



# Plans d'action pour une mise en oeuvre réussie des Big Data

Quatre scénarios pour réussir

## Introduction

Nul doute aujourd'hui que les Big Data entraînent une profonde mutation du paysage technologique. Selon IDC, le volume de données utiles sera multiplié par 20 entre 2010 et 2020, alors que 77 % des données pertinentes des entreprises seront non structurées jusqu'à l'horizon 2015.<sup>1</sup> À l'heure où ces tendances se confirment, tant en volume qu'en diversité, les entreprises se tournent de plus en plus vers Hadoop, NoSQL et d'autres outils pour surmonter les obstacles qu'elles n'arrivent pas à franchir avec les anciennes technologies, comme les bases de données relationnelles et les entrepôts de données.

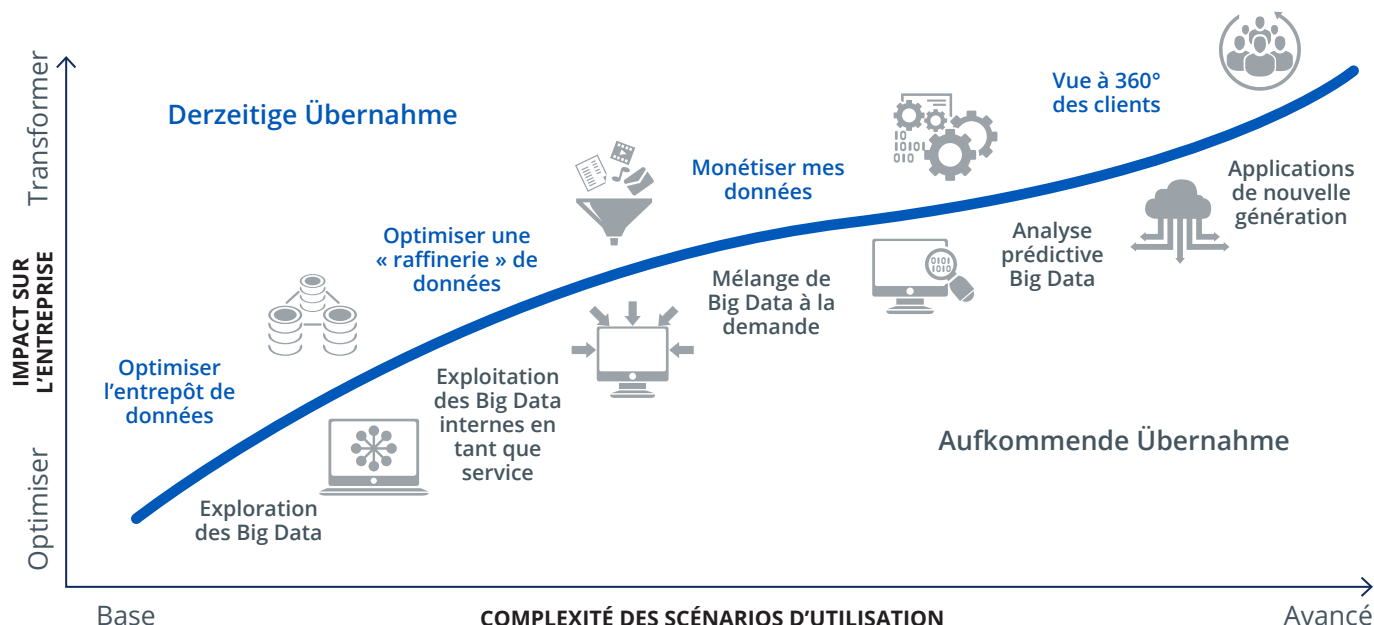
Bien que les opportunités d'exploitation des Big Data se multiplient rapidement, l'étude montre que les entreprises se posent deux grandes questions à cet égard : comment exploiter les données massives et comment élaborer une stratégie Big Data efficace ?<sup>2</sup> Face à ces enjeux, ce document a pour but d'identifier et de décrire les scénarios d'utilisation de Big Data susceptibles de générer des résultats commerciaux, mais aussi d'anticiper les futurs cas d'utilisation. Dans les pages suivantes, vous découvrirez quels sont ces scénarios, pourquoi les entreprises choisissent de les mettre en oeuvre et quelles sont les architectures de référence les plus courantes. L'illustration ci-dessous présente 10 scénarios d'utilisation de Big Data, classés selon leur impact sur l'entreprise (axe Y) et le degré de complexité de leur mise en oeuvre (axe X). L'impact de ces scénarios peut

aller d'une simple optimisation des processus actuels jusqu'à la transformation totale des modèles de gestion. Le degré de complexité varie selon qu'il s'agit de mises en oeuvre de base, reposant sur des technologies plutôt standard, ou de projets plus avancés faisant intervenir plusieurs technologies, dont certaines sont peu commercialisées. Les scénarios d'utilisation sont classés en deux catégories « Adoption actuelle » ou « Adoption à venir ». La première catégorie regroupe les mises en oeuvre les plus répandues qui suivent des consignes faciles à reproduire, alors que la deuxième regroupe des mises en oeuvre encore rares aujourd'hui, mais dont on sait qu'elles vont se développer. Ce livre blanc décrit en détail les scénarios d'utilisation appartenant à la catégorie « Adoption actuelle ».

---

1 Étude IDC sur l'univers numérique, 2012.

2 Rapport Gartner « Big Data Adoption in 2013 Shows Substance Behind the Hype », 2013.



## Brève description de chacun de ces scénarios :

### ADOPTION ACTUELLE

**Optimiser l'entrepôt de données** – L'entrepôt de données traditionnel est submergé par l'énorme volume des données, et les différentes parties prenantes n'obtiennent pas en temps voulu les analyses dont elles ont besoin. Augmenter la capacité de l'entrepôt de données peut se révéler très onéreux. C'est pourquoi les entreprises confient les données les moins utilisées à des entrepôts Big Data pour améliorer les performances de l'entrepôt de données.

**Optimiser une « raffinerie » de données** – Le entrepôt Big Data centralise et traite les données issues des différentes sources avant de les transmettre en aval à des outils d'analyse à faible latence (requêtes rapides dans une base de données analytique, par exemple). Les coûts liés à l'ETL et à la gestion des données diminuent considérablement et les Big Data deviennent un élément essentiel de l'analytique.

**Vue à 360° des clients** – Mélangez plusieurs sources de données opérationnelles et transactionnelles pour créer, à la demande, une vue analytique de vos principaux points de contact client. Les partenaires et les employés

en contact direct avec les clients peuvent accéder aux informations dont ils ont besoin directement depuis les applications métier qu'ils utilisent au quotidien.

**Monétiser mes données** – Dans ce cas d'utilisation, des jeux de données enrichies et dépersonnalisées peuvent être mis à la disposition de clients tiers sous forme de service. En tirant parti de puissants outils de traitement de données et d'une analytique intégrée, vous pouvez générer de nouvelles sources de revenus pour votre entreprise.

### ADOPTION À VENIR

**Exploration des Big Data** – Si les entreprises transmettent un volume colossal de données vers les entrepôts Big Data, elles ne savent pas pour autant quelles informations ces magasins renferment (« données obscures »), ni s'il existe un moyen d'exploiter efficacement ces données. Pour commencer, les analystes exécuteront des algorithmes de Data Mining de base et s'efforceront de corrélérer les tendances identifiées avec les données en provenance d'autres sources.

**Exploitation des données des équipements et des capteurs** – Jusqu'à présent, il était impensable, et surtout hors de prix, d'appliquer l'analytique aux données massives produites par les capteurs, les routeurs et les décodeurs de télévision. Aujourd'hui, en revanche, les Big Data permettent d'exploiter ces informations via le Data Mining et les analyses à faible latence. Les entreprises sont alors en mesure de corriger rapidement les problèmes de fonctionnement et de service.

**Analyse prédictive des Big Data** – Les Big data offrent un nouvel ensemble d'outils pour optimiser les algorithmes d'intelligence artificielle (formation et évaluation) et les utiliser pour prévoir ou influencer les résultats (notation). L'exécution d'analyses prédictives dans le entrepôt Big Data peut s'avérer utile pour la détection de fraudes, les moteurs de recommandation et l'optimisation des offres.

**Applications de nouvelle génération** – Si le cloud computing et le mode SaaS n'ont rien de nouveau, leur évolution dépendra très vraisemblablement des Big Data. Les fournisseurs d'applications ne cessent d'innover dans le domaine des architectures de données et d'analytique afin de rendre leurs produits plus puissants, plus intelligents et plus utiles aux clients. En

intégrant une interface analytique à leurs applications utilisateur, les fournisseurs peuvent tirer pleinement parti de leurs innovations. Mélange des Big Data à la demande – Une fois les entrepôt de Big Data mis en oeuvre, les équipes informatiques sont encore souvent soumises aux contraintes de temps de l'infrastructure d'entrepôt de données existante. Lorsque le facteur temps joue un rôle déterminant, il est nécessaire d'outrepasser l'entrepôt de données. Le fait de mélanger les données « à la dernière minute » permet d'éviter le stockage temporaire des données et d'alimenter l'analytique avec des données précises et actualisées, en provenance de toutes les sources.

**Exploitation des Big Data internes en tant que service** – Les entreprises exploitent les Big Data en tant que service de base de données partagé pouvant être diffusé à plusieurs équipes de développement d'applications afin de faciliter l'ingestion des données et leur accès. Le but est de réaliser des économies d'échelle et de réduire les coûts par rapport à une approche plus cloisonnée. Les solutions d'ETL et d'analytique sont intégrées sous forme de composants à la pile centralisée de l'entreprise.

# Optimiser l'entrepôt de données

## DESCRIPTION ET AVANTAGES POUR LES ENTREPRISES

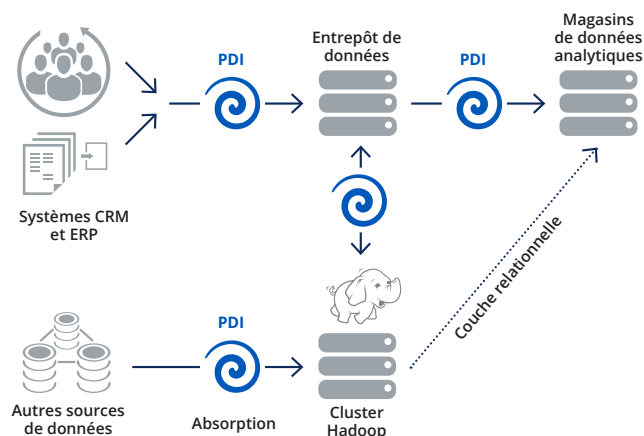
L'optimisation de l'entrepôt de données compte parmi les scénarios d'utilisation des Big Data les plus fréquents, principalement pour des raisons de coût et de performance opérationnelle. À mesure que le volume de données à stocker et à utiliser augmente, la capacité de l'entrepôt de données existant risque de devenir limitée. Les performances des requêtes finissent par se dégrader et les utilisateurs et informaticiens ont de plus en plus de difficultés à accéder aux données. Les entreprises n'ont souvent pas d'autre choix que d'accroître la capacité de stockage de leur entrepôt de données en s'adressant aux fournisseurs en place - solution très coûteuse et souvent temporaire puisque le volume des données ne cesse d'augmenter.

Par conséquent, les entreprises se sont tournées vers les Big Data, Hadoop plus particulièrement, pour alléger cette pression. Le modèle d'informatique distribuée d'Hadoop permet de disposer d'une énorme capacité de traitement sur du matériel standard. Il revient beaucoup moins cher de stocker les données dans HDFS (Hadoop Distributed File System) plutôt que dans un entrepôt de données traditionnel. Le coût de stockage Hadoop est d'environ 1 000 \$ par téraoctet (To) contre environ 5 000 à 10 000 \$ par To (voire plus) pour un stockage à pleine charge dans un entrepôt de données (matériel requis, serveur, etc. compris).<sup>3</sup> Ainsi, les services informatiques transféreront les données les moins utilisées de leur entrepôt de données vers Hadoop pour réduire les coûts de stockage, tout en continuant à livrer les données en temps et en heure conformément aux accords de niveau de service.

## EXEMPLE

Dans cet exemple, une entreprise exploite des données en provenance de systèmes CRM et ERP ainsi que d'autres sources. Un cluster Hadoop a été mis en place pour alléger l'entrepôt de données existant des données les moins utilisées. Dans la mesure où les analystes

accèdent aux informations depuis le entrepôt de données analytique, les coûts de stockage diminuent et les requêtes s'accélèrent.



## PRINCIPAUX POINTS À PRENDRE EN COMPTE

Si l'optimisation de l'entrepôt de données est l'un des scénarios d'utilisation des Big Data les plus fréquemment mis en oeuvre, il exige néanmoins du temps, des efforts et une planification. Hadoop est encore une technologie émergente et l'utilisation des outils « prêts à l'emploi » livrés avec les distributions Hadoop demandent certaines connaissances. Une expertise du codage Java est indispensable pour créer les routines qui transmettent les données de l'entrepôt de données vers Hadoop. Les entreprises ont du mal à embaucher des développeurs et des analystes qualifiés en nombre suffisant. Les experts Hadoop se font rares sur le marché et demandent des salaires près d'une fois et demie supérieurs à ceux des spécialistes en SQL ou autres outils plus traditionnels.<sup>4</sup>

Pour faciliter l'intégration des Big Data, Pentaho propose une interface utilisateur graphique intuitive qui élimine la programmation manuelle et rend Hadoop accessible à tous les développeurs de données. Tout ceci contribue à accélérer le retour sur investissement et à réduire les coûts de personnel. Si les entreprises disposent déjà d'une solution d'intégration de données, les plateformes existantes recourent encore à la programmation pour intégrer des sources et des bases de données existantes avec Hadoop.

3 Information Week, "How Hadoop Cuts Big Data Costs," 2012.

4 O'Reilly, "2013 Data Science Salary Survey," 2013.

# Optimiser une « raffinerie » de données

## DESCRIPTION ET AVANTAGES POUR LES ENTREPRISES

Face à l'explosion des volumes de données (transactions structurées, client et autres), les systèmes ETL traditionnels deviennent plus lents, rendant l'analytique inexploitable. Le concept de « raffinerie » permet de réunir en un seul lieu toutes les sources de données via un centre de traitement Big Data évolutif, en faisant intervenir Hadoop pour la transformation. Les données affinées sont ensuite transmises vers une base de données analytique en vue d'un traitement à faible latence.

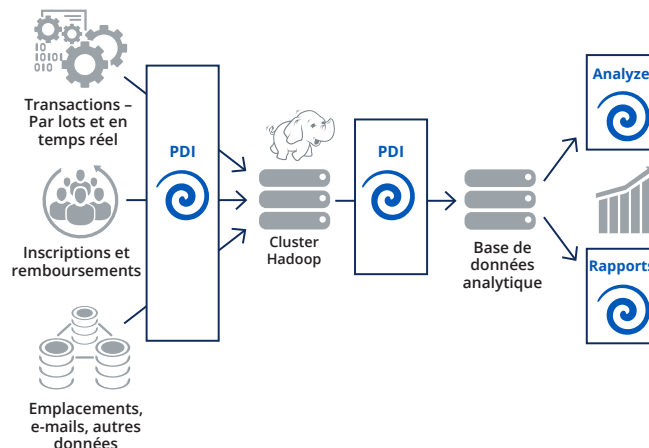
Ce scénario d'utilisation s'inscrit dans la logique d'économie et d'amélioration opérationnelle de l'optimisation de l'entrepôt de données. À ce stade, il est possible de charger dans Hadoop une grande quantité de données hétérogènes. Plus qu'un simple outil d'archivage, Hadoop devient une véritable mine de données métier multisources prêtes à être interrogées. Ce projet engendre davantage de transformation qu'une simple optimisation de l'entrepôt de données. L'entreprise peut exploiter l'analytique sur diverses sources de données massives grâce à des requêtes accélérées, une ingestion rapide et un traitement puissant rendu possible par la combinaison d'Hadoop et d'une base de données analytique (comme Vertica ou Greenplum). Par ailleurs, les équipes informatiques peuvent créer plus rapidement des jeux de données pour l'analyse prédictive.

## EXEMPLE

Une société de marketing électronique qui propose des offres personnalisées a créé une architecture de raffinerie. Les données des campagnes en ligne, des inscriptions et des transactions sont ingérées via Hadoop, traitées puis envoyées vers une base de données analytique. Grâce à un outil frontal d'analyse métier, les utilisateurs peuvent créer des rapports et des analyses ad hoc.

## PRINCIPAUX POINTS À PRENDRE EN COMPTE

Les enjeux liés au personnel et à la productivité ren-



contrés dans le scénario d'optimisation de l'entrepôt de données restent d'actualité dans ce cas d'utilisation. De toute évidence, le retour sur investissement peut être optimisé en évitant la programmation et en simplifiant l'intégration des entrepôt Big Data avec les différents systèmes relationnels. Ce projet d'intégration s'avère plus coûteux et plus fastidieux puisqu'il implique la consolidation de nombreuses connexions système point à point dans un modèle d'hub centralisé. Le projet devient plus complexe à exécuter au fur et à mesure que les types et les sources de données continuent d'augmenter. Il est donc important de sélectionner des plateformes d'intégration de données et d'analytique hautement flexibles, capables de se connecter à toute une gamme de systèmes de données, aujourd'hui comme demain. Compte tenu de l'importance des informations analytiques obtenues d'Hadoop dans ce scénario, les développeurs de données et l'analyste métier se doivent de collaborer. Seule une plateforme intégrée permet d'assurer la connectivité des données et l'analyse décisionnelle. La coordination est quasi impossible si l'équipe informatique et les utilisateurs exploitent des boîtes à outils distinctes.

Pour finir, cette architecture repose généralement sur une base de données analytiques. Ces bases de données sont optimisées pour la Business Intelligence, grâce à des requêtes accélérées, une plus grande évolutivité, des « cubes » d'analyse multidimensionnels et/ou une fonction en mémoire. Comparativement, les bases de données transactionnelles n'offrent ni les performances requises pour les requêtes ni la fonction analytique.

## Vue à 360° des clients

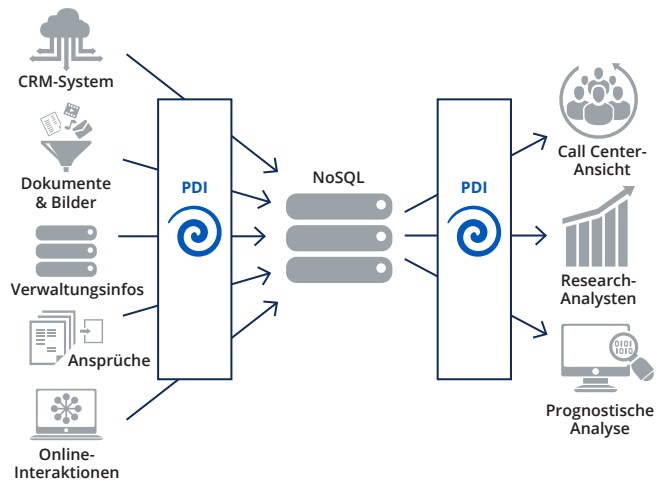
### DESCRIPTION ET AVANTAGES POUR LES ENTREPRISES

Les entreprises cherchent depuis longtemps à mélanger plusieurs sources de données opérationnelles et transactionnelles pour créer, à la demande, une vue analytique de leurs principaux points de contact client. En tirant parti des sources de données traditionnelles et des Big Data dans un environnement parfaitement intégré, les entreprises y parviennent et obtiennent un volume impressionnant d'informations clients à exploiter. Si les projets d'optimisation de l'entrepôt de données et de « raffinerie » de données sont essentiellement axés sur les coûts et l'efficacité, ce scénario vise clairement à dynamiser la valeur des clients sur toute leur durée de vie. Cette stratégie s'avère cruciale sur les marchés très compétitifs où la perte de clients est une réelle préoccupation (tels que les télécommunications, le tourisme et les services financiers grand public). Augmenter les offres de vente croisée et de montée en gamme et minimiser les risques de perte de clients sont les deux principales clés du succès. Côté back-end, ce scénario centralise les données issues de tous les points de contact client dans un référentiel unique de façon à accélérer les requêtes.

Côté frontal, il réunit les indicateurs pertinents et les met à la disposition des utilisateurs. En mélangeant des données précédemment cloisonnées, la vue à 360° des clients permet aux équipes commerciales de mieux cerner les habitudes d'achat et d'avoir une idée plus précise de la façon dont les produits et les services de la marque sont perçus. Disposant de ces informations au point d'interaction avec les clients, les employés peuvent prendre des décisions plus productives et rentables au moment opportun.

### EXEMPLE

Un établissement de services financiers absorbe les données de diverses sources dans un seul entrepôt Big Data, NoSQL dans cet exemple. À partir de là, les données sont traitées et synthétisées en fonction de l'ID unique du client de façon à avoir une visibilité totale. Des données client précises et contrôlées sont ensuite transmises aux vues analytiques de chaque rôle (notamment personnel



du centre d'appels, analystes de marché et spécialistes des données).

### PRINCIPAUX POINTS À PRENDRE EN COMPTE

Si cette mise en oeuvre peut radicalement transformer les entreprises, elle peut aussi s'avérer extrêmement complexe et mobiliser d'importantes ressources. Outre les enjeux rencontrés dans les scénarios précédents (manque de personnel qualifié et intégration point à point), les entreprises doivent prévoir une planification stratégique rigoureuse pour bénéficier d'une visibilité totale sur leurs clients.

Il convient tout d'abord d'intégrer au projet les objectifs de chiffre d'affaires. Les parties prenantes doivent identifier les facteurs potentiels de satisfaction client et les opportunités pour le personnel en contact direct avec la clientèle d'exploiter ces données. En parallèle, les utilisateurs concernés doivent participer au processus

**Augmenter les offres de vente croisée et de montée en gamme et minimiser les risques de perte de clients sont les deux principales clés du succès.**

de planification de façon à ce que les informations proviennent des bonnes sources et soient transmises de manière adéquate aux bonnes personnes. L'analytique doit être présentée aux utilisateurs de manière à remporter immédiatement leur adhésion. Il faut donc en faciliter l'accès, la rendre intuitive et l'intégrer dans des applications opérationnelles stratégiques.

D'un point de vue technique, une solution Big Data NoSQL, comme MongoDB, peut s'avérer être la solution idéale pour les entreprises voulant acheminer des flux de données client actualisées dans une collection unique pouvant être rapidement distribuée sur plusieurs serveurs. Hadoop convient davantage aux cas où les données peuvent être traitées par lots et doivent être stockées chronologiquement. Bien souvent, Hadoop et NoSQL coexistent au sein de la même architecture. Si, côté back-end, l'intégration d'un ou de plusieurs entrepôts de données Big Data avec toute une variété d'applications et de bases de données s'impose, il est très probable que, côté frontal, de nouvelles exigences apparaissent. Chaque type d'utilisateur d'analytique client nécessitera un type de BI différent, notamment :

- Tableaux de bord intuitifs et personnalisables pour les dirigeants
- Outils de division/répartition ad hoc sophistiqués et réactifs pour les analystes
- Création de rapports distribués facilitant le partage des informations entre les équipes
- Outils de Data Mining et d'analyse prédictive pour les spécialistes des données
- Analyses faciles à intégrer dans les applications opérationnelles telles que CRM et Service

Dans un souci de compatibilité technologique et de relation fournisseur, il est logique de rechercher des fournisseurs de données et d'analytique capables de réunir la plupart, sinon la totalité, de ces fonctions en une seule et unique plateforme. Parallèlement, les fournisseurs doivent s'engager à fournir des capacités d'intégration Big Data permettant de s'adapter aux évolutions technologiques afin de réduire les efforts de reconfiguration d'un projet déployé et de le pérenniser. Dans un projet aussi sophistiqué, il est primordial de pouvoir suivre l'évolution des besoins des utilisateurs et des architectures de données.

Dans un souci de compatibilité technologique et de relation fournisseur, il est logique de rechercher des fournisseurs de données et d'analytique capables de réunir la plupart, sinon la totalité, de ces fonctions en une seule et unique plateforme.



# Monétiser mes données

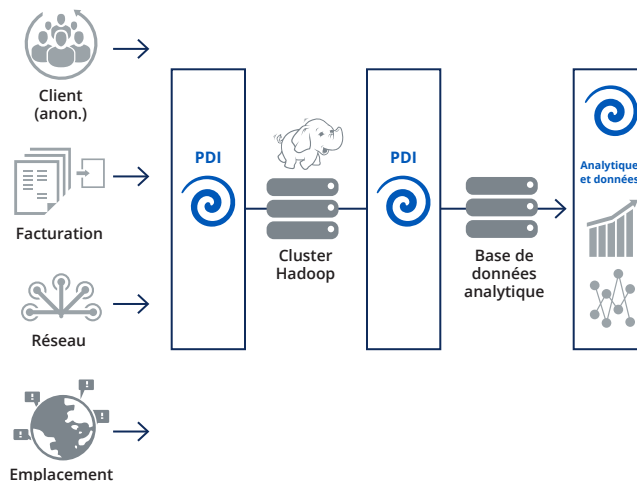
## DESCRIPTION ET AVANTAGES POUR LES ENTREPRISES

Grâce à leur capacité à traiter une grande variété de structures de données à moindre coût, les Big Data permettent non seulement aux entreprises d'optimiser leur cœur de métier, mais aussi de se créer de nouvelles sources de revenus. Le scénario Monétiser mes données s'inscrit dans cette logique en permettant aux entreprises de vendre leurs données.

À mesure que les entreprises collectent quotidiennement un volume de plus en plus important de données hétérogènes, la valeur que peuvent représenter ces données pour des tiers ne cesse d'augmenter. Dans ce scénario, les données sont classées, enrichies et dépersonnalisées (particuliers et entités sont rendus anonymes) avant d'être vendues, bien souvent à des acheteurs marketing externes. Par exemple, une société de télécommunications pourrait agréger les données de géolocalisation des appareils mobiles à différentes heures de la journée, les combiner aux données démographiques et vendre les jeux de données résultant à un détaillant qui s'en servirait pour déterminer quel est le meilleur emplacement pour implanter sa boutique. La société de télécommunications se crée ainsi une nouvelle source de revenus puisque ses données aident le détaillant traditionnel à mieux cibler son public.

## EXEMPLE

Dans l'exemple ci-dessous, la société de télécommunications s'appuie sur des données démographiques et de mobilité humaine pour offrir à des tiers un service analytique dédié, en tirant parti des visualisations géospaciales pour mieux cerner le potentiel d'achat. Cette société utilise aussi bien Hadoop qu'une base de données analytiques.



## PRINCIPAUX POINTS À PRENDRE EN COMPTE

Selon Gartner, 30 % des entreprises monétiseront directement leurs données d'ici 2015. Il s'agit là d'une véritable opportunité et seule une exploitation efficace des outils et approches Big Data permettra d'en libérer tout le potentiel. Dans le cadre de la monétisation des données, Hadoop, en tant que plateforme de traitement, réduit les coûts et augmente les marges par rapport aux anciennes solutions d'entreposage des données bien plus chères (prix réduit de 5 à 10 fois par To - Voir le paragraphe « Optimiser l'entrepôt de données »). Grâce à ses fonctions d'analytique et d'intégration de Big Data sans codage, Pentaho permet d'optimiser la rentabilité et le retour sur investissement. Par ailleurs, pour pouvoir proposer des services analytiques à des tiers, il peut s'avérer nécessaire pour les entreprises d'intégrer des outils de création de rapports et de visualisation à leur site Web. Avec son architecture ouverte et sa large offre de visualisations, Pentaho est un choix tout naturel pour une telle approche.

5 Gartner, „Gartner sagt voraus, dass die Datenbestände von Unternehmen im Jahr 2016 zu 30 Prozent monetarisiert sind“, 2013.



## En savoir plus sur Pentaho Business Analytics

[pentaho.fr/contact](http://pentaho.fr/contact)

+1 (866) 660-7555.

### Siège

Citadel International - Suite 340  
5950 Hazeltine National Dr.  
Orlando, FL 32822, USA  
tel +1 407 812 6736  
fax +1 407 517 4575

### Bureaux aux États-Unis et Dans le Monde

353 Sacramento Street, Suite 1500  
San Francisco, CA 94111, USA  
tel +1 415 525 5540  
toll free +1 866 660 7555

### Royaume-Uni, Reste de l'Europe, Moyen-Orient et Afrique

Londres, Royaume-Uni  
tel +44 7711 104854  
Numéro gratuit (Royaume-Uni) 0 800 680 0693

#### FRANCE

Bureaux - Paris, France  
tel +33 97 51 82 296  
Numéro gratuit (France) 0800 915343

#### ALLEMAGNE, AUTRICHE, SUISSE

Bureaux - Francfort, Allemagne  
tel +49(0)89 / 37 41 40 81  
Numéro gratuit (Allemagne) 0800 186 0332

#### BELGIQUE, PAYS-BAS, LUXEMBOURG

Bureaux - Anvers, Belgique  
tel +31 6 52 69 88 01  
Gebührenfrei (Belgique) 0800 773 83

#### ITALIE, ESPAGNE, PORTUGAL

Bureaux - Valence, Espagne  
numéro gratuit (Italie) 800-798-217  
numéro gratuit (Espagne) 900-868522  
numéro gratuit (Portugal) 800-180-060

Be social  
with Pentaho:



Copyright ©2015 Pentaho Corporation. Redistribution autorisée.  
Toutes les marques appartiennent à leurs propriétaires respectifs.  
Pour obtenir les dernières informations, consultez notre site Web à l'adresse [pentaho.com](http://pentaho.com).