, schémas et illustrations	126

A. Liste des illustrations	126
B. Liste des tableaux	127
Table des matières	128
Bibliographie.....	134
A. DATAMINING	134
B. Élément Techniques lié à l’environnement de développement	136
C. Gestion de projet informatique et Bâtiment	137
Annexes.....	138
ANNEXE 1 : Synthèse en Anglais.....	139
ANNEXE 2 : Liste d’indicateurs de suivi de projet.....	144
ANNEXE 3 : REDMINE et les données	145
ANNEXE 4 : Les modèles de données	148

Bibliographie

A. DATAMINING

Adem Karahoca, « Chapter 3: Data mining Applied to the improvement of Project Management », Data Mining Applications in Engineering and Medicine, Août 2012

Les applications du datamining à l'ingénierie et à la médecine ont pour but d'aider les analystes dans leur travail quotidien. Ce livre aborde des études de cas visant à démontrer l'apport de ces techniques dans les domaines de gestion de projet et de la médecine.

Le chapitre est disponible en ligne et gratuitement, à l'adresse suivante :

<http://www.intechopen.com/books/data-mining-applications-in-engineering-and-medicine/data-mining-applied-to-the-improvement-of-project-management>

Alain Rakotomamonjy, Gilles Gasso, « Introduction au DataMining »

Il s'agit d'une présentation de deux supposés étudiants à l'INSA Rouen (laboratoire PSI) ayant pour but de faire une introduction à la fouille de données, et aux techniques liés à l'apprentissage.

La présentation est accessible en ligne à l'adresse suivante :

https://moodle.insa-rouen.fr/pluginfile.php/1336/mod_resource/content/0/Parties_1_et_3_DM/IntroDM.pdf

Antoine crochet Darnais, «Le Web Analytics combiné au datamining permet de suivre le ROI des projets digitaux », journaldunet.com, Février 2012.

L'article recueille les réponses de Didier RICHAUDEAU, de l'entreprise EQUANCY. Il traite notamment du datamining et ses applications dans le domaine du WEB.

L'article est disponible sur le lien suivant :

<http://www.journaldunet.com/solutions/analytics/didier-richaudeau-vers-une-convergence-entre-web-analytics-et-data-mining.shtml>

Aude Demoulin, « L'Email-Marketing & le datamining ou comment développer un réel avantage concurrentiel » journaldunet.com, Juin 2013

La journaliste traite de l'utilisation massive du datamining dans les processus de marketing afin de toujours mieux cibler les clients.

L'article se trouve sur le lien suivant :

<http://www.journaldunet.com/ebusiness/expert/54372/l-e-mail-marketing---le-data-mining-ou-comment-developper-un-reel-avantage-concurrentiel.shtml>

Georges El Helou, Charbel Abou Khalil « Datamining : Techniques d'extraction des connaissances », publié par le Laboratoire de recherche informatique de l'université Paris sud, Février 2004

L'étude, porte sur le datamining, les modalités de mise en place de ces techniques et aussi sur le Data warehouse et les données qui y sont stockées.

J.L. ALVAREZ, J. MATA, J.C. RIQUELME, «Data mining for the management of software development process », International Journal of Software Engineering and Knowledge Engineering

Article scientifique, il présente une étude portant sur une nouvelle méthode d'application de techniques de datamining lors de la gestion de projet informatique (développement de logiciel). Cette méthode combine deux outils : le premier basé sur l'apprentissage supervisé et le second l'apprentissage non supervisé. L'objectif de cette méthode est d'induire un ensemble de règles de gestion qui rendent facile le processus d'élaboration pour les gestionnaires. Selon la façon et l'ensemble sur lequel sera appliquée la méthode, cette dernière permettra une première analyse, un suivi du projet ou analyse finale lors du bilan.

L'article est disponible en ligne, à l'adresse suivante :

<http://www.lsi.us.es/~riquelme/publicaciones/3.8%20ijseke.pdf>

Richi Nayak, and Tian Qiu, « Data Mining Application in a Software Project Management Process »

Au cours du développement d'un logiciel, informatique, de nombreux problèmes surviennent. La résolution de ces derniers peut parfois être longue, coûteuse en temps et

ressources et surtout engendrer encore plus de difficulté qu'à l'origine. Cet article met en avant l'utilité des techniques de datamining afin de trouver et résoudre ces problèmes. De plus, ces méthodes permettent de trouver des connaissances précieuses sur le processus de développement personnel de chaque membre de l'équipe.

La publication est disponible à l'adresse suivante :

<http://eprints.qut.edu.au/1472/1/1472.pdf>

Stéphane TUFFERY, Data Mining et statistique décisionnelle, l'intelligence des données, 2010

Ce livre aborde le Datamining selon différents aspects : ses buts, ses domaines d'applications, son déroulement, etc.

Le livre peut être acheté en magasin ou alors visualiser, en partie, en ligne à l'adresse suivante :

http://books.google.fr/books?id=AyIYAA4a2kC&printsec=frontcover&hl=fr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

B. Élément Techniques lié à l'environnement de développement

Brian W. Fitzpatrick, Ben Collins-Sussman, Team Geek : À software Developer's Guide to working Well With Others, Août 2012

Brian W. Fitzpatrick dirige le front de libération de donnée de Google et Ben Collins-Sussman est l'un des développeurs fondateurs de SVN, qui dirige actuellement une équipe d'ingénierie pour le réseau d'affiliation Google.

Le livre a pour but d'aider les programmeurs à devenir plus efficaces lors de la création de logiciels. Il vise à améliorer le processus de développement personnel de chacun en donnant des astuces et axes de réflexion. Cela tourne autour du développement, mais aussi de la communication et de la collaboration avec l'équipe. Disponible seulement en Anglais, ce livre peut être commandé sur des sites tels qu'AMAZON.fr

Philip Johnson, Hongbing Kou, « Hackystat : a framework for collection, analysis, visualization, interpretation, annotation, and dissemination of software development process and product data »,code.google.fr

Il s'agit d'un blog regroupant l'ensemble des instructions et informations nécessaire à l'installation et l'utilisation du plugin Hackystat avec l'IDE Eclipse.

Des vidéos de Philip Johnson sont d'ailleurs disponibles sur youtube.fr pour plus de précision.

L'ensemble des informations relatives à ce plugin se trouve à l'adresse suivante :

<https://code.google.com/p/hackystat-sensor-eclipse/>

C. Gestion de projet informatique et Bâtiment

Agence Laurent Bansac Architectures, « Étapes de conception et construction de bâtiment », www.architecte-batiments.fr

L'article présente les étapes de conception et de construction d'un bâtiment avec la collaboration d'un architecte.

L'article est disponible à l'adresse suivante :

<http://www.architecte-batiments.fr/etapes-de-conception-et-construction-de-batiment/>

Antoine Crochet-Damais « Gestion de projet informatique : décryptage », www.journaldunet.com

L'article traite de la gestion de projet appliqué a domaine informatique. Tout en faisant une présentation de ce domaine, il présente aussi les problématiques qui lui sont propres.

L'article se trouve sur le lien suivant :

<http://www.journaldunet.com/solutions/dsi/gestion-de-projet-informatique-decryptage.shtml>

Denis Morand, « Thèse : Liaison entre la conception et la gestion de projet de bâtiments : PROJECTOR, un prototype pour la planification », [Université de savoie](http://www.univ-savoie.fr)

Cette thèse présente une solution afin d'améliorer la liaison entre la conception et la gestion de projet de bâtiments en intégrant les connaissances nécessaires dans le système informatique PROJECTOR.

La thèse est disponible sur le lien :

http://tel.archives-ouvertes.fr/docs/00/68/99/88/PDF/ThA_se_21-01-05.pdf

Annexes

ANNEXE 1 : Synthèse en anglais

ANNEXE 2 : Liste d'indicateurs de suivi de projet

ANNEXE 3 : REDMINE et les données

ANNEXE 4 : Les modèles de données

ANNEXE 1 : Synthèse en Anglais

"Being always more efficient in always less time" as seems to be the perfection that seeks to reach the contractors and youth professionals these days.

More and more connected, thanks to new technologies, the companies are facing global competition which is becoming increasingly fierce. To carry out their projects, whether it's for a national market or not, they need to achieve an intense Strategic foresight at all levels: monitoring of competitors, dialogue with multiple stakeholders from all horizons, study of the market and its trends, etc.

This watch generates a mass of data to evaluate and analyse very consistent. More important than 10 years ago, thanks to the internet that allows rapid exchange, this volume must be treated and husked enough quickly to allow time for reflection to decision-makers.

In summary, to complete a project well and make decisions by limiting the risks, a company needs, on one hand, to centralize its data and update it regularly. And on the other hand, it needs efficient and effective tools for filtering, calculations and analysis. Thus, leaders need solutions to help them in their day-to-day work: the project management softwares and the business intelligence softwares.

The purpose of this paper is to highlight the possible interaction between these two domains, and the possible applications of data mining techniques in project management. We could thus optimize the work early and allow better strategic business management.

The race for innovation and performance is more than necessary for businesses today. In fact, any company wishing to be competitive, regardless of its niche, must master its production and therefore its project management.

It is the first pillar of a company. Generally, it consists of 4 steps: the design (conception), the planning, the implementation and the ending. Whatever the area studied, we always find these steps, with sometimes variations. If we consider two distinct domains, the buildings and the computing, we will see, that for the first, the realization is the most critical phase, while for the second, it is the conception phase.

Indeed, in the first case, if a building is poorly built, the project falls into the water. Cannot not be inhabited or corrected, the building must be destroyed to be rebuilt which has an important cost.

On the contrary, in the computer science, there is a whole process of verification during the realization of the project, which prevents it to fail. The most critical phase is the analysis (initialization) because if it is poorly thought, then the project would be badly estimated, which will cause an overflow of budget, time and certainly a dissatisfaction of the customers and the providers.

Although different, computing and the building have two points in common. The first is the need to control information and the second the insufficiency of the existing indicators and their prefixed character.

Whether it's before launching a building restoration or the creation of a software, the decision-makers need to know under what conditions the project is profitable, and under which it is not. This implies the study of the data of past and present projects in order to predict future situations. But between the staff turnover, the multiplicity of tools and stakeholders, it is not always easy to have the right information at the right time.

The second common point, concerning the inadequacy of the current indicators, highlights the lack of reference and tools for decision-makers. The indicators are based on relevant data but past. They therefore provide a statement (a presentation) of the past or present, but not future. To predict the future and make prognoses, leaders rely on their experience and intuition. Nothing factual, which is not without risk. Finally, the prefixed character of indicators can be problematic, since it hides the environmental parameters.

Project management tools, will allow companies to centralize and improve their project management. Managers, Team members, and Project leaders will be more effective because the information will flow between them and that they will be more organized.

From the initialization to the end, everything will be centralized, stored and communicated to stakeholders. The cohesion of the team and the success of the projects are ensured.

Therefore, it is the need of a global and transversal vision that pushes leaders to be equipped with the most sophisticated solutions.

Business Intelligence (BI) is a pillar for decisions-makers. It offers a transverse vision of the company at any moment. Through methods of calculation, the company can establish indicators of progress, or explain its turnover per point of sale or per city for example. The explanatory character of BI sufficed for years to policy makers. They could then make their future strategic choices only by relying on their experience. However, faced a dynamic economic and technological context, today, they need more certainty. Due to the explosion of the data volumes, the decision-makers had to find efficient and rapid techniques for the analysis, comprehension and the detection of opportunities. They were looking for more security in their process: they wanted the tools to help them to make more prediction but based on facts with a controlled risk. Those are the main reasons of the appearance of the data mining.

New science of the past 20 years, the Data mining is the exploration of large amounts of data in order to find correlations between distinct elements a priori unrelated.

By exploring data, Data Mining techniques reveal relevant information. More or less reliable, that information would allow managers to make predictions more serene.

At the crossing of artificial intelligence, statistics and database treatments, the data mining is the process of knowledge discovery into the data. By using different techniques, it studies all the data in order to find a correlation. If it find one, it presents it to an expert which will validate or no the discovered relation. If it not, a new data study cycle will be done, and so on until the expert validates one of the relations.

Once one of that is validated, we can, therefore, consider that a new knowledge was discovered. On top of being used for the reporting, the new information would be add in the original batch (the database), in order to create a new batch to analyze.

This is how work the mechanism of the knowledge discovery.

It is therefore an iterative and interactive process composed of four phases: identification of the field of study, the data preparation, the action on these ones and then the study results.

Each iteration requires human intervention because the results are submitted to the appreciation of people.

The difference is that those one are not experts, but simple users with a good knowledge of the profession or the area of study. The Data mining, therefore, demystifies data analysis by opening it to all those of the job.

Declined in several subdomains, the data mining can be applied to all possible data formats. Text Mining, Image Mining, Multimedia mining and web mining represent respectively the exploration of text, images, multimedia and web content. The latter is widely used for the email marketing and web analytics. It allows in the first case, to make of the marketing a competitive advantage and an efficient real-time process. And in the second case, it allows to track the return of investment (economic indicator) of digital projects and, thus, to consolidate strategies.

Ally the data mining and the project management is possible. Several solutions and studies exist. The chosen one was wanted accessible to all. It is an architecture composed by two environments. The first is linked to the activity of the company (the production environment), the second, to the storage and the management of data.

If the first context depends on the company and its wishes, it is nevertheless the starting point of the architecture. Regardless of the size, type and business corporations, the generated data must be large, detailed, different and diverse. Having different visions of the same project at the same time would be also very interesting.

All of these production data will serve as bases for data mining techniques.

The second half of the architecture is the key component of a BI solution: it is composed by an ETL and a data warehouse. If the production environment was free from any constraint except generate a huge amount of data, the storage is just the opposite. In BI computing, complex queries and analyses of several terabyte of data must have a low answer times. For this a data warehouse is necessary. The structure of this tool was thought in order to allow quick access to complex information. To feed it, it is necessary to use an ETL. Its name means : Extract, Transform and Load. Therefore, as its name supposed it, it is the responsible tool for the extraction of data from operational (production) databases and the transformation of the data so that they can suit the data warehouse format. Once everything is done, the ETL will load the transformed data in the Data warehouse.

Thus, this tool is the intermediary between the production and decision-making environment: it is the interpreter which makes possible the connection between the two domains.

Once all things set up, the data mining is then possible.

Many studies conducted around the world, have proven that the “data mining-project management” alliance is a powerful competitive advantage: by improving the process of personal development of each developer, the entire team and therefore the projects improved significantly. The discovery of bug and their location could be anticipated by the data mining process, which facilitated their correction. The process of estimation of time, budgets and thus predictions become safer: the quality of the project arise then.

By detecting the disparate teams or those who matched, a reorganization of the teams is possible in order to improve group cohesion.

All these advantages, demonstrated scientifically, highlight the beneficial contribution of the data mining in the project management.

However, we need to qualify because this architecture, although simple in appearance, has two main constraints: the volume of data and the price of the Data warehouse.

Although many companies have a lot of activities, they do not generate all large volumes of data that requires the use of data mining. The risk is manageable, and the prediction in their scale is possible without the need for major human and material investments.

In addition, although most have a need in decision-making, all cannot invest thousands of euros for a data warehouse, especially since it requires maintenance and therefore resources that are fully dedicated to it.

Thus, the benefits of the data mining are undeniable in the context of project management; however, the necessary investments must be made according to the size and the real needs of the company. Long and costly in resources, the choice of setting up such an architecture has to be made in full knowledge of the facts, and only if the return of investment is good enough.

ANNEXE 2 : Liste d'indicateurs de suivi de projet

Lors de la réalisation d'un projet informatique, il existe des indicateurs détaillés permettant de suivre l'avancement de projet tâche par tâche, ainsi que leur consommation de ressource et de temps.

- Soit les charges prévues, **PRV**, l'ensemble des charges prévues pour une tâche ou un projet donné.
- Soit les charges consommées, **CC**, l'ensemble des charges consommées pour une tâche ou un projet donné.
- Soit le reste à engager, **RAE**, le reste à faire sur une tâche donnée. Ce dernier correspond à l'estimation de la charge du travail restant pour finir le projet ou la tâche considérée.

Ces trois indicateurs vont permettre de suivre précisément l'évolution d'un projet :

- L'estimation globale du projet **EG** : $EG = CC + RAE$
- La production, **PRO** : $PRO = PRV - RAE$
- La productivité des équipes, **ProdEq** : $ProdEq = PRO / CC$
- L'écart de production, **ECT** : $ECT = PRO - CC$
- Le pourcentage d'avancement de la production, **%AvProd** : $\%AvProd = PRO / EG$
- Le pourcentage d'avancement de consommation, **%AvCon** : $\%AvCon = CC / EG$

Ces indicateurs sont les plus communément utilisés dans la gestion de projet et cela, quel que soit le métier de l'entreprise. Connus aussi sous d'autres noms, ils peuvent aussi bien être calculés par mois ou par semaine, ou pour chaque tâche du projet, chaque équipier ou le projet en lui-même.

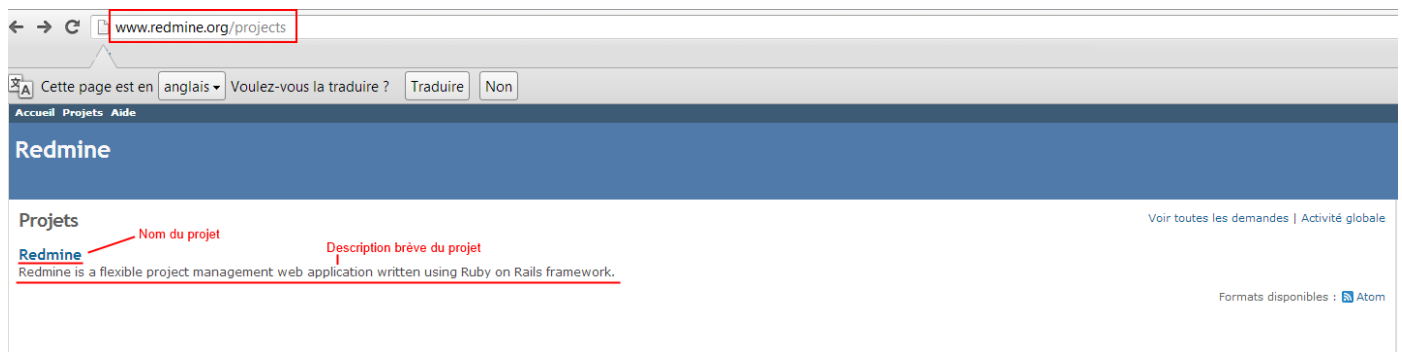
Il est ainsi possible, régulièrement, de comparer l'avancement réel du projet, par rapport à ce qui a été prévu et surtout par rapport aux charges prévues de départ.

ANNEXE 3 : REDMINE et les données

Les illustrations sont issues du site : www.redmine.org/

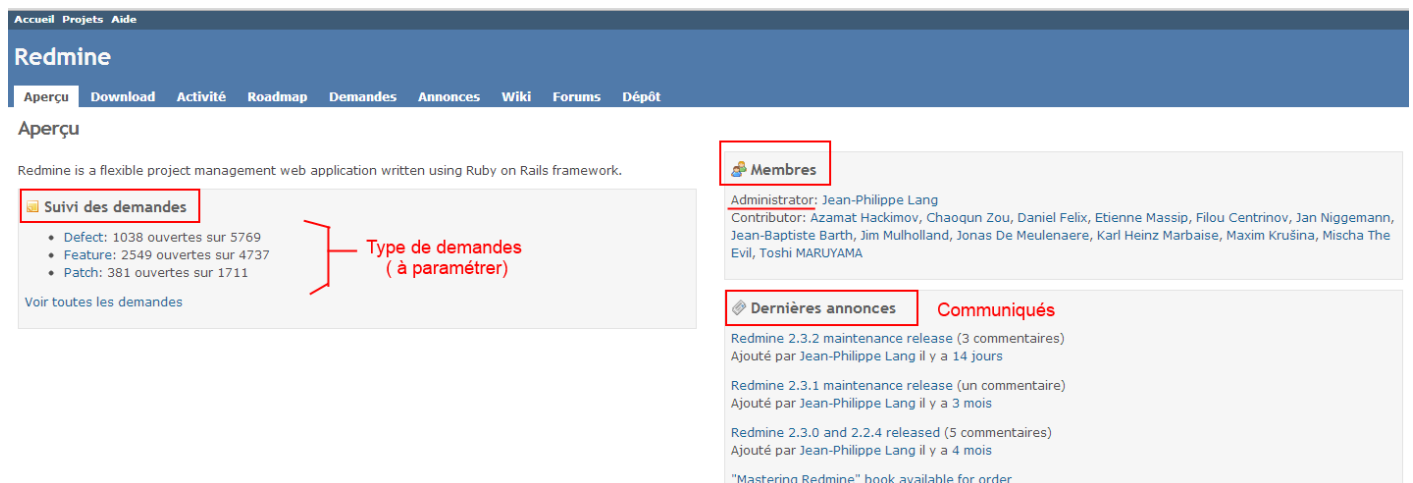
Les créateurs de cet outil le gèrent comme un projet informatique et utilisent cet outil pour cela.

1) Liste des projets sur REDMINE



Dans ce cas-là, il n'y a qu'un seul projet à gérer : le projet REDMINE

2) Liste des éléments d'un projet



L'ensemble du style et des onglets affichés sont modifiables. Le projet étant open source*, les personnes le souhaitant peuvent afficher ce qu'il souhaite.

146

Accueil

Projets

Aide

Redmine

Aperçu

Download

Activité

Roadmap

Demandes

Announces

Wiki

Forums

Dépôt

Demandes

▼ Filtres

☒ Statut

► Options

Appliquer

Effacer

ouvert ▼

ouvert

égal

différent

fermé

tous

→ Possibilité de filtrer les demandes selon leur statut

Ajouter le filtre

#	Tracker	Statut	Sujet	Mis-à-jour	Catégorie
14562	Defect	New	dif of CJK (Chinese/Japanese/Korean) is broken on Ruby 1.8	2013-07-28 13:06	
14557	Defect	Reopened	Error (stack level too deep) when creating new project	2013-07-28 15:22	Projects
14555	Feature	New	Custom fields UserList Watchers per user	2013-07-27 10:22	Custom fields
14554	Defect	New	Diffs from subprojects return 404	2013-07-28 13:32	SCM
14553	Feature	New	Filter for custom field of type version	2013-07-26 15:28	Custom fields
14548	Feature	New	Add one more option for Users display format	2013-07-26 09:17	UI
14547	Feature	New	Change the location between project column and Subject column	2013-07-26 06:16	UI
14546	Defect	New	Maximum number of items in Atom Feeds of redmine.org (25) is too little.	2013-07-26 00:53	Website (redmine.org)
14545	Defect	New	Email notifications sent for recent projects only (created after setting up the SMTP connection) can add "atom","csv","json","psd" option in setting for different level user to be shown for them	2013-07-25 20:40	Email notifications
14540	Feature	New	Running redmine on rails 4	2013-07-25 10:03	
14534	Feature	New	"Close completed versions" button not working	2013-07-24 13:54	Rails support
14533	Defect	New	mySQL2 gem under Windows is failing for version 0.3.13	2013-07-24 12:38	Roadmap
14528	Defect	New	Git-smart-http authentication by repository	2013-07-26 11:28	Gems support
14527	Patch	New	Email subject with Task or just Watcher	2013-07-24 00:16	SCM extra
14523	Feature	New	Add view hook to associated revision partial	2013-07-24 05:07	Email notifications
14509	Patch	New	Workflow administration screen - loss of overview when working with many statuses	2013-07-21 16:41	Issues
14508	Feature	New	Multiple repositories not visible if main repository is empty	2013-07-21 08:52	Issues workflow
14506	Defect	New	Link to a file in a named repository in another project	2013-07-20 13:24	SCM
14505	Feature	New	Permission check on an unused custom field leads to unwanted restriction for the ticket status	2013-07-27 19:25	Text formatting
14496	Defect	New	MailHandler: Unable to determine target project (when allow_override=project and project=unassigned is used)	2013-07-19 10:57	permissions and roles
14491	Defect	New	Wrong CSS for large menu	2013-07-19 20:11	Email receiving
14489	Defect	New	Delete issue history	2013-07-18 14:39	UI
14487	Feature	New	Add column "cumulated spent time" to issue filters	2013-07-26 05:41	Issues
14483	Patch	New	Fix for Issue #13544	2013-07-18 02:02	Issues
14476	Patch	New		2013-07-17 14:06	SCM

1 2 3 ... 159 Suivant » (1-25/3458)

Identifiant

Statut

Date de dernière modification

Description/Sujet de la demande

Formats disponibles : Atom | CSV | PDF

4) Visualisation d'une demande

Defect #14562 ——— Identifiant de la demande

diff of CJK (Chinese/Japanese/Korean) is broken on Ruby 1.8 ——— Titre de la demande

Ajouté par Toshi MARUYAMA il y a environ 7 heures. Mis à jour il y a environ 7 heures.

Statut:	New	Début:	
Priorité:	Normal	Echéance:	
Assigné à:	-	% réalisé:	0%
Catégorie:	-		
Version cible:	2.3.3		
Affected version:	2.3.2	Resolution:	

Description

From #13644.

Correct:

```
1 --- a.txt 2013-07-27 06:03:49.133257759 +0900
2 +++ b.txt 2013-07-27 06:03:58.791221118 +0900
3 @@ -1,3 +1,3 @@
4  aaaa
5  - 日本記
6  + 日本娘
7  bbbb
```

Broken:

View differences: ☒ inline ☐ side by side

```
b.txt 2013-07-27 06:03:58.791221118 +0900
1 1 aaaa
2 2 日本000
3 2 日本000
3 3 bbbb
```

correct.png (11,338 ko) Toshi MARUYAMA, 2013-07-28 12:46
broken.png (8,569 ko) Toshi MARUYAMA, 2013-07-28 12:46
issue-13644-3.diff (142 octet) Toshi MARUYAMA, 2013-07-28 12:46

Historique

Mis à jour par Toshi MARUYAMA il y a environ 7 heures

#1

- Sujet changé de *diff of CJK(Chinese/Japanese/Korean) is broken on Ruby 1.8* à *diff of CJK (Chinese/Japanese/Korean) is broken on Ruby 1.8*

Révisions associées

Chaque révision est
rattaché à un commit

Révision 12046
Ajouté par Toshi MARUYAMA il y a environ 7 heures
fix diff of CJK is broken on Ruby 1.8 (#14562)
Contributed by Jun NAITOH.

Révision 12047
Ajouté par Toshi MARUYAMA il y a environ 7 heures

Ensemble d'information
détaillant une demande

L'ajout de commentaire ou de remarque est aussi possible pour une demande en particulier. Cela permet par exemple d'expliquer les contraintes techniques rencontrées ou les solutions adoptées.

ANNEXE 4 : Les modèles de données

1. Le modèle relationnel

C'est le modèle de structuration des informations respecté par les systèmes de Gestion de bases de données relationnelles (SGBDR). Dans les SGBDR, les informations sont rangées dans des tables.

Ex. : la table « Employé » possède des attributs « date de naissance » « Sécurité sociale », « nom », « prénom » qui la caractérise.

Une base de données relationnelle est un ensemble de tables relationnelles.

Les tables sont reliées entre elles par des relations soit du fait de clé primaire ou de clé étrangères.

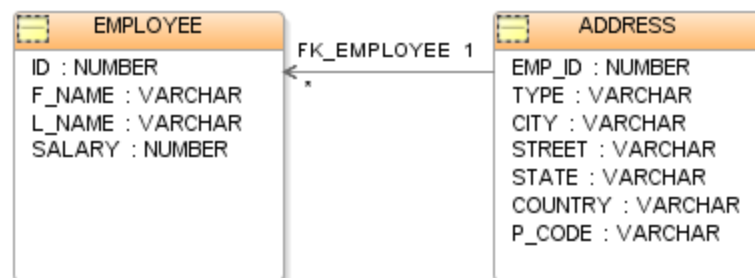
Une clé primaire dans une table est un ensemble d'attributs qui détermine tous les autres.

Ex. : Dans la table « Employé », le numéro de sécurité sociale sera notre clé primaire, car on ne peut avoir deux personnes avec un même numéro. Ainsi si l'on a un numéro de sécurité sociale on peut retrouver la personne.

Une clé étrangère est une clé primaire dans une autre table.

Ex : une adresse appartient à un employé, ainsi dans la table adresse on retrouvera un attribut identifiant désignant la personne.

On dit de cet attribut « identifiant » (noté ID) qu'il est une clé étrangère dans la table adresse.



Modèle relationnel

Ce modèle est le plus répandu pour les bases de production des entreprises.

2. Le modèle en étoile

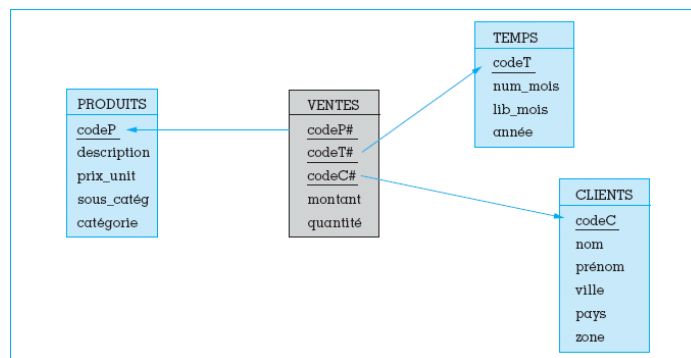
Dans les datawarehouse ou datamart, le modèle de données « en étoile » est typique des structures multidimensionnelles stockant des données atomiques ou agrégées.

Le modèle en étoile est souvent considéré (à tort) comme un modèle dénormalisé, ce qui permet une économie de jointures à l'interrogation. Il est ainsi optimisé pour les requêtes d'analyse.

La table située au centre de l'étoile est la table des faits ou mesures (ou encore métriques) : ce sont les éléments mesurés dans l'analyse comme les montants, les quantités, les taux, etc.

Les tables situées aux extrémités de l'étoile sont les tables de dimensions (ou encore, axe d'analyse) ou niveaux de suivi. Ce sont les dimensions explorées dans l'analyse par exemple le temps (jour, mois, période, ...), la nomenclature des produits (référence, famille, ...), la segmentation clientèle (tranche d'âge, marché, ...), etc.

Le principe d'optimisation de ce modèle en étoile est le suivant : une clé calculée "technique" (clé générique) sert de jointure relationnelle entre les tables de dimensions et la table des faits. La requête SQL réalise d'abord sa sélection sur les tables de dimensions (peu volumineuses) et ensuite seulement, à partir des clés ainsi sélectionnées, la jointure avec la volumineuse table des faits.

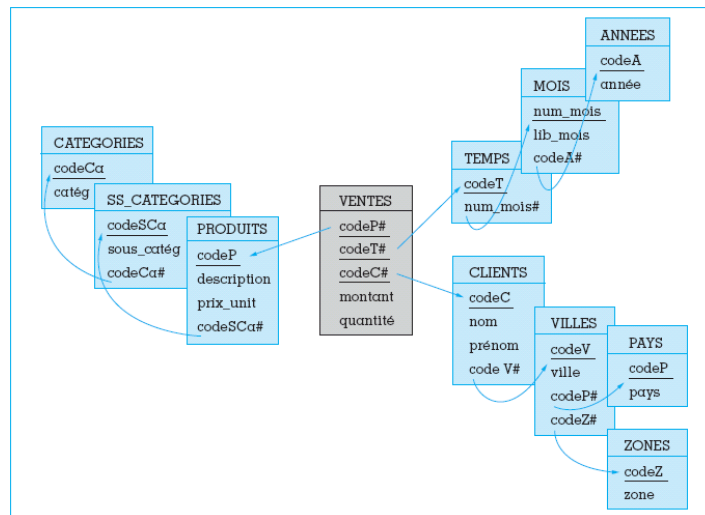


Modèle en étoile

Attention, un modèle flocon n'a qu'une seule table de dimension par axe, contrairement au modèle en flocon qui peut en posséder plusieurs d'où son nom.

3. Le modèle en flocon

Le modèle de données dit « en flocon » est une variante du modèle en étoile : chaque table de dimension est renormalisée pour faire apparaître la hiérarchie sous-jacente (nomenclature, ...). La normalisation n'est pas indispensable, car ni mises à jour ni suppressions ne sont effectuées directement sur l'entrepôt de données. L'intérêt principal du modèle en flocon réside dans le gain en espace de stockage qui est de l'ordre de 5 à 10 %. Un ensemble d'étoiles ou de flocons dans lequel les tables de faits se partagent certaines tables de dimensions forme un modèle « en constellation ».



Modèle en flocon

L'ensemble de ces modèles peuvent être implémenté sur une base de données SQL server, Oracle, Mysql. Il faut juste prendre en compte le volume présent et futur des données, afin d'être prêt à toutes les éventualités.