DASSAULT SYSTEMES

EXALEAD

APPRIVOISER LE "BIG DATA" : GUIDE PRATIQUE

Opportunités, challenges et outils

3DS.COM/EXALEAD

"Donnez moi un point d'appui et un levier assez grand, et je soulèverai le monde" [1]

Archimède

À PROPOS DE L'AUTEUR

Diplômée de l'Université du Maryland et titulaire d'un PhD, Laura Wilber rejoint au début de sa carrière, la division d'ingénierie des systèmes fédéraux de Bell Atlantic (désormais Verizon) basée à Washington, DC. Au début des années 2000, elle crée puis dirige à San Diego AVENCOM, sa propre entreprise de logiciels spécialisés dans les bases de données en ligne. Après la cession de l'entreprise à Red Door Interactive en 2004, elle devient VP Marketing de Kintera, Inc., un fournisseur de logiciels SaaS destinés au secteur public. Laura travaille aujourd'hui en tant qu'analyste chez EXALEAD. Laura Wilber a plusieurs cordes à son arc, elle a mené des travaux de recherche autour de la propriété intellectuelle dans le cadre du Legal Arts Multimedia, LLC et elle a enseigné l'Ingénierie des Affaires, la Gestion des Systèmes d'Informations ainsi que le E-Commerce à l'ISG (Institut Supérieur de Gestion) basé à Paris. Aux côtés de son confrère d'EXALEAD, Gregory Grefenstette, elle a récemment co-rédigé *"Search-Based Applications : At the Confluence of Search and Database Technologies"*, publié en 2011 par Morgan & Claypool Publishers.

À PROPOS D'EXALEAD

Fondée en 2000 par deux ingénieurs pionniers dans le domaine des moteurs de recherche, EXALEAD® est un fournisseur leader de logiciels de recherche d'accès à l'information pour les entreprises et les organisations gouvernementales. EXALEAD compte parmi ses clients à travers le monde, de grandes entreprises comme PricewaterhouseCooper, Voyages-sncf.com, la Bibliothèque nationale de France, Orange et GEFCO ou encore la Banque Mondiale et Sanofi Aventis R&D. Plus de 100 millions d'utilisateurs uniques par mois se servent de la technologie d'EXALEAD® pour effectuer leurs recherches. Aujourd'hui, EXALEAD a pour ambition de redessiner le paysage du contenu numérique en proposant EXALEAD CloudView™. Cette plate-forme logicielle industrielle utilise le meilleur des technologies sémantiques pour accéder, structurer et donner du sens au patrimoine de données, structurées ou non structurées, qui constituent le "nuage informationnel" de l'entreprise. CloudView collecte des données provenant de pratiquement toutes les sources, dans n'importe quel format, afin de les transformer en briques d'informations métier, universelles, structurées et contextualisées, directement accessibles aux recherches ou pouvant servir de base à un nouveau type d'applications d'accès à l'information.

EXALEAD est une société du groupe Dassault Systèmes depuis juin 2010.

SYNTHÈSE

Comment définir le Big Data ?

Bien entendu, avec la virtualisation progressive des processus métiers de l'entreprise, sont apparus un peu partout des entrepôts de données aux tailles de plus en plus importantes, relevant du teractet d'abord puis du petaoctet pour atteindre des niveaux jamais égalés. Cependant, il n'existe pas de seuil de taille prédéfini qui permettrait de qualifier ce type d'entrepôt de "Big Data". La seule définition qui semble prévaloir est la suivante : un entrepôt de données est considéré comme étant du "Big Data" lorsqu'il prend une ampleur telle qu'il ne peut plus être géré (ou exploité) de manière efficace ou abordable par les outils de gestion traditionnels comme les systèmes classiques de gestion de bases de données relationnelles (SGBDR) ou les moteurs de recherche standards. Ce phénomène peut en réalité apparaître autant à une échelle de 1 Téraoctet qu'à une échelle de 1 Pétaoctet.

Une problématique ancienne mais de nouvelles opportunités

Le phénomène du Big Data est aujourd'hui plus que d'actualité. Il est cependant loin d'être un phénomène récent. Les professionnels de l'IT travaillant dans des domaines tels que la banque, les télécommunications et les sciences physiques ont cherché pendant des décennies à le surmonter.[2] Régulièrement confrontés à des bases de données dépassant la capacité des systèmes en place, leurs réactions, face à de telles situations furent certes pragmatiques mais parfois peu adéquates.

Par exemple :

- Pour pouvoir accéder à ces gros volumes de données, il fallait les compartimenter (en silos)
- Pour pouvoir les traiter, il fallait acquérir un superordinateur.
- Pour pouvoir les analyser, il fallait se rabattre sur un système d'échantillonnage
- Les stocker relevait de l'impossible. Il fallait se contenter de les utiliser, les purger avant de passer à autre chose et aller de l'avant

> *"Dans l'ère du Big Data, la quantité fait toute la différence."*[3]

Aujourd'hui, l'émergence de nouvelles technologies que nous décrirons dans ce livre blanc offre aux initiés du Big Data des solutions beaucoup plus avantageuses et permet aux organisations de toute taille d'accéder et d'exploiter le Big Data pour la toute première fois.

Parmi les données qui étaient jusqu'ici trop volumineuses, complexes et instables pour être utilisées, nous pouvons citer les relevés de compteurs et capteurs, les journaux d'événements, les pages Web, le contenu des réseaux sociaux, les messages électroniques et les fichiers multimédia. Le monde du Big Data s'ouvre progressivement à de nouveaux usages et modes de consommation personnelle et professionnelle de l'information. Il remet en question ce que nous pensions connaître de nous-mêmes, les organisations dans lesquelles nous évoluons, les marchés dans lesquels nous opérons – pour ne pas dire, tout l'univers numérique dans lequel nous vivons.

L'Internet : foyer de l'innovation du Big Data

La plupart de ces technologies innovantes sont nées sur le Web. Internet constitue en effet un terrain d'expérimentation rêvé : de larges volumes de données, des millions d'utilisateurs non captifs et donc de nombreuses contraintes quasi impossibles à surmonter ou tout du moins à concilier :

- Un trafic impossible à prévoir
- Une disponibilité de 99,999%
- Un temps de réponse quasi-instantané
- La gratuité des sessions
- La nécessité d'innover à un rythme soutenu

Pour faire face à ces contraintes, les spécialistes du Big Data ont développé des systèmes de gestion de données avec des capacités équivalentes à des superordinateurs, mais à un coût réduit, notamment en généralisant les achitectures distribuées sur de larges fermes de serveurs. Plus de performance donc mais aussi plus d'agilité grâce à des modèles de données dont la flexibilité dépasse ceux des SGBDR traditionnels. Les plus connues de ces technologies dérivées du Web, sont les bases de données non-relationnelles (autrement nommées "NoSQL" à savoir "Not-Only-SQL", SQL étant le terme commun donné à l'interrogation et la gestion des SGBDR), parmi lesquelles Hadoop (inspiré par Google et développé par Yahoo), Cassandra (Facebook) et les plates-formes de moteurs de recherche telles que CloudView (EXALEAD) et Nutch (Apache).

Parmi les autres approches émergentes, les systèmes "NewSQL" constituent une autre catégorie de solutions qui tente de répondre aux attentes liées au Big Data sans pour autant abandonner le principal modèle de bases de données relationnelles. [4] Afin d'optimiser la performance et l'agilité, ces systèmes emploient des stratégies inspirées de celles des vétérans de l'Internet (telles que la distribution à grande échelle, le traitement en mémoire vive ainsi que des modèles de données encore plus flexibles et directement inspirés des NoSQL). Ils peuvent également avoir recours à des stratégies SGBDR devenues plus courantes telles que les architectures en mémoire et les traitements analytiques de bases de données. Enfin depuis mi-2011, est apparue une nouvelle sous-catégorie

de solutions techniques qui va un cran plus loin en combinant les performances des systèmes RDBMS avec du NoSQL et/ou des plates-formes d'indexation pour proposer des applications décisionnelles nouvelle generation mêlant données structurées et non-structurées.

Le bon outil pour la bonne tâche

Ces différentes technologies, dans leur ensemble, peuvent répondre à pratiquement tous les besoins d'accès, d'analyse et de stockage des gros volumes de données. Il suffit simplement de savoir quelle technologie est la mieux adaptée au type de tâche entreprise et de comprendre les avantages et inconvénients que peuvent représenter certaines solutions (utilisabilité, maturité, coût, sécurité, etc.).

Des outils complémentaires et non-concurrents

Dans la plupart des situations, les technologies NoSQL, d'indexation et NewSQL jouent des rôles non-concurrents et complémentaires. Seule exception à la règle, l'analyse exploratoire où ces trois alternatives peuvent répondre mais de manière différente selon le contexte. Ainsi le recours à une plate-forme de recherche peut suffire 1) si les utilisateurs attendus du service sont des non spécialistes qui attendent de l'analytique très operationnelle, mixant structuré et non-structuré 2) ou encore s'il s'agit d'explorer des sources de données encore peu exploitées de type logs ou médias sociaux.

De même, pour l'analyse des données et le reporting opérationnel, il est possible d'utiliser une plate-forme NewSQL ou de recherche. Cependant la plate-forme de recherche suffit, là-encore, si l'application en question ne vise que des décideurs qui ont besoin d'une latence de mise à jour de données en secondes ou en minutes. La plate-forme NewSQL serait en revanche plus appropriée pour toute analyse à latence faible (inférieure à une milliseconde) et/ou pour le traitement d'événements complexes.

Alors que seule une plate-forme de recherche suffit dans certains cas – en allant parfois jusqu'à constituer la seule solution pour certaines activités telles que l'Analyse des Sentiments (Sentiment Analysis) – il est en réalité logique de déployer simultanément, en matière de Big Data, un moteur de recherche et un système NoSQL ou NewSQL. En effet il n'y a pas plus efficace que la technologie de recherche pour donner du sens au phénomène du Big Data et le rendre accessible à tout utilisateur.

C'est pourquoi nous avons décidé de rédiger ce document. Notre objectif est de mettre en évidence le rôle, souvent négligé et incompris, que jouent les technologies d'indexation dans les environnements touchés par le Big Data. Aussi, nous souhaitons fournir une vue d'ensemble concrète sur tous les outils disponibles pour relever les défis posés par ce nouveau et fascinant phénomène. Notre propre expérience avec nos clients et nos partenaires nous a fait réaliser que, malgré tout ce qui a déjà été exprimé au sujet du Big Data, cette notion reste encore très floue. Nous espérons que ce document saura mieux vous éclairer et vous aider à exploiter avec succès, votre propre Big Data.

TABLE DES MATIÈRES

1) PAR-DELÀ LA FRONTIÈRE ZETTA

Alimentés par une puissance informatique sans précédent, par des capteurs et compteurs omniprésents, des terminaux et services répondant aux attentes grandissantes des consommateurs, des stockages peu onéreux et une capacité de réseau ultra flexible, nous nous retrouvons aujourd'hui, nous et nos machines, à produire une quantité d'informations numériques à un rythme qui dépasse l'entendement.

Selon une étude menée par IDC, dans la seule année 2010, nous avons généré à l'échelle mondiale suffisamment d'informations numériques pour constituer l'équivalent de la capacité de stockage d'une pile de DVD qui mesurerait deux fois la distance entre la Terre et la Lune. Ceci représente 1,2 Zettaoctets, ou plus d'un trillion de Gigaoctets – soit une augmentation de 50% depuis 2009.[5] IDC prévoit de plus, qu'à partir de 2011, la somme des données mondialement produites doublera tous les deux ans.

Il est de ce fait loin d'être surprenant que les scientifiques aient choisi de créer le terme spécifique de "Big Data" pour exprimer cette échelle extraordinaire d'entrepôts de données désormais amassées au sein des organisations publiques et privées et sur le Web.

A. Qu'est ce que précisément le "Big Data" ?

Le Big Data est plus un concept qu'un terme précis. Certains emploient le qualificatif uniquement pour décrire les entrepôts de données à une échelle perçue en Pétaoctets (> à un million de Go). Pour d'autres, un entrepôt de Big Data ne couvre que quelques douzaines de Téraoctets de données. La plupart du temps cependant, le Big Data se définit plus en fonction de sa situation que de sa taille.

Mesurer le Big Data

Mémoire du disque*

1000 Gigaoctets (Go)	≈	1 Téraoctet (To)
1000 Téraoctets	≈	1 Pétaoctet (Po)
1000 Pétaoctets	≈	1 Exaoctet (Eo)
1000 Exaoctets	≈	1 Zettaoctet (Zo)
1000 Zettaoctets	≈	1 Yottaoctet (Yo)

* Pour le processeur ou le stockage virtuel, remplacer 1000 par 1024.

BIG DATA

Un entrepôt de données trop volumineux pour être géré de façon efficace et abordable par les technologies traditionnelles.

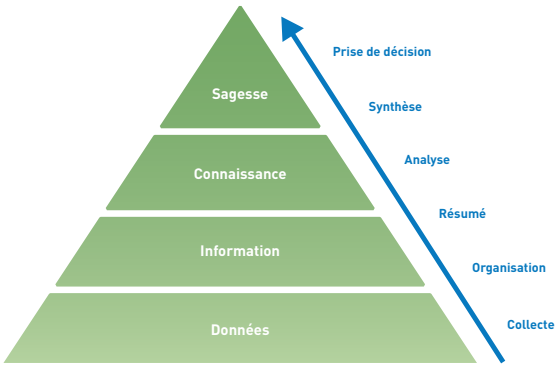
De manière spécifique, un entrepôt de données est considéré comme étant un "Big Data" lorsqu'il prend une ampleur telle qu'il ne peut plus être géré (ou exploité) de façon efficace ou abordable par les outils de gestion traditionnels comme les systèmes classiques de gestion de bases de données relationnels (SGBDR) ou les moteurs de recherche classiques. Ce phénomène peut apparaître autant à une échelle de 1 Téraoctet qu'à une échelle de 1 Pétaoctet.

B. Qui est touché par le Big Data ?

Le Big Data a déclenché un vif intérêt pour les organisations évoluant depuis un certain temps dans des secteurs précis tels que les sciences physiques (météorologie, physique), les sciences de la vie (la génomique, la recherche biomédicale), le secteur public (défense, trésor), la finance et la banque (traitement des transactions, analyse industrielle), la communication (archivage d'appels, données du trafic de réseaux) et, bien évidement, l'Internet (indexation de moteurs de recherche, réseaux sociaux).

Aujourd'hui, cependant, du fait de notre productivité numérique, les quantités volumineuses créées posent un réel problème pour les organisations, quelle que soit leur typologie. En 2008, la plupart des entreprises se retrouvaient déjà à devoir gérer une moyenne de 100 To et plus de contenu numérique.[6] Le Big Data est désormais également au centre des préoccupations des individus. Ces derniers prennent en effet conscience de la profondeur et l'ampleur du phénomène, notamment lorsqu'ils considèrent la notion d'information personnelle.

C. Le Big Data : Aubaine ou Fléau ?



L'objectif de la gestion classique de l'information : transformer les données brutes en savoir prêt à l'emploi. Dans l'ère du Big Data, le défi est de trouver des méthodes automatisées et industrielles qui permettent d'effectuer cette transformation..

Pour certains, le Big Data est un véritable fléau qui entraîne de nombreux problèmes tant au niveau du coût, de l'échelonnement et de la performance des systèmes d'informations que de la sécurité, la confidentialité et la

propriété des données. Cependant, le Big Data peut avoir le pouvoir d'entraîner de nouvelles découvertes et de faire avancer le monde de l'entreprise, la science, la médecine et le secteur public si nous parvenons à vaincre ces obstacles.

2) LES OPPORTUNITÉS DU BIG DATA

Beaucoup d'organisations mettent progressivement en évidence la possibilité de transformer ces quantités volumineuses de données brutes en véritable savoir exploitable pour le public ou les métiers. Il y a incontestablement beaucoup de valeur inexploitée dans la "zone grise" de données, qui constitue le Big Data. Cette zone grise regroupe les données non utilisées (ou sous-utilisées) :

1) Soit parce qu'elles sont trop volumineuses, non structurées et/ou brutes (c'est à dire structurées de façon minimale) pour pouvoir être exploitées par des systèmes d'informations traditionnels.

2) Soit parce qu'elles sont, dans le cas des données hautement structurées, trop coûteuses ou complexes à intégrer et exploiter (par exemple, en tentant d'agréger des données à partir d'une douzaine de bases de données différentes).

Ces mêmes organisations inventent de nouveaux modes d'analyse et d'exploration du Big Data qui exploitent les spécificités de tout type de données, structurées (telles que le contenu de bases de données), semi-structurées (tels que les fichiers d'ouverture de session ou fichiers XML) et le contenu non-structuré (tels que les fichiers textes ou les pages Web).

Parmi ces modes exploratoires agnostiques de la notion de source d'information et de volume :

- La navigation multi-dimensionnelle à grande échelle
- Le traitement du multimédia
- L'analyse de sentiments
- L'enrichissement automatique des bases de données
- L'analytique exploratoire
- L'analytique opérationnelle nouvelle génération

Nous allons maintenant nous intéresser de plus près à ces différentes propositions en présentant, pour chacune, des exemples de réalisation avec des moteurs de recherche, qui constituent une technologie souvent sous-estimée ou incomprise dans le contexte du Big Data. Nous passerons ensuite en revue l'ensemble des alternatives du marché, pour finir sur des illustrations concrètes de mise en oeuvre d'applications Big Data avec des technologies d'indexation.

A. La Navigation Multi-Dimensionnelle à Grande Échelle

La navigation multi-dimensionnelle permet d'affiner, à tout moment, une requête en sélectionnant (ou en excluant) des ensembles ou catégories de résultats. À la différence de la méthode de pagination traditionnelle qui consiste à présenter les données à travers de simples listes, la navigation multi-dimensionnelle (autrement nommée recherche paramétrique ou navigation à facettes) offre des moyens remarquablement efficaces de recherche et d'exploration de larges volumes de données. Elle le fait d'autant plus efficacement lorsque elle est combinée avec des outils tels que la suggestion automatique, le correcteur orthographique, et le "fuzzy matching" ou recherche floue (approximations phonétiques et/ou approximatives).

Jusqu'à récemment, la navigation multi-dimensionnelle ne pouvait se mettre en oeuvre que sur des petites et moyennes séries de données car elle reposait la plupart du temps, sur des processus manuels tels que la classification de données et les méta-tags (ou balises) de description. Aujourd'hui cependant, avec les technologies industrielles de Programmation Neuro Linguistique (PNL), la donne n'est plus du tout la même, car la classification et la catégorisation peuvent s'effectuer sur de très gros volumes de données au contenu non structuré et offrir ainsi une navigation multi-dimensionnelle à grande échelle.

EXEMPLE DE NAVIGATION MULTI-DIMENSIONNELLE

EXALEAD CloudView™ utilise des algorithmes industriels de traitement sémantique et statistique afin de catégoriser automatiquement les résultats de recherche sur un index de 16 milliards de pages Web (soit approximativement 6 Pétaoctets de données brutes).

Grâce aux facettes, les utilisateurs souffrent moins du volume et de la diversité des données disponibles et naviguent dans le Big Data du Web de manière beaucoup plus simple et naturelle.

LA PROGRAMMATION NEURO LINGUISTIQUE (PNL)

Intimement liée à l'intelligence artificielle, la PNL – autrement nommée linguistique computationnelle – utilise des outils tels que les algorithmes statistiques et l'apprentissage automatique afin de permettre aux ordinateurs de comprendre les différents aspects du langage humain (tels que les transcriptions de la parole, les fichiers texte et les SMS). Alors que le PNL se concentre sur les aspects structurels d'un énoncé, la linguistique va au-delà de toute structure en cherchant à identifier et mieux comprendre les significations et les relations.

EXALEAD CloudView™ constitue une illustraton parfaite de ce système industriel de facettes. Il en va de même pour les moteurs de recherche Internet tels que Google, Yahoo! et Bing qui s'y sont mis depuis quelques années. À des degrés variés d'automatisation et d'échelle, les moteurs de recherche en entreprise tels que Autonomy, Endeca, MS Fast et Lucene/Solr sont aussi des promoteurs de ce type d'expérience. La navigation multi-dimensionnelle constitue une véritable tendance pour l'exploration de contenus de type Big Data non-structuré.

B. Le Traitement du Multimédia

Parmi les nouveaux types de données générées par les utilisateurs, le contenu multimédia se trouve être le plus prolifique, avec des milliards de photos, fichiers audio et vidéo téléchargés au quotidien à la fois sur Internet et les serveurs d'entreprises. Exploiter ce type de contenus à une échelle de Big Data relevait jusqu'ici de l'impossible car pour y accéder et l'assimiler, nous nous basions uniquement sur les balises manuelles ou sur la simple association de métadonnées tels que les noms de fichiers.

Cependant, les récentes technologies telles que le traitement de la transcription phonétique automatique ou la reconnaissance d'objet (appelée Recherche d'Image par le Contenu, Computer-Based Image Retrieval ou CBIR en anglais) nous donnent désormais la possibilité de structurer entièrement ce type de contenu et d'offrir une nouvelle accessibilité à ces larges volumes de données multimédia. Cette tendance aura, pour sûr, une influence sans précédent dans les domaines de la santé, des médias, de l'édition, de la médecine légale ou encore celui de la gestion des actifs numériques.

Exemple de Recherche Multimédia:
France 24 est une chaîne d'actualité internationale diffusée, 24/24 et 7/7, en langue française, anglaise et arabe. En partenariat avec EXALEAD, Yacast Media et Vecsys, France 24 génère automatiquement des transcriptions en quasi temps réel de ses diffusions et utilise l'indexation sémantique de ses transcriptions afin d'offrir une recherche « plein texte » au sein de ses vidéos. La technologie de segmentation numérique permet de plus aux utilisateurs de trouver directement dans la diffusion, le moment précis où le terme est mentionné.

C. Le "Sentiment Analysis" ou "L'Analyse de Sentiments"

Le "Sentiment Analysis" ou "l'Analyse de Sentiments" repose sur des technologies sémantiques qui détectent, extraient et résument les émotions et les attitudes exprimées dans les contenus non-structurés. L'Analyse de Sentiments s'applique généralement à des contenus propriétaires, de type mails, enregistrements d'appels ou encore enquêtes de consommateurs/électeurs. Néanmoins, le Web est devenu en quelques années le premier et le plus grand recueil de Big Data au monde. Le Web est en effet le premier dépositaire du sentiment public, stockant de manière infinie toutes les émotions, avis, opinions du grand public sur les sujets de société, les personnes, les produits et les entreprises.

Le Web : Le plus large recueil de Big Data au monde.

Ce type d'analyse du Big Data a le vent en poupe depuis plusieurs mois car à l'heure de l'hégémonie du Web Social, il constitue une source d'information considérable dans des domaines aussi variés que la conception de produit ou la politique. Il permet en effet une efficacité et une réactivité sans précédent dans :

- La surveillance et la gestion de la perception par le public d'un problème, d'une marque, d'un organisme, etc. (connues sous le terme de "Management de la réputation")
- L'analyse des feedbacks des consommateurs sur un produit ou service, nouveau ou renouvelé
- L'anticipation ou la réaction face aux problèmes de qualité, de prix ou de conformité
- L'identification des opportunités de marché et des différentes tendances en matière de demandes des consommateurs

EXEMPLE D'ANALYSE DE SENTIMENTS

Un important constructeur d'automobiles fait appel à l'Analyse des Sentiments sur le Web afin d'améliorer la gestion de la qualité de ses produits. L'application utilise la plate-forme EXALEAD CloudView™ pour extraire, analyser et organiser des informations pertinentes liées aux problèmes de qualité, à partir de différentes sources telles que des forums de consommateurs d'automobiles. L'entreprise peut ainsi, en amont, détecter et réagir aux éventuels problèmes. Des processeurs sémantiques permettent de structurer automatiquement ces données en fonction du modèle, de la fabrication, de l'année, du type de symptôme et bien plus encore.

D. L'Enrichissement de Bases de Données

À partir du moment où une organisation collecte, analyse et organise un Big Data non-structuré, il est possible de l'utiliser afin d'améliorer et de contextualiser les ressources de données structurées existantes telles que les bases de données ou de connaissances. Par exemple, si l'on considère un système de Gestion de la Relation Client (GRC), il est possible d'extraire des informations à partir de sources ultra volumineuses telles que les logs de mails, de chats, de sites Internet et les réseaux sociaux pour pouvoir enrichir les profils clients. Ou encore, un catalogue numérique de produits peut, par ce biais, être élargi par un contenu Web complémentaire (descriptions de produits, photos, spécifications, informations données par le fournisseur). Enfin, un tel contenu peut également être employé afin d'améliorer la qualité des bases de données d'une organisation, en utilisant le Web pour vérifier les détails et compléter les caractéristiques manquantes.

EXEMPLE D'UN ENRICHISSEMENT DE BASE DE DONNÉES

La première agence de voyages en ligne et le premier site marchand français, Voyages-sncf. com, utilise des sources Web non-structurées (telles que les événements locaux, les attractions, les articles de tourisme et l'actualité) pour perfectionner le contenu de ses bases de données internes relatives au transport et à l'hébergement. Il en résulte un site complet d'organisation de voyages, dit moteur d'inspiration qui donne la possibilité à tout utilisateur de suivre de près chaque étape du cycle d'achat, autant d'opportunités supplémentaires d'accroître le nombre de ventes croisées et de faire de Voyages–sncf.com la plus grande référence française en matière d'organisation de voyages.

E. L'Analyse Exploratoire

L'analyse exploratoire a été pertinemment définie comme étant *"le processus d'analyse de données permettant de répondre aux questions qu'on ne sait pas poser "*. [7] Il s'agit d'un type d'analyse qui requiert une ouverture d'esprit prononcée. En pratique, il s'agit d'engager un véritable dialogue entre l'analyste et les données, pour réaliser des découvertes et trouver de nouveaux éclairages en allant simplement d'un point intéressant à un autre (d'où le terme d'analyse exploratoire autrement nommée "analyse itérative").

Ce type d'analyse est à l'extrême opposé de l'analyse traditionnelle nommée Traitement Analytique En Ligne ou OLAP (Online Analytical Processing en anglais). Le schéma classique de l'OLAP implique la volonté de proposer des réponses à des questions précises et préposées sur la base d'un univers de données connu et ordonné. Employé pour confirmer ou réfuter des hypothèses existantes, il peut parfois apparaître sous le terme d'Analyse Confirmatoire de Données ou CDA (Confirmatory Data Analysis en anglais).

> *"Les enjeux du Big Data sont énormes et peuvent engendrer de nouvelles vagues de croissance de la productivité, de développement de l'innovation et de la consommation – ce, à condition de mettre en place les politiques et les structures appropriées."*
>
> McKinsey Global Institute [8]

Découvrir les Significations Cachées et les Relations

Il est un fait certain que les larges volumes de données que nous amassons désormais, contiennent les réponses à des questions auxquelles nous n'avions pas pensé. Il est incroyable d'imaginer le nombre de révélations que recèlent les 100 Pétaoctets de données météorologiques logées au coeur du DKRZ (Centre de recherches météorologiques allemand). Il en va de même pour les 10 Pétaoctets de données produites tous les ans par l'accélérateur de particules Large Hadron Collider (LHC) ou encore les 200 Pétaoctets de données présentes à travers l'ensemble des 43 000 serveurs (bientôt 60 000) de Yahoo!.

Les choses peuvent aller encore plus loin si l'on commence à penser référencement croisé. Par exemple, un référencement croisé de données génomiques, démographiques, chimiques et biomédicales, pourrait participer à trouver un remède contre le cancer. Dans un tout autre domaine, le référencement croisé de données à grande échelle pourrait ouvrir la voie à une meilleure gestion des inventaires. L'entreprise américaine de grande distribution Walmart en est un bon exemple. Son choix de croiser les données météorologiques avec les données de consommation lui a permis de découvrir que les alertes cycloniques augmentaient non seulement les ventes de torches et de piles (ce qui n'est pas surprenant) mais aussi celles d'une marque particulière de gâteaux fruités (ce qui surprend), et que la bière constituait le produit "pré-cyclonique" le plus vendu (surprenant là encore).

En 2004, à l'approche de l'ouragan Frances, la direction de Walmart a pu réaliser un approvisionnement approprié des produits précédemment cités. Pour cela, elle a choisi d'extraire les données de consommation relatives aux quelques journées qui ont précédé l'ouragan Charley (intervenu peu de temps avant) à partir de son entrepôt de données (qui couvrait à l'époque 460 Téraoctets).[10] Certes, Walmart a découvert ce fait en faisant appel à de l'analyse traditionnelle et non à de l'analyse exploratoire, contrairement a ce qui est souvent dit, mais ce qui est intéressant, c'est que cet exemple nous pousse à imaginer ce qui pourrait se passer si nous donnions aux machines la liberté de découvrir, par elles-mêmes, de telles corrélations. Aujourd'hui, il est possible d'effectuer ce type d'analyse de deux façons, en mode Push ou Pull.

> *"Les techniques (d'analyse exploratoire) permettent d'explorer des milliers de bottes de foin plutôt que de se contenter de chercher une aiguille dans une seule."*
>
> PricewaterhouseCoopers 2010 [9]

Avec la stratégie "pull", nous donnons aux outils sémantiques en place la liberté d'identifier les relations, les schémas et les significations présentes dans les données. Nous pouvons ensuite recourir aux outils de visualisation, aux facettes (groupes et catégories dynamiques) et aux questions en langage naturel afin d'explorer ces connections de façon totalement ad hoc. Dans le cas de la stratégie "push", nous pouvons séquentiellement demander aux données de répondre à des questions précises ou leur ordonner d'effectuer certaines opérations (comme le tri) afin de voir ce qu'il en ressort.

Améliorer la Précision et la Rapidité des Prédictions

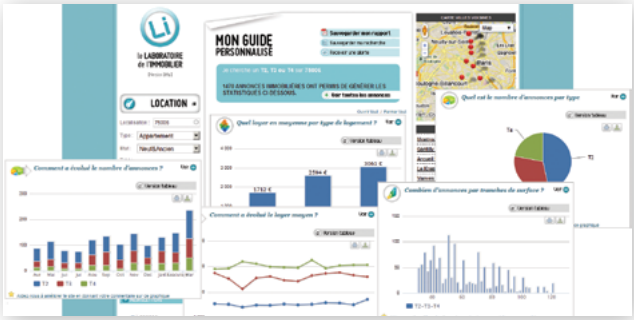
"Voyons ce qu'il en découle", tel est l'esprit de l'analyse exploratoire. Son but réside pratiquement toujours dans le fait de susciter des prédictions précises et concrètes. Dans les OLAP classiques, ces prédictions sont réalisées en appliquant des modèles statistiques complexes qui nettoient des séries de données échantillonnées par le biais d'un processus formel et scientifique d' "hypothétisation, de modélisation et de test ".

"Les décisions métiers se feront de plus en plus ou, du moins, seront de plus en plus supportées par des algorithmes informatiques plutôt que par des intuitions humaines." [11]

"La combinaison "modèles simples" et "gros volumes de données" est plus efficace que la combinaison "modèles élaborés" et "petite quantité de données." [12]

Alon Halevy, Peter Norvig & Fernando Pereira

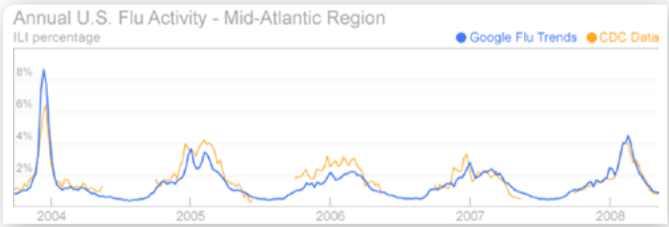
L'analyse exploratoire accélère ce processus formel en offrant un grand nombre de modèles "prêts à tester" qui auraient pu ne jamais être découverts. L'analyse traditionnelle des prédictions n'est pas, pour autant, menacée. Néanmoins, il est aujourd'hui avéré que le fait d'utiliser des algorithmes afin de remédier au désordre du Big Data peut produire des prédictions aussi précises que l'analyse complexe de séries de données échantillonnées ultra nettoyées et statistiquement structurées.



Les applications de décisionnel de type Search-Based, permettent de générer des tableaux de bords à partir d'informations totalement non strcuturées et notamment du Web, ici un observatoire de l'immobilier conçu à partir de dizaine de sites Web, labo-immo.org.

Par exemple, en analysant la fréquence de milliards de recherches Internet relatives aux symptômes de la grippe, Google a démontré qu'il était possible de prédire les épidémies de grippe avec une précision proche de celle des prédictions réalisées par les Centres de Contrôle et de Prévention des Maladies (CDC, Centers of Disease Control and Prevention) américains et basées sur une compilation de données provenant des cliniques et des praticiens. En effet, les personnes ont tendance à effectuer des recherches sur Internet avant de se rendre chez le médecin, l'analyse des requêtes sur Internet révèle ainsi les tendances en amont avant même la prise de rendez-vous. Autant dire que ce type d'information offre aux organismes de la Santé une marge de manoeuvre précieuse afin de se préparer aux épidémies.

Aujourd'hui, le CDC, ainsi que d'autres organismes de la Santé, tels que l'Organisation Mondiale de la Santé, exploitent les Tendances de Grippe Google comme un outil complémentaire de surveillance des maladies.



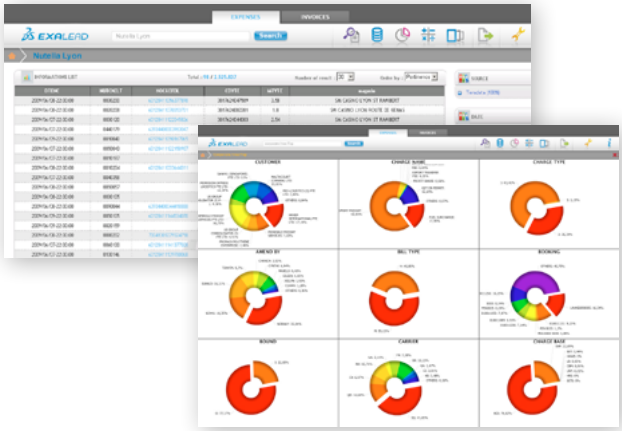
Les Tendances comme outil complémentaire de surveillance des maladies

Il est important de préciser que le CDC et les directeurs de cliniques sont plus qu'heureux de pouvoir obtenir de telles informations à travers Internet. Ils ne cherchent pas à savoir ni pourquoi ni comment les prédictions de ce type se rapprochent autant de celles présentées par le CDC. Tel est le potentiel de l'analyse exploratoire du Big Data : il suffit de tout échantillonner, de voir ce qui en découle et, en fonction de la situation, d'agir ou de présenter les informations à des spécialistes afin qu'il puissent les analyser ou les valider. [13]

EXEMPLE D'ANALYSE EXPLORATOIRE

L'une des plus grandes entreprises de distribution européenne utilise une SBA (Search-Based Application) EXALEAD pour permettre à des utilisateurs non spécialistes d'effectuer des interrogations en langage naturel sur les détails des reçus de caisse. Auparavant, ces mêmes reçus (stockés dans des entrepôts de données atteignant les 18 Téraoctets) ne pouvaient être consultés que par les utilisateurs du système décisionnel à travers des requêtes enregistrées ou personnalisées de manière complexe.

Une seconde SBA permet aux utilisateurs du département Marketing d'effectuer une analyse exploratoire reposant sur le référencement croisé à grande échelle des données "Tickets de caisse" et du "Programme de fidélité". Les utilisateurs peuvent ainsi effectuer la requête "nutella et paris" pour extraire une catégorie inédite de clients et inventer les opérations marketing les plus personnalisées.



F. L'Analyse Opérationnelle

Alors que l'analyse exploratoire est vue comme un sérieux atout pour planifier, l'analyse opérationnelle apparaît comme la méthode idéale pour agir. Le but de ce type d'analyse est de générer des indicateurs permettant une intelligence d'action en temps réel ou quasi-temps réel.

La tâche est cependant loin d'être facile car beaucoup de ces métriques sont enfouies dans des flux de données produits par des appareils en réseau tels que les compteurs "intelligents", les lecteurs RFID, les lecteurs de code-barres, les moniteurs d'activité de sites Internet et les unités de suivi GPS. Il s'agit là initialement de données de machines conçues pour être utilisées par d'autres machines et non par les individus.

Les organisations qui utilisent ce type d'infrastructure rencontrent des difficultés techniques et économiques pour rendre ces données accessibles aux individus. Avec l'émergence de nouvelles technologies, elles surmontent peu à peu ces obstacles et disposent maintenant d'outils de reporting et d'analyse de flux de données en temps réel (voir Chapitre 4) exploitables directement par des utilisateurs "humains ".

Prenons l'exemple du Docteur Carolyn McGregor de l'université d'Ontario. En poursuivant ses recherches au Canada, en Australie et en Chine, elle fait appel à l'analyse opérationnelle en temps réel du Big Data pour détecter en amont les infections potentielles chez les nourrissons prématurés. La plate-forme d'analyse surveille en temps réel les flux de données comme les lectures de la respiration, du rythme cardiaque et de la tension, saisies par des matériels médicaux (les électrocardiogrammes à eux seuls, sont capables de générer 1000 lectures par seconde).

Le système peut détecter les anomalies éventuellement annonciatrices d'une infection. Ce, bien avant l'apparition des symptômes et en devançant l'analyse de séries limitées de données effectuées sur papier et toutes les heures par un médecin.

EXEMPLE DE L'ANALYSE OPÉRATIONNELLE

L'un des plus grands fournisseurs privés d'électricité et des plus importants opérateurs d'énergie renouvelable au monde a déployé une Search-Based Application CloudView pour pouvoir mieux gérer sa production d'énergie éolienne. Plus précisément, il emploie CloudView pour automatiser les processus d'analyse de consommation et de prévisions en temps et en heure.

Cette SBA offre en effet une comparaison en quasi temps réel des données produites par des matériels de compteurs (Oracle) et les données de prévisions produites par un application MS SQL. Avant de faire appel à CloudView, les données étaient stockées séparément et donc devaient être comparées manuellement, ce qui avait pour conséquence un processus inefficace et sujet à l'erreur ainsi qu'une marge de manoeuvre limitée.

Ce type d'analyse prédictive améliore considérablement la capacité de l'entreprise à obtenir un parfait équilibre entre la production et la consommation minimisant ainsi les déficits ou les surplus onéreux. L'utilisation d'une SBA offre également la possibilité de puiser, de façon ad hoc, dans toutes les facettes de données par des affinages illimités par géolocalisation (pays, région, ville, etc.) et par période (heure, jour, semaine, mois, etc.).

Autre avantage non négligeable de ce type d'infrstructure reposant sur de l'indexation, le database offloading qui permet d'optimiser la réceptivité de l'ensemble des systèmes d'information en allégeant les systèmes Oracle et MS SQL déjà encombrés.

Le Dr McGregor précise d'ailleurs : *"Ces éléments ne peuvent être perçus à l'œil nu. Seul un ordinateur est capable de faire cela"*. [14]

3) L'INNOVATION RÉVOLUTIONNAIRE DE L'INTERNET

Les exemples du Chapitre 2 démontrent qu'il est possible de relever les défis techniques et financiers que posent le Big Data. La résolution de ces challenges repose sur l'exploitation de technologies qui ont été développées tout au long de ces quinze dernières années par les créateurs d'Internet. Parmi eux, on distingue :

- Les moteurs de recherche tels que EXALEAD, Google et Yahoo ! qui ont travaillé à rendre Internet (constituant le plus grand recueil de Big Data au monde) accessible à tous.
- Les sites de réseaux sociaux tels que LinkedIn et Facebook.
- Les géants de l'e-Commerce (comme Amazon).

Ces organisations, comme tant d'autres du même type, ont compris que les technologie traditionnelles de bases de données relationnelles engendraient trop de rigidité et/ou trop de coûts pour le traitement, l'accès et le stockage de données liées à l'univers ultra-mouvementé et volumineux de l'Internet.

> *"La fiabilité à grande échelle est un des plus grands challenges auquel nous faisons face chez Amazon.com... La moindre petite panne a des conséquences financières considérables et impacte la confiance de nos clients."* [15]
>
> Amazon

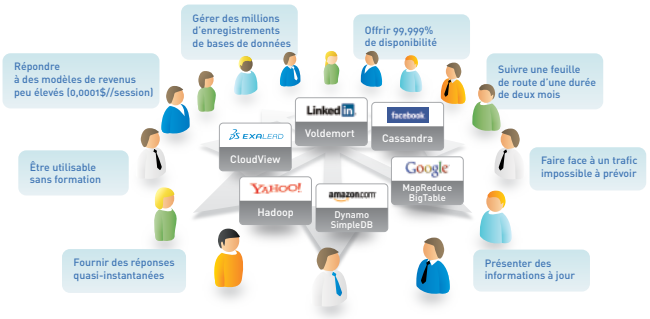
Les systèmes de Bases de Données Relationnelles (SGBDR) furent conçus à la base (il y a cinquante ans) pour enregistrer, de manière fiable et précise, les paiements, commandes et autres transactions effectuées dans le cadre d'entreprises "brick-and-mortar" (entreprises non-virtuelles). Pour protéger la précision et la sécurité de ce genre d'informations, ces systèmes faisaient en sorte que les données entrantes puissent intégrer les modèles et spécifications de données élaborés et précieusement structurés. Pour cela, ils construisaient des contraintes de protection de données, nommées ACID (Atomicité, Cohérence, Isolation et Durabilité des données).

Ces contraintes ACID ont fait preuve d'une grande efficacité lorsqu'il s'agissait de garantir la précision et la sécurité des données. Cependant, elles passent difficilement à l'échelle, ne permettent pas certains types d'interactions (comme celles nécessaires dans les réseaux sociaux ou dans l'analyse exploratoire), voire ne sont pas nécessaires. Parfois, il est plus important d'optimiser la disponibilité et la performance d'un système que de chercher à garantir la cohérence et l'intégrité des données.

EXEMPLES DE SGBDR LES PLUS UTILISÉS
Serveur MS SQL
MySQL
Postgre SQL
Oracle 11g
IBM DB2 & Informix

De ce fait, les géant du Web ont développé de nouveaux systèmes de gestion de données qui adoucissent les contraintes ACID et leur offrent les moyens d'accroître leur volume d'opérations de façon rentable tout en préservant une disponibilité et une performance optimales.

CONTRAINTES ACID
Atomicité
Cohérence
Isolation
Durabilité



L'Internet moteur d'innovation pour le traitement du Big Data

A. Les Architectures Distribuées et le Traitement Parallèle

L'innovation est en grande partie due à la généralisation du recours à la distribution en parallèle des tâches de traitement et d'accès, sur un ensemble (souvent géographiquement dispersé) de fermes de serveurs peu coûteux et en couplage faible (loosely-coupled).

Agissant en parallèle, ces fermes de serveurs d'entrée de gamme rivalisent avec les superordinateurs en fournissant de la puissance de calcul à un moindre coût et en garantissant une disponibilité continue de service en cas de panne.

Il s'agit là d'une architecture inspirée du multitraitement symétrique (SMP), du traitement parallèle massif (MPP) ainsi que des stratégies et technologies de grid informatique (voir les descriptions ci-dessous).

LES DIFFÉRENTS TYPES DE TRAITEMENT PARALLÈLE

En matière de traitement parallèle, les tâches de programmation sont divisées en sous-tâches et exécutées en parallèle à travers de multiples processeurs informatiques pour optimiser la performance. Le traitement parallèle peut s'effectuer dans un seul ordinateur à processeurs multiples ou à travers des milliers d'ordinateurs à processeur unique ou à processeurs multiples.

Le SMP est un traitement parallèle qui se fait à travers un petit nombre de processeurs à couplage serré (ex : la mémoire partagée, les bus de données, parfois les disques de stockage, les systèmes OS, etc.).

Le MPP est un traitement parallèle qui s'effectue sur un large nombre de processeurs à couplage faible (chaque noeud informatique possédant sa propre mémoire locale, son disque de stockage, sa copie OS, etc.). Il s'agit là d'une architecture qui oppose le "non-partage" à la "mémoire partagée" ou au "disque partagé". Les noeuds MPP communiquent généralement dans un réseau privé. Ils sont généralement constitués de machines homogènes qui se situent dans une seule et même localisation.

La Grid Computing utilise également des noeuds à couplage faible dans une structure non-partagée mais, contrairement aux SMP et aux MPP, un grid n'est pas conçu pour agir comme un ordinateur unique mais pour fonctionner de la même manière qu'un groupe d'individus qui collabore pour résoudre un problème unique (tel que la modélisation d'une protéine ou l'affinage d'un modèle climatique).

Les grids reposent sur des collaborations inter-organisationnelles qui mettent en commun des ressources pour créer une infrastructure informatique partagée. Elles sont généralement hétérogènes, dispersées et communiquent en utilisant des technologies WAN standards. Nous pouvons citer comme exemple de grid, les grids à la demande (ex : Amazon EC2), les grid peer-to-peer (ex : SETI@Home) et les grid de recherche (ex : DutchGrid). en utilisant des technologies WAN standards. Nous pouvons citer comme exemples de grilles, les grilles à la demande (ex : Amazon EC2), les grilles peer-to-peer (ex : SETI@Home) et les grilles de recherche (ex. DutchGrid). that pool resources to create a shared computing infrastructure. They are usually heterogeneous, widely dispersed, and communicate using standard WAN technologies. Examples include on-demand grids (e.g., Amazon EC2), peer-to-peer grids (e.g., SETI@Home), and research grids (e.g., DutchGrid).

B. Une Cohérence Assouplie et des Modèles Flexibles de Données

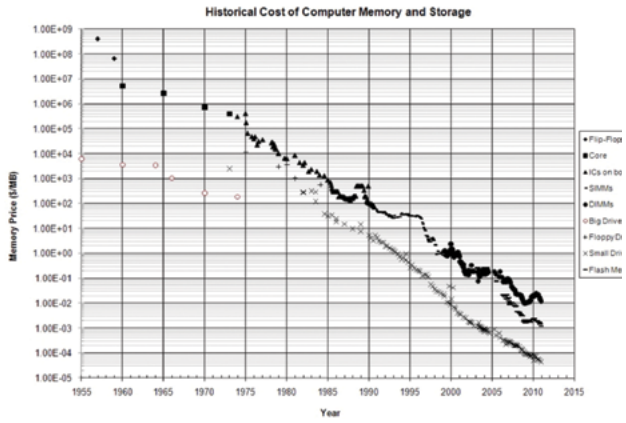
Ces innovateurs de l'Internet ont également réussi à garantir des niveaux de performance, de disponibilité et d'agilité plus importants, en concevant des systèmes capables d'assimiler et de traiter des données en perpétuelle évolution. Ces modèles flexibles, accompagnés de technologies sémantiques ont aussi joué un rôle majeur dans l'exploitation croissante des fameuses "zones grises" décrites plus haut (ces modèles sont analysés au Chapitre 4, Section B, Traitement & Interaction de données).

C. Le Traitement Cache & en Mémoire

Les systèmes les plus sophistiqués font maintenant appel au cache des données, ou aux stockage et traitement en mémoire. Dans les architectures en mémoire, les données sont stockées et traitées à grande vitesse – à savoir en mémoire vive – éliminant ainsi les aller-retours entre l'input et l'output qui peuvent diminuer la performance. Cette évolution repose en partie sur la chute drastique du coût de la mémoire vive (RAM) (voir le graphique ci-dessous) ainsi que la généralisation des architectures distribuées.

(Notons que si le prix de la mémoire vive a baissé considérablement, il est toujours moins cher d'acheter un ensemble d'ordinateurs en combinant leur mémoire vive pour qu'elles puissent atteindre 1 Téraoctet, que d'acheter un seul ordinateur avec une mémoire vive de la même capacité).

Graph of Memory Prices Decreasing with Time (1957-2010)



Même si peu d'organisations cherchent à traiter des données à l'échelle de l'Internet, ces innovations nées sur le Web ont donné naissance à des infrastructures logicielles propriétaires ou open-source accessibles aujourd'hui à toute personne qui souhaite relever les défis et saisir les opportunités offertes par le Big Data. Intéressons-nous, dès à présent, à cette boîte à outils.

4) LA BOÎTE À OUTILS DU BIG DATA

Certains organismes de recherche ont les moyens de s'équiper de superordinateurs pour répondre aux besoins du Big Data. Notre boîte à outils, en revanche, est composée d'outils accessibles aux organisations de toute taille et de tout genre. Ces outils se décomposent selon les catégories suivantes :

A. La Capture et le Prétraitement de Données

1. Les outils ETL (Extraire, Transformer, Charger)
2. Les connecteurs API
3. Les Crawlers
4. Les systèmes de messagerie

B. Le Traitement et l'Interaction des Données

1. Les systèmes NoSQL
2. Les systèmes NewSQL
3. Les Moteurs de Recherche

C. Les Outils Supplétifs

1. Les Services Cloud (ou en Nuage)
2. Les Outils de Visualisation

Chaque outil joue un rôle particulier dans la capture, le traitement et l'accès ou l'analyse du Big Data. Evaluons, tout d'abord, les outils de capture et de prétraitement.

A. La Capture et le Prétraitement des Données

1. LES OUTILS ETL

Fonctions Principales

- La consolidation des données (en particulier lorsqu'il s'agit de charger des bases de données)
- Le prétraitement/la normalisation des données

Définition

Les outils ETL (Extraire, Transformer, Charger) sont utilisés pour extraire et déplacer de larges volumes de données d'un système à un autre. Ils sont plus fréquemment utilisés comme des outils d'intégration de données. Plus précisément, ils sont employés pour consolider des données, issues de bases de données multiples, dans un nouvel entrepôt de données.Les plates-formes ETL comprennent généralement des mécanismes qui "normalisent" les données sources avant de les transférer. Pour ainsi dire, ces mécanismes réalisent le traitement minimal nécessaire pour aligner les modèles de données à l'entrée sur le modèle de la source de données cible. Ils permettent également de nettoyer les systèmes des données dupliquées ou qui présentent des anomalies.

Exemples

Ces solutions peuvent aller des plates-formes open source aux offres commerciales plus coûteuses. Certains outils ETL peuvent apparaître en tant que modules embarqués dans les systèmes décisionnels ou de bases de données. Les solutions commerciales sont celles qui sont les plus à même d'offrir les caractéristiques utiles et adaptées aux contextes Big Data. Parmi ces caractéristiques, nous pouvons citer le "pipelining" et le partage des données ainsi que la compatibilité avec les environnements de SMP, MPP et de grids. Parmi les outils d' ETL les plus connus :

- Ab Initio
- CloverETL (open source)
- IBM Infosphere DataStage
- Informatica PowerCenter
- Jasper ETL (open source – alimenté par Talend)
- Les services d'Integration de serveurs MS SQL
- Le Constructeur d'Entrepôts Oracle (implanté dans Oracle 11g) et l'Intégrateur de Données Oracle
- Le Talend Open Studio

Mises en garde

Dans les environnements de Big Data, le processus d'Extraction peut parfois constituer un poids insupportable pour les systèmes sources. L'étape de Transformation représente une menace si les données sont structurées de manière minimale ou sont extrêmement brutes. Le processus de Chargement peut être lent et ce, même si le code est optimisé pour pouvoir répondre à de larges volumes. C'est pourquoi les transferts ETL,

qui sont largement employés pour alimenter les entrepôts de données, ont tendance à être exécutés en dehors des heures de travail – la plupart du temps, durant la nuit. Il en résulte, dans certaines situations, une latence inadmissible dans la disponibilité des données. Il faut préciser cependant, qu'un grand nombre de fournisseurs d'outils ETL sont en train de développer – certains l'ont même déjà fait – des éditions spéciales capables de pallier ce type de limitation. Parmi ces éditions, nous pouvons citer Edition Real Time proposé par le PowerCenter d'Informatica (leur nouvelle version 9.1 est d'ailleurs conçue pour être spécifiquement adaptée aux environnements Big Data).

2. LES API

Fonctions principales

- L'échange et l'intégration de données.

Définition

Une API (Interface d'Application de Programmation) est une interface de logiciel-à-logiciel qui permet d'échanger tous les types de services ou de données imaginables. Nous nous concentrons plus ici, sur l'utilisation d'API en tant qu'outils d'échange et de consolidation de données. Dans ce contexte, une API peut offrir à un système hôte la possibilité de recevoir des données provenant d'autres systèmes (API "Push"), ou inversement, d'extraire des données du système hôte (API de publication "pull"). Les outils API emploient des langages de programmation et des protocoles standards qui facilitent les échanges (ex. : http/REST, Java, XML). Les packagings d'API spécifiques pour une source portent souvent le nom de "connecteurs" et peuvent être généralistes telle que l'API de Connectivité de Bases de Données Java (JDBC) qui traitent les SGBDR standards. Ils peuvent également être spécifiques (en fonction du vendeur ou de la plate-forme) comme le connecteur pour IBM Lotus Notes.

Exemples

Des API sont disponibles pour la majorté des grands sites Web, comme Amazon, Google (ex. : AdSense, Maps), Facebook, Flickr, Twitter et MySpace. Elles le sont aussi pour la plupart des applications métiers d'entreprises et des systèmes de gestion de données. Les moteurs de recherche d'entreprise offrent un package de connecteurs qui regroupent la plupart des types de fichiers et des systèmes d'entreprise communément utilisés (ex. : les dépôts XML, les fichiers de serveurs, les annuaires, les plates-formes de messagerie, les systèmes de gestion des contenus et des documents).

Mises en garde

Le recours à des API peut ralentir le processus de chargement de Big Data, notamment si le design choisi est imparfait ou si les ressources informatiques ou de réseaux sont inadaptées.

Mais ils ont, en général, fait preuve de flexibilité en ce qui concerne l'échange de grands volumes de données et de services. D'ailleurs, certains diront que la prolifération des API publiques et privées a grandement participé à la création de l'actuel monde du Big Data.

Cependant, le recours à un ETL garantit une performance beaucoup plus sûre que pour des APIs. Dans le cas des flux de données, l'architecture de messagerie apparaît elle aussi comme une solution performante et positive. (voir les systèmes de messagerie ci-dessous).

De plus, les outils API sont loin d'être idéaux pour collecter les données à partir d'Internet. Un crawler est un outil beaucoup plus adapté à cette tâche (voir Crawlers ci-dessous). Dans le contexte du Web, les outils API présentent trois principaux inconvénients :

- Malgré leur foisonnement, seul un pourcentage minime de sources de données en ligne est actuellement accessible par le biais d'une API.
- Les API offrent, habituellement, un accès à une portion limitée de données provenant d'un site Internet.
- L'utilisateur est libre de choisir ses formats et méthodes d'accès. Ceux-ci peuvent donc changer à tout moment. Du fait de ce climat d'incertitude et de variabilité, l'élaboration et l'entretien de liens API individuels sont parfois très consommateurs de temps. Cette situation peut devenir quasiment impossible à gérer dans l'univers du Big Data.

3. LES CRAWLERS

Fonction principale

- La collecte des données non-structurées (souvent du contenu Web) ou de petits paquets de données.

Définition

Un "crawler" ou robot d'indexation est un programme logiciel capable de se connecter à un système source, d'en extraire méthodiquement le contenu et les métadonnées avant de transmettre le contenu de cette extraction à un système hôte afin qu'il puisse l'indexer.

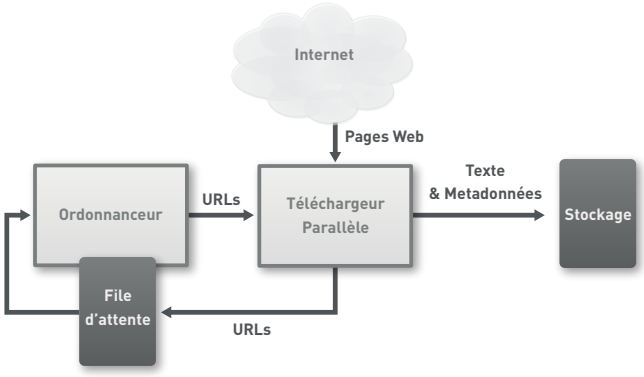
Parmi les différents types de crawlers, il existe le crawler de systèmes de fichiers. Celui-ci s'aventure à travers les annuaires, sous-annuaires et fichiers d'un ordinateur, pour recueillir les contenus de fichiers et les métadonnées (tels que le parcours, le nom, la taille et la date de dernière modification des fichiers). Les crawlers de systèmes de fichiers sont utilisés pour collecter le contenu non-structuré (comme les fichiers texte), le contenu semi-structuré (comme les logs) ainsi que le contenu structuré (comme les fichiers XML).

Le Web (HTTP/HTTPS) Crawler est un autre type de crawler qui est chargé, cette fois-ci, d'accéder à un site Web, de capturer et transmettre le contenu des pages ainsi que les métadonnées disponibles (titres des pages, légendes des contenus, etc.) avant de suivre les liens (ou une liste instaurée de liens visités) pour pouvoir ensuite passer à un autre site. Généralement, le traitement, le stockage et l'accès aux contenus capturés par les crawlers, s'effectuent au travers d'un moteur de recherche. Cependant, l'utilisation des crawlers peut être associée à d'autres types de systèmes de gestion de données (DMS ou Data Management System).

Exemples

Les crawlers de systèmes de fichiers sont généralement embarqués dans d'autres logiciels (tels que les moteurs de recherche, les systèmes opérationnels, les bases de données, etc.). Cependant, certains se présentent sous forme de programme séparé et unique : RiverGlass EssentielScanner, Sonar, Methabot (qui jouent aussi le rôle de Web Crawler).

Les crawlers Web sont, de même, la plupart du temps rattachés à des moteurs de recherche mais il existe également des Crawlers open source totalement indépendants. Les plus connus sont ceux exploités par les moteurs de recherche WWW publics.



Architecture type d'un Web Crawler standard. Source : Wikipedia

Parmi les exemples de Web crawler, on distingue :

- Bingbot
- crawler4j
- EXALEAD Crawler
- Googlebot
- Heritrix
- Nutch
- WebCrawler
- Yahoo! Slurp

Mises en garde

Si on a recours à un crawler, comme à tout outil de collecte de données, il est primordial de respecter ses spécificités et de ne pas charger démesurément les systèmes sources ou le crawler lui-même. La qualité du crawler détermine la qualité de la gestion des chargements.

Il est aussi important de préciser que les crawlers ne reconnaissent qu'un nombre limité de formats de fichiers (ex. : HTML, XML, texte, PDF, etc.). Pour pouvoir recueillir, avec un crawler, des fichiers aux formats incompatibles il faut convertir les données. Cette conversion s'effectue par le biais d'outils tels que les API (compatibles avec la plupart des moteurs de recherche commerciaux), les outils d'exportation des systèmes sources, les plates-formes ETL ou les systèmes de messagerie.

Il faut aussi bien tenir compte des problèmes rencontrés dans le Web crawling :

- Contenu Manquant

Un volume important de données Web existe sous des formes structurées, semi-structurées et non-structurées qui supposent un crawl particulièrement exigeant. Il existe peu de moteurs capables d'accéder à ces données et d'assimiler leur structure sémantique.

- Contenu de Faible Qualité

Les crawlers sont conçus pour couvrir un large champ et s'appuient sur la capacité des moteurs de recherche (ou des DMS) à filtrer les données. Cependant, le niveau de qualité peut être amélioré si le crawler est configuré dans le but de pouvoir effectuer un filtrage qualitatif préliminaire. Ce type de filtrage pourrait supposer une exclusion de certains types de documents, la réduction du traitement d'un site Web à une seule page pour éviter de noyer d'autres sources pertinentes, la détection et mise en place de règles qui s'appliqueraient au contenu dupliqué ou quasi-dupliqué, etc.

- Performance

La gestion du chargement est primordiale en matière de Web crawling. La performance du crawler peut être fragilisée si la largeur et la profondeur du crawl ne sont pas proportionnelles aux besoins et aux ressources de l'entreprise. Il est donc important de réguler, si possible, son action. Il peut aussi y avoir des problèmes liés à de mauvaises stratégies de mise à jour. Pour remédier à cela, il faut instaurer une stratégie de mise à jour de qualité qui implique de se concentrer sur le contenu pertinent (nouveau ou modifié) plutôt que de procéder à un re-crawling et à une ré-indexation de la totalité du contenu.

Indépendamment de la taille du crawl, il faut éviter avant tout de placer une charge trop excessive sur le site visité ou de violer les règles de propriété et de confidentialité de données.

Ces infractions sont généralement involontaires et dûes à la fragilité du crawler utilisé. Cependant, elles mènent aux blocage ou au "blacklistage" du crawler sur les sites Web publics. Dans les cas des crawls intranet, ce type de mauvaise gestion peut engendrer des problèmes de performance et de sécurité.

Il est possible d'éviter tous ces pièges en s'appuyant, dans le cas du Web public, sur un RSS "Really Simple Syndication" ou "Rich Site Summary" (des formats XML utilisés pour la syndication de contenu Web) qui livre uniquement les nouveaux contenus Web autorisés. Cependant, ces RSS ne sont disponibles que pour un nombre limité de sites et certains peuvent être incomplets ou dépassés.

MIEUX COMPRENDRE LE WEB

Parfois, un moteur de recherche peut percevoir le contenu HTML comme un arbre dont les branches seraient les tags HTML et les noeuds seraient le texte. Soit il utilise des règles écrites en langage de requête XML standard (XPath) pour extraire et structurer le contenu. Avec cette stratégie, le crawler joue un rôle prédominant dans le cadre du prétraitement du contenu. Soit, il collecte le HTML en texte brut et s'appuie sur ses modules sémantiques pour donner une structure au contenu.

La première approche peut donner de très bons résultats, mais elle implique l'élaboration et la surveillance laborieuses de règles spécifiques appliquées à chaque source (dans l'univers en perpétuel changement du Web, une règle XPath a une durée de vie de seulement 3 mois). La deuxième approche peut être appliquée à tous les sites mais elle est complexe et sujette à l'erreur. Cependant, il existe une stratégie idéale qui vient équilibrer ces deux approches en exploitant les schémas de structures existantes et en se basant sur la sémantique pour vérifier et améliorer ces schémas.

4. LES SYSTÈMES DE MESSAGERIE

Fonctions principales

- L'échange de données (événements et paquets de données)
- L'intégration d'applications/de systèmes
- Le prétraitement et la normalisation de données (rôle secondaire)

Définition

Les systèmes Message-Oriented Middleware (MOM) constituent un pilier efficace pour l'intégration d'applications d'entreprises. Souvent déployées dans le cadre d'une architecture orientée services (SOA), les solutions MOM relient les systèmes et les applications à travers une passerelle de type bus de messages. Les messages (petits paquets de données) gérés par ce bus peuvent être configurés pour la livraison d'un point à un autre (système de fil d'attente de messages) ou diffusés à de multiples destinataires (messagerie publication/souscription). Ils soutiennent, à différents degrés, la sécurité, la pérennité et l'intégrité des messages.

Les échanges entre les systèmes dispersés sont possibles car tous les systèmes connectés ("peers") partagent un schéma de messages, une série de commandes et une infrastructure communs. Les données issues des systèmes sources sont transformées à un degré suffisant pour permettre aux autres systèmes de les consommer. Par exemple, les valeurs binaires peuvent avoir besoin d'être converties en leurs équivalents textuels (ASCII) ou encore les adresses de session ID et IP peuvent être extraites des fichiers log et encodées en registres XML. Les API qui gèrent ces traitements de données peuvent être embarquées dans des systèmes individuels connectés au bus, ou implantées dans une plate-forme MOM.

Le Traitement d'Evénements Complexes (CEP)

Les systèmes MOM sont souvent employés pour gérer l'échange asynchrone d'événements et de paquets de données (tels que les scans de code-barres, les cours d'action, les données météorologiques, les logs de session et les lectures de compteurs) entre différents systèmes. Dans certaines situations, un moteur CEP peut être déployé pour analyser les données en temps réel. Pour cela, il doit permettre la détection de tendances complexes, le filtrage par motifs et règles et la modélisation des flux de données. Par exemple, un moteur CEP peut appliquer des algorithmes complexes aux flux de données tels que les retraits des guichets automatiques et les achats par cartes de crédit afin de détecter et signaler toute activité suspecte en temps réel ou quasi temps réel. Si un CEP propose un traitement historique, les données doivent être capturées et stockées dans un DMS.

Exemples

Les plates-formes MOM sont des applications indépendantes, ou font partie de SOA plus larges. En voici quelques exemples :

- Apache ActiveMQ
- Oracle/BEA MessageQ
- IBM WebSphere MQ Series
- Informatica Ultra Messaging

- Microsoft Message Queuing (MSMQ)
- Solace Messaging & Content Routers
- SonicMQ from Progress Software
- Sun Open Message Queue (OpenMQ)
- Tervela Data Fabric HW & SW Appliances
- TIBCO Enterprise Message Service & Messaging Appliance

La plupart des organisations citées ci-dessus proposent également un moteur CEP. Il existe aussi des fournisseurs CEP spécialisés comme SteamBase Systems, Aleri-Coral8 (faisant désormais partie du groupe Sybase), UC4 Software et EsperTech. En plus de ceci, de nombreuses plates-formes NewSQL (évoquées dans la prochaine section) intégrent la technologie CEP, ce qui compromet d'ailleurs l'avenir de l'indépendance de la technologie CEP.

Mises en garde

Les systèmes de messagerie étaient spécifiquement conçus pour répondre aux besoins des secteurs tels que la finance, la banque et les télécommunications face aux entrepôts de données ultra-volumineux et ultra-volatiles. Les volumes du Big Data peuvent cependant surcharger certains systèmes MOM, notamment si le MOM exécute le traitement des données (filtrage, agrégation, transformation, etc.) au niveau même du bus de messages. Dans de telles situations, la performance peut être accrue en déchargeant le plus possible les tâches de traitement vers les systèmes sources ou de destination. Il est aussi possible d'acquérir des solutions de type messaging encore plus performantes telle que IBM WebSphere MQ Low Latency Messaging ou Informatica Ultra Messaging ou encore les applications intégrées de messagerie Solace, Tervela ou TIBCO (cette dernière application ayant été développée en partenariat avec Solace).

B. Le Traitement et l'Intéraction des Données

Aujourd'hui, les SGBDR classiques sont complétées par une série abondante de différents DMS spécifiquement conçus pour supporter le volume et la variabilité du Big Data. Ces DMS incluent les NoSQL, les NewSQL et les systèmes de recherche. Tous sont capables d'assimiler des données fournies par les outils de capture et de prétraitement évoqués dans les sections ci-dessus (ETL, API, crawlers et systèmes de messagerie).

- Les NoSQL

Les systèmes NoSQL sont des bases de données distribuées et non relationnelles conçues pour faire du stockage et du data crunching à très grande échelle en s'appuyant sur des architectures distribuées sur un très large nombre de serveurs. Ils permettent d'effectuer plusieurs types de traitement, incluant l'analyse exploratoire et prévisionnelle, la transformation du type ETL des données et les OLTP (Traitement Transactionnel en Ligne). Leurs principaux inconvénients sont leur caractère encore insolite et, en ce qui concerne les plus récentes de ces solutions largement open-source, leur instabilité.

- Les NewSQL

Les NewSQL sont des bases de données relationnelles conçues pour répondre aux contraintes ACID, tout en offrant des capacités d'OLP temps réel, et d'OLAP à l'échelle du Big Data. Ces systèmes surpassent largement les performances des RDBMS traditionnels en s'appuyant sur les mêmes caractéristiques que le NoSQL, comme le stockage de données en colonnes et les architectures distribuées, et en ayant recours à des technologies telles que le traitement en-mémoire (in-memory, le SMP ou MPP). Leur principal inconvénient réside dans leur coût.

- Les plates-formes de Recherche

Les plates-formes de recherche aptes à faire face au Big Data, trouvent leur origine dans Internet tout comme leurs équivalents NoSQL (architectures distribuées, modèles de données flexibles, "caching"). Elles emploient donc les mêmes stratégies et technologies. Certains disent d'ailleurs qu'elles sont elles-mêmes des solutions NoSQL. Cependant, si nous les plaçons dans cette même catégorie, nous courons le risque de passer à coté de leur principal point de différenciation : le traitement en langage naturel (NLP). C'est en effet la technologie NLP qui donne aux plates-formes de recherche le pouvoir de collecter, analyser, classifier et corréler les différents sources de données structurées, non structurées ou semi structurées.

Les NLP et les technologies sémantiques offrent également aux moteurs de recherche des capacités inédites : l'analyse de sentiments, l'apprentissage automatique, l'analyse de texte, etc. Les plates-formes de recherche apparaissent comme un complément aux systèmes NoSQL et NewSQL, offrant à tout type d'utilisateur un moyen simple et familier d'analyser ou d'explorer leur Big Data. Dans certaines situations, les Search-Based Applications (SBA) offrent même une alternative plus facile et abordable aux systèmes NoSQL et NewSQL en matière de reporting et d'analyse.

Comme précisé dans la synthèse, le challenge avec ces technologies est de déterminer laquelle est la mieux adaptée au type de tâche effectuée et de comprendre les avantages et les inconvénients de ces solutions (utilisabilité, maturité, coût, sécurité, aptitudes techniques, etc.). Le tableau suivant explique comment choisir le bon outil pour la bonne tâche.
Note : ce tableau exprime une vue générale des meilleures utilisations et non des seules utilisations.

Pour les deux catégories dans lesquelles sont proposées plusieurs options – à savoir l'analyse exploratoire et l'analyse opérationnelle – le choix entre NoSQL, Recherche ou NewSQL se fait selon l'utilisateur cible, à savoir 1) une machine ou 2) un individu.

S'il s'agit d'un individu, le choix dépend également du type d'individu : un utilisateur standard ou un expert en analyse, un statisticien ou un programmeur.

Le deuxième critère de choix repose sur le besoin de latence dans la mise à disposition de l'information, temps réel, quasi temps réel ou en décalé.

Pour mieux comprendre, analysons maintenant ces trois différents types de DMS (Data Management Systems).

Tâche relative au Big Data	Outils relatifs au Big Data (DMS)		
	NoSQL	Recherche	NewSQL
Mémoire			
Données structurées			X
Données non-structurées, semi structurées et structurées en petits paquets	X		
Traitement			
Transformation/Crunching (analyse) basique des données	X		
Langage Naturel / Traitement Sémantique / Analyse de Sentiments		X	
Traitement des Transactions (ACID OLTP et Traitement du Flux des Evènements)			X
Accès et Interaction			
Extraction des Informations de Machine à Machine (IR)	X		
Exploration d'Individu à Machine		X	
Développement Agile d'applications métiers		X	
Analyses			
Analyse Conventionnelle (OLAP)			X
Analyse Exploratoire	X	X	X
Analyse/Reporting opérationnels		X	X

1. Les Systèmes NoSQL

Fonctions Principales

- Traitement de données à grande échelle (traitement parallèle des systèmes distribués)
- Recherche embarquée (recherche et extraction basiques d'informations de machine à machine)
- Analyse exploratoire de données semi structurées (niveau d'expert)
- Mise en mémoire ou stockage de données (non structurées ou structurées en petits paquets)

Définition

Les NoSQL ("Not Only SQL" voulant dire "Pas seulement SQL") constituent une catégorie encore atypique de systèmes de gestion de bases de données (ex. : Hadoop, Cassandra et BerkleyDB). Parmi leurs caractéristiques principales : le recours à des architectures distribuées avec un traitement parallèle effectué sur un grand nombre de serveurs, la flexibilité du modèle de données qui tient compte de la variabilité des sources et de l'utilisation de caches et/ou de stratégies en-mémoire pour renforcer le niveau de performance. Ils utilisent également des langages et mécanismes non SQL pour interagir avec les données (même si certains présentent parfois des API capables de convertir les requêtes SQL au langage de requête de base du système).

Par conséquent, ils offrent une possibilité de stockage à faible coût et à grande échelle de larges volumes de données comme des paquets de données historisées tels que les logs d'enregistrements de données d'appel, de lectures de compteurs et les clichés de téléscripteur (c'est à dire un stockage "par gros morceaux"). Cette même possibilité de stockage s'applique aux données semi-structurées ou non-structurées (archives de mails, fichiers XML, documents, etc.). Leur architecture distribuée fait d'eux la solution idéale pour le traitement de gros lots de données (regroupement, filtrage, tri, analyse algorithmique - statistique et programmatique - etc.).

Ils sont efficaces également pour l'extraction et l'échange de données de machine à machine ainsi que pour le traitement de larges volumes de transactions à condition que les contraintes ACID soient allégées ou renforcées au niveau applicatif, plutôt qu'au coeur du DMS.

Ces systèmes sont également capables d'analyse exploratoire de données semistructurées ou hybrides. Cependant, les analystes qui y ont recours doivent ici être des statisticiens expérimentés, voire travaillant en tandem avec des programmeurs aguerris.

En ce qui concerne les systèmes NoSQL, si l'on souhaite effectuer des recherches plein texte, des requêtes ad hoc, un ajout d'applications métiers ou, tout simplement si l'on veut rendre accessibles les données à des utilisateurs non experts, il est nécessaire d'utiliser un moteur de recherche en plus.

Les DMS NoSQL se présentent sous quatre différentes formes, chacune étant adaptée à un type de tâche: [16]

- "Key-Value stores" ou Table de trancription Clés/Valeurs
- Bases de données orientées Document
- Bases de données orientées Colonnes (Wide Colomn ou Family Colomn)
- Bases de données orientées Graphe

LES CLÉS ET VALEURS DE REGISTRES

Typiquement, ces DMS stockent des éléments tels que les identificateurs alphanumériques ("clés") et les valeurs associées dans des tableaux indépendants (appelés "table de transcription"). Ces valeurs sont soit de simples chaînes de texte soit des listes et séries plus complexes. Le recherche de données ne peut se faire qu'au niveau des "clés", non des "valeurs", et est limitée à des correspondances exactes.

Fonctions Principales

La simplicité des Key Value Stores fait d'elles la solution idéale en matière d'extraction rapide, efficace et à grande échelle de données : gestion de profils ou de sessions d'utilisateurs ou encore extraction de noms de produits. C'est pourquoi Amazon n'a de cesse de faire appel à son propre système de Key Value Store, nommé Dynamo, pour gérer ses paniers de shopping.

Key Value Store

Clé	Valeur
Prod_123	Zapito Scooter
Prod_124	Walla Scooter
Prod_125	Super Cruiser

Pour réaliser une extraction rapide, la plupart des Key Value Stores associent les simples clés de chaînes aux valeurs des chaînes.

Exemples de Clés et valeurs de Registres

- Dynamo (Amazon)
- Voldemort (LinkedIn)
- Redis
- BerkeleyDB
- Riak
- MemcacheDB

Base de Données Orientées Document

Clé	Valeur
Prod_123	Type: Scooter, Name: Zapito Scooter, Price: 1000, Color: Silver
Prod_124	Type: Scooter, Name: Walla Scooter, Color: Blue
Prod_125	Type: Scooter, Name: Super Cruiser , Price: 2500

Les bases de données orientées Documents contiennent des valeurs semi structurées qui peuvent être interrogées. Le nombre et le type d'attributs par rangée peuvent varier, offrant ainsi une flexibilité plus grande que celle rencontrée dans les modèles de données relationnelles.

BASES DE DONNÉES ORIENTÉES DOCUMENTS

Inspirées de Lotus Notes, les bases de données orientées Document étaient, à la base, conçues pour gérer et stocker des documents. Ceux-ci sont encodés dans des formats standards d'échange de données tels que XML, JSON (Javascript Option Notation) ou BSON (Binary JSON). Contrairement aux simples Key Value Stores décrits ci-dessus, les colonnes de valeurs des bases de données de documents contiennent des données semi-structurées – plus spécifiquement, des attributs de paires nom/valeur. Une seule colonne peut regrouper des centaines de ces attributs. Le nombre et le type d'attributs enregistrés peuvent variés d'une rangée à l'autre. Dans les bases de données orientées Document, les clés et les valeurs sont cherchables, ce qui les distingue là encore des Key Value Stores.

Fonctions Principales

Les bases de données orientées Document sont appropriées pour stocker et gérer les entrepôts ultra volumineux de documents textuels tels que les documents textes, les messages mails, les fichiers XML ainsi que les documents conceptuels (comme les représentations dénormalisées ou agrégées d'une entité de bases de données tel qu'un produit ou un consommateur). Elles peuvent également être efficaces pour stocker des données éparses ou irrégulières (semi-structurées) (qui requièrent généralement l'usage de 0 dans un SGBDR).

Exemples de Bases de Données Orientées Documents

- CouchDB (JSON)
- MongoDB (BSON)
- MarkLogic (bases de données XML)
- Berkeley DB XML (bases de données XML)

LES BASES DE DONNÉES ORIENTÉES COLONNES OU (WIDE COLOMN OU FAMILY COLOMN)

Tout comme les bases de données orientées Document, les bases de données orientées Colonnes (WC pour "Wide-Column"/CF pour "Column Family") reposent sur une structure de données distribuée et orientée colonnes qui contient de multiples attributs par clé. Alors que certaines WC/CF possèdent un ADN de type Key Value Store (comme Cassandra inspirée par Dynamo), la plupart repose plutôt sur le modèle du "Google Bigtable", à savoir le système de stockage de données distribuées que Google a développé pour son index de recherche, puis pour Google Earth et Google Finance.

Ces bases de données reprennent non seulement la structure de données de la Bigtable de Google mais aussi son système de fichiers distribué (GFS) et son système de traitement parallèle nommé MapReduce (c'est le cas avec Hadoop qui comprend le Système de fichiers Hadoop – "HDFS" basé sur son GFS, le Hbase (un système de stockage type Bigtable) et le MapReduce.

Fonctions Principales

Ce type de DMS est très efficace pour :

- Le stockage des données distribuées, notamment celles converties aux fonctions d'horodatage des WC/CF
- Le traitement à grande échelle et par lots de données : tri, analyse, conversion (conversion entre valeurs hexadécimales, binaires et décimales), crunching algorithmique, etc.
- les analyses exploratoire et prévisionnelle effectuées par des statisticiens et programmateurs expérimentés

Notons que si vous utilisez un cadre de type MapReduce, il est important de savoir que MapReduce est une méthode de traitement par lots. C'est pourquoi Google a depuis peu réduit son rôle en renforcant l'indexation par flux et en temps réel. C'est ce que propose Caffeine, sa toute nouvelle infrastructure.

Dans les tableaux SGBDR orientés colonnes, chaque attribut est enregistré dans une colonne séparée et chaque rangée – ainsi que chaque colonne figurant dans cette même rangée – doit être séquentiellement lue pour extraire des informations. Cette méthode est plus lente que celle du modèle NoSQL orienté colonnes avec lequel un grand montant d'informations peut être extrait, en une seule action de "lecture", d'une large colonne unique.

Exemples de Bases de Données Orientées Colonne

- Bigtable (Google)
- Hypertable
- Cassandra (Facebook, Digg, Twitter)
- SimpleDB (Amazon)
- Hadoop (Yahoo)
- Cloudera

LES BASES DE DONNÉES ORIENTÉES GRAPHE

Les bases de données orientées Graphe remplacent les tableaux relationnels par des tableaux relationnels Clé/Valeur interconnectées. Elles se rapprochent des bases de données orientées objet car les graphes se présentent sous forme de réseau orienté objet de noeuds (objet conceptuel), de relations entre noeuds ("edges") et de propriétés (attributs d'objet exprimés en paires clé/valeur). Parmi les quatre types de NoSQL évoqués ici, elles sont les seules qui s'intéressent aux relations et leur intérêt pour la représentation visuelle d'information les rend plus accessibles aux individus que les autres DMS NoSQL.

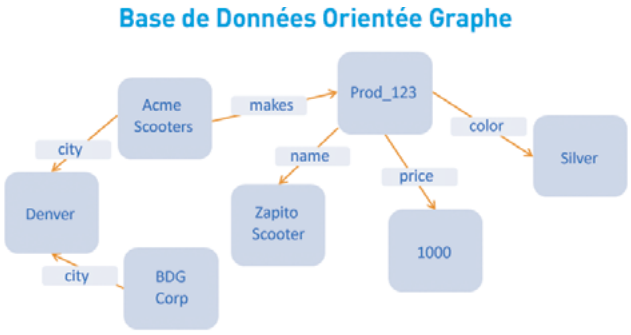
Fonctions Principales

En général, les bases de données orientées Graphe sont d'une grande utilité pour ceux qui s'intéressent plus aux relations entre les données qu'aux données en elles-mêmes. Par exemple, ceux qui souhaitent générer des recommandations (ex. : propositions de montée en gamme ou de ventes croisées) ou de mener des investigations (ex. : détection de modèles).

Notons que ces DMS sont optimisés pour "traverser" les données et non les interroger. Si vous souhaitez explorer les relations tout en interrogeant et en analysant les valeurs qu'elles regroupent (et/ou pour pouvoir utiliser les requêtes en langage naturel pour analyser ces mêmes relations), il vaut mieux choisir un DMS de recherche ("search-based").

Exemples de Bases de Données orientées Graphe

- Neo4j
- InfoGrid
- Sones GraphDB
- AllegroGraph
- InfiniteGraph



Les bases de données orientées Graphe s'attachent plus aux relations entre les entités de données qu'aux entités en elles-mêmes.

MISES EN GARDE

Ces systèmes offrent des solutions abordables et à grande échelle pour répondre aux besoins de stockage de données ultra volumineuses, de traitement et d'analyse. Cependant, dans l'évaluation de ces solutions NoSQL, il est important de prendre en compte les inconvénients fréquents suivants :

- **Un manque de maturité**

  Beaucoup d'entre elles sont des solutions open source avec un niveau normal de volatilité inhérent aux méthodes de développement. De plus, leur degré de support, de standardisation et de packaging varie grandement. De ce fait, les services professionnels qui doivent s'y rattacher sont onéreux.

- **Un manque d'expertise**

  Seul un petit nombre d'ingénieurs est capable de mettre en place et de gérer ces systèmes. De même, il existe peu de constructeurs et d'utilisateurs à même de maîtriser le langage de requêtes et les outils qui s'y rattachent : fonctions MapReduce (dans Erlang, Java, JavaScript, etc.), HQL, Lua, JRuby, SparQL, XQuery, LINQ, JSON/BSON, etc. Si disponible, il vaut mieux choisir une version d'entreprise ou commerciale du système – qui contient tous les outils de gestion et/ou la passerelle SQL – afin de réduire la complexité des conditions de fonctionnement.

- **Inaccessibilité**
  Les systèmes NoSQL ne prévoient pas l'indexation plein texte (et par conséquent la recherche plein texte) et la plupart ne fournissent pas la catégorisation et le groupage automatique. Seule l'installation d'un moteur de recherche à part saura satisfaire ces fonctions.
- **Sécurité**
  En matière de droit d'accès, beaucoup d'entre elles présentent une sécurité faible ou quasi-nulle. Seule une couche d'application externe permet de renforcer cette sécurité.

Pour ce qui est de la sécurité physique ou concrète, beaucoup de systèmes compromettent le degré de recouvrement des données pour pouvoir améliorer la performance (comme par exemple, Memcached ou MongoDB) même si certains fournissent une aide pour contrôler le compromis (comme Redis, MongoDB, Riak, Cassandra, Voldemort, etc.). Par conséquent, il est conseillé d'éviter d'utiliser les NoSQL comme solution principale de stockage à moins d'être certain que ce système est configuré pour répondre aux besoins de pérennité des données.

2. Les NewSQL

Fonctions Principales

- OLTP (Traitement Opérationnel en Ligne) obéissant aux ACID et prêts pour de gros volumes de données.
- Analyse SQL en temps réel des évènementielles (ex. : données de machines, données spatiales, logs Web).
- Stockage à échelle Big Data/OLAP des données structurées et hybrides (CRM, ERP, etc.)

Définition

Tout comme leurs équivalents NoSQL, ces nouveaux (et non si récents) SGBDR basés sur un SQL, permettent de faire face à l'échelle du Big Data. Pour cela, ils emploient des architectures distribuées (notamment MPP), un traitement en-mémoire, la technologie SSD et/ou ils incorporent dans leurs modèles de données, une flexibilité inspirée par les NoSQL. D'autres ont recours à l'analyse en-bases de données. Il s'agit là d'une stratégie qui combine dans un seul et même système, l'entreposage de données et les fonctions analytiques. Ceci permet de réduire la lenteur et d'empêcher le trop plein de flux aller-retour de données entre la base de données et une plate-forme d'analyse séparée.

Les systèmes qui obtiennent une efficacité grâce aux technologies en-mémoire et SSD, ont tendance à être des solutions obéissant aux ACID qui se fixent sur l'OLTP. Ceux qui atteignent cette efficacité avec des technologies type bases de données et/ou MPP (telles que des techniques innovantes de parallélisation et MapReduce) sont généralement tournés vers le décisionnel et limitent les contraintes ACID pour améliorer la performance (parmi les exceptions, Oracle Exadata qui supporte à la fois ACID/OLTP et OLAP).

Contrairement aux solutions NoSQL, les systèmes NewSQL sont plutôt commerciaux qu'open source. Ils atteignent leur capacité MPP grâce à un traitement symétrique sur un large nombre de processeurs implantés dans un seul et unique ordinateur (appartenant, souvent, à un individu), ou sur un petit groupe de ces mêmes ordinateurs (VoltDB constituant la seule exception). Dans certains cas, ils peuvent fournir une capacité de Big Data sur une seule machine puissante. Par exemple, la Teradata Extreme Data Appliance présente une capacité de stockage des données utilisateurs de 45TB par noeud, pouvant ainsi aller jusqu'à 186PB pour un ensemble de 4096 noeuds grâce à l'architecture Teradata MPP.

Les systèmes NewSQL orientés OLTP sont parfaits pour traiter de larges volumes de données, dans les situations où une milliseconde peut faire toute la différence : le commerce à haute fréquence, les commandes et contrôles des champs de bataille, les détections d'intrusion, le routage de réseaux, etc.

Exemples

Exemples de systèmes haute performance reposant sur des technologies enmémoire/SSD (orientées OLTP pour la plupart) :

- eXtreme DB-64 (db implanté)
- IBM SolidDB
- Oracle TimesTen In-Memory
- Teradata Extreme Performance (orienté OLAP)
- VoltDB (en-mémoire)

Exemples de systèmes haute performance reposant sur des technologies d'analyse en-bases de données et MPP (orientés analytique pour la plupart):

- Aster Data (acquis par Teradata)
- Greenplum
- IBM DB2 (avec la fonction de partitionnage des données)
- MS DATAllegro/ SQL Server 2008 R2 Parallel Data Warehouse
- Netezza (IBM)
- Oracle Exadata (OLTP + OLAP)
- ParAccel Analytic Database
- Teradata Extreme Data
- Vertica

Mises en garde

Ces solutions coûtent chères. En plus des coûts de licence et de développement, beaucoup ont besoin d'être installées sur des serveurs haut de gamme, beaucoup sont aussi des solutions de matériel/logiciel intégrés, ce qui rend particulièrement onéreux le passage à échelle. Il existe d'autres solutions possibles, comme des NoSQL moins coûteux, une passerelle SQL, un moteur CEP indépendant ou un DMS de recherche qui constituent des alternatives plus abordables. Cependant, dans les contextes critiques de Big Data, les NewSQL constituent le meilleur choix.

3.LES PLATES-FORMES DE RECHERCHES

Fonctions Principales

Traitement :

- Traitement en langage naturel (NLP)/sémantique (exploitation de texte, tagging automatique, classification et groupage, topographie des relations, etc.)
- Agrégation de données (types/sources de données hétérogènes)

Accès/Intéractions :

- Recherche plein-texte, en langage naturel
- Navigation à facettes
- Développement rapide d'applications métiers (service clients, logistique, MRO, etc.)

Analyses :

- Analyse des Sentiments
- Analyse Exploratoire (pour les utilisateurs commerciaux)
- Reporting/analyse opérationnels en quasi temps réel (là encore, destiné aux utilisateurs commerciaux)

Définition

Une "plate-forme de recherche" se définit comme étant un système complet de moteur de recherche, à savoir une plate-forme capable de gérer l'agrégation, l'accès et l'analyse d'information en plus de répondre aux besoins classiques de recherche en entreprise ou sur le Web. De telles plates-formes (connues également sous le terme d' "Accès Unifié aux Informations" - UIA) englobent tous les moyens de gestion des données de base, avec un traitement en langage naturel et une indexation. Ces fonctions supposent :

- La capture de données (crawlers, connecteurs, API)
- Le stockage de données (copies caches du contenu source et l'index en lui-même)
- Le traitement des données (NLP et construction et maintenance de l'index)
- L'accès aux données (recherche d'individus et d'objets, navigation à facettes, tableau de bord analytique)

Un système de recherche est de ce fait un DMS comme les NoSQL et NewSQL. Il est capable, autant qu'eux, de passer à l'echelle, c'est à dire d'employer des architectures distribuées, un traitement parallèle, des modèles de données orientées colonnes, etc. Cependant, les technologies NLP situées au coeur de ces DMS de recherche, font d'eux des compléments parfaits pour les systèmes NoSQL et NewSQL.

Tout d'abord, les DMS de recherche effectuent une recherche plein texte des systèmes NoSQL et NewSQL (ils apportent, en cela, une valeur significative). Ensuite, ils offrent une automatisation industrielle à la tâche de structuration des données, ce qui est primordial pour pouvoir pleinement tirer profit du Big Data.

Une plate-forme de recherche peut :

- Structurer efficacement le contenu non structuré
- Enrichir tout type de données avec des significations et des relations qui ne se démarquent pas dans les systèmes sources
- Agréger le contenu hétérogène et multi-sources (non structuré et/ou structuré) en un ensemble cohérent

Pour structurer les données non structurées, la plate-forme de recherche fait parcourir le contenu à travers des processeurs NLP qui l'enrichissent immédiatement avec des attributs et des valeurs structurels et sémantiques. Prenons comme exemple le traitement d'une page HTML. D'abord, avec le traitement de texte-centrique (voir page 17), un crawler capture sur la page des informations structurelles basiques telles que la taille de la page, le type de fichier, et l'URL. Il transmet ensuite ces informations, accompagnées du texte, à un indexeur.

L'indexeur viendra compléter cette base d'informations avec les résultats de l'analyse sémantique afin de créer un document holistique à indexer. Les étapes majeures de cette analyse incluent la détermination du langage utilisé dans le texte, suivie de l'analyse grammaticale du contenu pour permettre l'indexation de mots-clés (puis de phrases). Cette analyse détermine, au cours du processus, la forme grammaticale de chaque mot clé et ses possibles variantes grammaticales et sémantiques. Les indexeurs plus sophistiqués peuvent ensuite analyser le texte et identifier les synonymes et les termes associés, signaler les personnes, les endroits ou les organisations (en utilisant des listes standard, type entités nommées, ou des listes de clients), déterminer le sujet général traité, décider si le ton général est positif ou négatif, etc. Des règles commerciales peuvent être appliquées pour guider l'analyse et réaliser des transformations de données variées de style ETL. Ceci peut supposer l'extraction de seulement quelques attributs pour pouvoir transformer de gros volumes de données en une sous-série pertinente et manipulable.

Une fois que cette version structurée d'un document non structuré a été réalisée, les technologies sémantiques identifient les liens que ce document peut avoir avec d'autres documents, que ces documents soient issus de sources structurées telles que les bases de données, de sources semi-structurées comme les logs Web ou d'autres sources non structurées comme les serveurs.

Lorsque l'utilisateur cible est un utilisateur standard et non un développeur ou statisticien expérimenté, cette plate-forme de recherche présente un avantage significatif. Il n'existe en effet aucune autre technologie qui soit aussi efficace que la recherche pour rendre le Big Data utile et accessible aux utilisateurs individuels ordinaires

La recherche en langage naturel, la navigation à facettes et la visualisation de données sont des outils qui offrent aux utilisateurs de tout niveau de compétence, un moyen instantanément familier d'explorer et d'analyser le Big Data. Ils leur donnent la possibilité de lancer toute recherche ou toute tâche analytique de la même façon que sur Internet : en entrant une phrase ou quelques mots-clés dans un champ de saisie. Ensuite, avec ces outils, l'utilisateur peut réaliser des analyses exploratoires en affinant sur des clusters ou groupes de données dynamiques (souvent représentés sous forme de menus texte ou sous forme visuelle comme les tableaux et les graphiques).

Cette facilité d'utilisation, ajoutée aux temps de réponses quasi instantanés des plates-formes de recherche, encourage l'exploration : si les utilisateurs reçoivent des réponses instantanées aux questions qu'ils posent à leur manière, ils vont vouloir aller plus loin et explorer encore plus. S'il est difficile de formuler les questions et si les réponses sont lentes à arriver, les utilisateurs décideront de chercher ailleurs ou simplement, d'abandonner leur recherche.

Pour toutes ces raisons, les plates-formes de recherche sont vues comme étant des compléments idéaux pour les NoSQL et les NewSQL. Dans certains contextes, elles vont même jusqu'à constituer une solution alternative pragmatique à ces deux systèmes.

Exemples

Les plates-formes citées ci-dessous sont disponibles sous forme de systèmes indépendants que vous pouvez utiliser de manière polyvalente au sein de votre organisation. Il existe d'autres plates-formes qui ne sont pas vendues sous forme indépendante mais qui fournissent aussi aussi une infrastructure d'applications métiers de type Search-Based.

Notons cependant que, même si ces plates-formes sont conçues pour une utilisation qui peut adresser de larges volumes de données, nous ne pouvons nous porter garants de leur performance à l'égard des séries de Big Data (à l'exception, bien sûr, de CloudView). Nous vous recommandons donc de bien analyser ces produits, en suivant la procédure habituelle : références des SBA et test de leur performance pour des utilisations similaires, Pilotes (POC) en utilisant vos propres données, etc.

- Attivio Active Intelligence Engine
- EXALEAD CloudView™
- Expert System's Cogito
- Fabasoft Mindbreeze
- Isys Search Software
- Lucene/Nutch/Solr (Apache)
- Vivisimo Velocity

Notons que la plupart des grands noms de la recherche tels que Google, Baidu, Bing Yahoo! et Ask ne figurent pas dans cette liste. L'explication réside dans le fait qu'il n'y a pas de moteur de recherche du type recherche "end-to-end"/platesformes UIA qui fasse l'objet d'une licence commerciale pour entreprises. EXALEAD constitue l'exception à la règle. Sa plate-forme pour entreprise, CloudView, offre également un service public de recherche sur le Web.

Google propose, cependant, une offre de recherche pour entreprises, nommée Google Search Appliance mais il s'agit principalement d'une boîte noire, "plug-and-play" conçue pour répondre aux besoins basiques des entreprises en matière de recherche et non d'un DMS de recherche complet. Google propose, néanmoins et de manière encore timide, des DMS de traitement Big Data sur le Cloud (voir la prochaine section relative aux Outils Annexes).

Microsoft possède également une offre de recherche pour entreprises, nommée MS FAST, mais elle est conçue comme un outil classique de recherche au coeur d'un environnement Sharepoint. Il ne s'agit donc pas d'une plate-forme DMS générale (la même chose s'applique aux outils proposés par IBM, Oracle et SAP).

Pour finir, il est important de préciser que Apache Lucene peut se déployer sur le Web. Il n'est cependant pas associé aux grands noms des moteurs de recherche. L'indexeur de recherche Lucene et le crawler Nutch ont été développés pour compléter le moteur de recherche open source Nutch (conçu par Doug Cutting, le créateur de Lucene, Nutch et Hadoop).

Lucene, Nutch et le serveur de recherche Solr (ou les composants équivalents) doivent être combinés pour engendrer une plate-forme de recherche complète. Toutefois, ces trois systèmes étant open source, il faut un fort degré d'expertise pour garantir une intégration réussie. Ceci est particulièrement vrai en ce qui concerne les SBA (Search-Based Applications) car ces composants ne possèdent aucun outil integré pour les administrer.

Mises en garde

Les plates-formes de recherche constituent des solutions matures parfaites pour grouper, accéder et analyser de larges volumes de données aux formats multiples et issues de sources multiples. Elles sont aussi formidables pour permettre le développement rapide d'applications métiers sécurisées et efficaces construites sur de gros volumes de données. Elles ne conviennent cependant pas au stockage d'archives de données, aux OLTP ou aux OLAP complexes et/ou historisés.

Il est important de retenir que les moteurs de recherche ne sont pas tous égaux. En plus de la procédure d'analyse des produits recommandés ci-dessus, il est primordial de disposer d'une check-list telle que celle proposée ci-dessous pour s'assurer que le produit répond bien aux besoins en matière de consolidation, d'accès, de découverte et d'analyse de l'information à l'échelle du Big Data.

Est-ce que la plate-forme de recherche…
- collecte et traite les données non structurées, structurées ou semi-structurées ?
- présente un cadre ouvert et basé sur des API et connecteurs ?
- permet le regroupement de données en plus d'une recherche fédérée, de mashups et d'une méta-recherche ?
- utilise des technologies sémantiques pour analyser et enrichir les sources de données efficacement ?
- catégorise automatiquement le contenu pour permettre une recherche à facettes, une navigation et un reporting ?
- fournit une recherche API ou des outils internes de tableau de bord pour la visualisation et l'analyse des données ?
- offre une architecture distribuée avec un traitement parallèle, ou une architecture équivalente pour assurer une performance, une scalabilité et un coût satisfaisants pour les environnements de données ?

Idéalement, la plate-forme de recherche devrait également être suffisamment sophistiquée pour pouvoir automatiser les tâches essentielles de configuration, d'application et de gestion.

C. Les Outils Annexes

1. LES SERVICES CLOUD (OU "EN NUAGE")

Fonctions Principales

Il y a aujourd'hui des offres en Cloud ou "en nuage" pour chaque besoin relatif à la gestion des données :
- L'acquisition de données
- Le traitement/calcul par lots des données
- L'accès aux données
- L'analyse de données
- Le stockage de données

Définition

Représenté par une icône de nuage utilisée pour illustrer l'Internet dans les schémas de réseaux informatiques, le "Cloud Computing" (ou "Informatique en Nuage") renvoie à tout service ou technologie informatique livrée via Internet et sur la base d'une souscription ou d'une tarification à l'utilisation.

Les applications métiers sur le Cloud, labélisées "Software-as a Service" (SaaS) "Logiciel vu comme un Service" sont les plus connues avec, entre autres, Salesforce et Google Apps.

Aujourd'hui, quasiment tous les fournisseurs de logiciels commerciaux B2B proposent des déclinaisons de leurs produits en SaaS. Nous assistons de plus, en ce moment, à un essor fulgurant des application SaaS de type Big Data.

Les "Infrastructure comme un Service" ("Infrastructure-as-a-Service" ou IaaS) constituent la deuxième tendance. Les entreprises ont recours depuis longtemps à des solutions IaaS très abondantes pour faire, par exemple, de l'hébergement isolé (et souvent virtuel) de sites Web.

> *"L'informatique en nuage et les nouvelles catégories d'algorithmes permettront de garder plus de détails de transactions, de les garder plus longtemps et de les faire se mélanger avec d'autres larges séries de données (ex. annuaires téléphoniques et registres de propriété)."* [17]

En matière de Big Data, les trois plus importantes offres IaaS concernent :
1. Le stockage de données
2. Le traitement de données (services informatiques statistiques),
3. L'acquisition de données (également appelée "Data-as-a-Service" ou "DaaS")

En matière de stockage de données, de nombreux fournisseurs spécialisés dans les solutions de stockage/sauvegarde/récupération et les grands du Web comme Amazon et Microsoft offrent désormais des solutions NoSQL en Cloud à des prix abordables.

Au niveau du traitement des données, Amazon, Google, Microsoft et d'autres encore permettent progressivement aux entreprises d'utiliser leurs infrastructures industrielles MapReduce sur le Cloud pour traiter leurs données.

Pour ce qui est de l'acquisition des données, les sociétés commerciales offrent aussi une sélection de plus en plus grande de recueils de données spécialisées, accessibles depuis le Cloud. Ces propositions largement commerciales sont complétées par un nombre grandissant de répertoires de données publiques publiées sur le Web par les organisations publiques, d'éducation et scientifiques.

Cette grande diversité de services Cloud permet aux organisations de toute taille et en tout genre de dépasser les limites techniques et financières de l'exploitation du Big Data.

Exemples

Services de stockage :

- Amazon S3
- EMC Atmos
- Nirvanix
- Google Storage (projet Labs)

Services Informatiques :

- Amazon Elastic Compute Cloud (Amazon EC2)
- Google Prediction API & BigQuery (ces deux offres faisaient initialement partie du programme Google Labs, Google choisira peut-être de les commercialiser. Il en va de même pour les applications geo-spatiales Earth Builder).

Recueils de données :

- Factual (divers)
- InfoChimps (divers)
- Windows Azure Marketplace DataMarket (divers)
- Hoovers (commercial)
- Urban Mapping (geographique)
- Xignite (finance)

Il existe également un grand nombre d'entreprises qui offrent des systèmes de bases de données en nuage (relationnelles pour la plupart) plus optimisés pour les applications sociales ou mobiles que pour le stockage, le traitement et l'analyse du Big Data. Parmi elles :

- Database.com
- Amazon Relational Database Service (RDS)
- Microsoft SQL Azure
- Xeround

Mises en garde

En plus des mises en garde classiques du modèle Cloud (confidentialité, efficacité, interactivité, etc.), il est aussi important de prendre en compte la spécificité des défis en matière de Big Data qui supposent de travailler de façon isolée sur de très volumineuses séries de données. Ces séries sont coûteuses et lentes à déplacer et peuvent impacter les capacités du réseau, aussi importantes soient-elles.

Par exemple, avec une connexion T1 (1,544 Mbps), le téléchargement d'un téraoctet de données peut prendre minimum 82 jours. Avec une connexion à 10Mbps, il peut prendre minimum deux semaines. Ceci explique en partie pourquoi Amazon AWS préfére renforcer son réseau interne à haute vitesse (qui contourne Internet), avec du matériel de stockage supplémentaire.[18]

2. LES OUTILS DE VISUALISATION

Fonctions principales

- Reporting et Analyse

Définition

Le fait de représenter le Big Data sous forme visuelle aide à le rendre plus accessible aux individus. L'aide est telle que la plupart des logiciels relatifs aux sciences, à l'industrie et à la Business Intelligence (BI) fournissent des outils de visualisation et de navigation.

Pour la plupart des vendeurs BI – dont SAP Business Objects, IBM Cognos, MicroStratégie, SAS Institute et Information Builders – les capacités de visualisation incluent par exemple, les histogrammes interactifs, les tableaux de bord, les graphiques circulaires, et la topographie géographique. Les moteurs SBA tels que CloudView offrent également cette possibilité, en proposant des représentations navigables telles que les heatmaps, les diagrammes de dispersion, les graphiques, les graphiques de relations, les nuages de tags, les "sliders", les cartes routières et géospatiales.

En plus des fonctions de traçage en 2D et 3D et les fonctions de visualisation de volume en 3D, de nombreux outils de visualisation ont la capacité d'exporter des résultats vers des formats graphiques communément utilisés.

Exemples

Les exemples d'outils de visualisation incluent :

- Advizor
- Gephi
- JMP
- Panopticon
- Spotfire
- Tableau

Mises en garde

Ces outils sont extrêmement efficaces pour synthétiser de grandes séries de données et pour découvrir et explorer des relations et tendances inattendues mais il est important de sélectionner le bon type de représentation en fonction de la série de données et de l'analyse recherchée. Dans le cas contraire, le risque est de se retrouver avec une représentation qui induit en erreur, mène à la confusion ou tout simplement, qui peut ne pas être lue. Pour éviter ce risque et pour faire bon usage de la visualisation en général, il est recommandé de choisir des outils qui proposent une très grande intéractivité et permettent de jouer pleinement avec les données.

Notons aussi que la transformation graphique est parfois extrêment consommatrice de ressources, notamment face à de très larges recueils de données. Si vous souhaitez employer un outil

indépendant de visualisation, il faudra vous attendre à patienter pour pouvoir exploiter le moteur de visualisation.

5) ETUDES DE CAS DE RECHERCHE

A. GEFCO
Dépasser les Limites de Performance

Avec plus de 10 000 employés dans 100 pays, GEFCO est l'un des dix premiers groupes européens de transport et de logistique. L'entreprise fournit des services en matière de transport multimodal et de logistique amont et aval à des clients industriels spécialisés dans les secteurs automobile, deux roues, électronique et biens de consommation.

GEFCO est responsable au quotidien de la localisation de 7 millions de véhicules. Leur base de données Oracle consolide tous les mouvements logistiques de ces voitures à travers le monde ainsi que les 100 000 événements logistiques quotidiens. Après avoir mis en oeuvre sur deux ans des projets coûteux d'optimisation, GEFCO rencontrait encore des problèmes de performance avec l'ancien système Track & Trace : un temps de réponse aux questions supérieur à une minute, un accès restreint pendant les heures de travail pour éviter les conflits entre demandes d'information et traitement interne des opérations, ainsi qu'un délai de mise à jour pouvant aller jusqu'à 24 heures. Avec 3 Téraoctets de données, GEFCO faisait soudainement face à un réel problème de "Big Data".

Plutôt que de continuer à s'appuyer sur les SGBDR existants ou d'installer un système NewSQL, GEFCO a décidé de transformer son portail Track & Trace en une SBA CloudView. Cette refonte du système, récompensée pour son innovation, offre beaucoup plus de performance, d'agilité, d'utilisabilité et de sécurité – à un coût moins élevé – que leur précédent modèle de centre de données.

Les grands points d'amélioration :

- Un temps de réponse aux questions quasi-instantané ;
- Une fréquence de mise à jour passant de 24 heures à 15 minutes (avec un système capable, si besoin est, d'actualiser les données en quasi-temps réel)
- Une réduction des coûts de 50% par utilisateur
- Une nette amélioration au niveau de l'accessibilité des informations – sans nécessité de former les utilisateurs destinataires
- Une disponibilité de 99,98% avec un investissement matériel limité
- Une production automatisée de reporting opérationnel multi-dimensionnel

Aussi, le prototype initial de cette application fût développé en seulement 10 jours et la première version complète fût diffusée trois mois plus tard.

> *"Chaque jour, je vois des milliers d'événements consolidés en temps réel, ce qui me pousse même à me loger sur le système pour le croire !... La stabilité et la performance de cette application sont incroyables, une véritable innovation."*
>
> Guillaume Rabier, Directeur des Etudes et Projets IT, GEFCO

GEFCO : les formulaires longs et complexes ont été remplacés par une boite de recherche simple interrogeable en langage naturel et des fonctionnalités de navigation multi-dimensionnelle et cartographique. Ici, la déclinaison de cette application sur une tablette numérique.

La nature des activité de GEFCO et les caractéristiques innovantes de la nouvelle application (actualisation des données, réactivité instantanée, disponibilité élevée et ergonomie maximum) ont naturellement conduit à la création d'une version

mobile de cette même application enrichie de fonctions de routage et de cartographie.

B. Yakaz
L'innovation par la Recherche et les NoSQL

Fondée en 2005, Yakaz est un moteur de recherche vertical réputé et spécialisé dans les petites annonces (logements, voitures, motos, emplois, biens et services). Ce site offre un accès unifié à 60 milllions d'annonces, diffusées en 50 langues et issues de dizaines de milliers de sites Internet, offrant à ses visiteurs la possibilité de parcourir tout type d'annonces. C'est un service informatif et rapide qui accueille chaque mois 15 millions de visiteurs uniques.

Les créateurs de cette entreprise, deux anciens directeurs d'AOL, ont décidé de construire ce nouveau service avec la plate-forme EXALEAD CloudView™. Ils étaient convaincus que CloudView accélèrerait leur processus commercial et leur permettrait de gérer rapidement des volume de données considérables.

Aujourd'hui, Yakaz est présent dans le monde entier et s'adresse aux clients de 193 pays. L'application repose sur une infrastructure innovante autour de quatre composants clés :

- **EXALEAD CloudView™**
  CloudView est utilisé pour "crawler" les petites annonces sur le Web, pour automatiquement structurer les données Web extraites par le crawler, pour traiter les annonces choisies via RSS et XML, pour mettre à jour l'index et pour effectuer un traitement à l'échelle des requêtes.
- **Cassandra**
  Cassandra est un projet open source Apache. C'est une base de données NoSQL qui marie l'architecture distribuée Dynamo avec le modèle de données orientées colonnes Bigtable. Il gère et stocke les données d'utilisateurs et d'applications.
- **Ejabberd**
  Edjabberd est un serveur de messagerie instantanée Jabber/XMPP figurant sous la licence GPLv2 (Free and Open Source) et écrit en Erlang/OTP. Il est utilisé pour aider Yakaz à incorporer les interactions sociales dans son portail, en commençant par le nouveau service de messagerie instantanées d'utilsateur à utilisateur.
- **OpenStreetMap**
  OpenStreetMap est un service open source de géo-données. Ses cartes sont créées par l'utilisation de données issues des matériels portables GPS, de photographies aériennes, de sources de données publiques et de propositions d'utilisateurs.

Cette unique infrastructure offre à Yakaz une plate-forme agile et évolutive qui répond parfaitement à son esprit d'innovation.

C. La Poste
Construire des Applications Métiers pour le Big Data

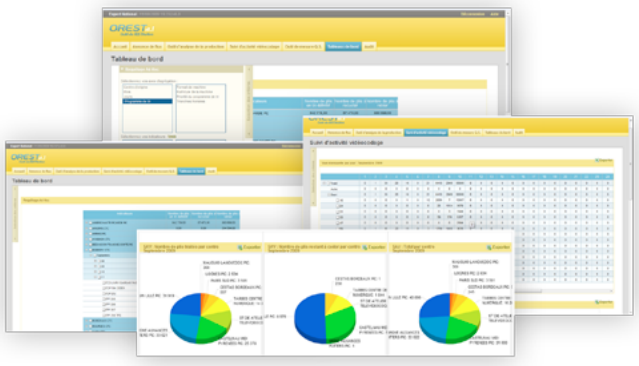
Intéressons-nous à deux SBA CloudView déployées par cette entreprise qui réalise plus de 20 milliards d'euros de chffre d'affaires et compte près de 15 000 agences locales :

1. Une plate-forme d'analyse opérationnelle en quasi temps réel du courrier
2. Un système multicanal de Gestion de la Relation Client

Reporting et Analyse Opérationnels

Cette entreprise utilise une application CloudView de reporting et d'analyse opérationnels pour surveiller en temps réel, plus de 62 milliards d'événements annuels impliquant 18 000 personnes. Il s'agit là d'un véritable environnement de Big Data avec :

- 55 millions de courriers traités au quotidien
- 3 à 5 évènements/traitements par courrier écrit
- 300 millions de registres créés chaque jour, pouvant aller jusqu'à 7000 registres par seconde
- Un préalable de rétention de données de 21 jours
- Un index de 6,3 millions de registres représentant 9 Téraoctets sur les 90 Téraoctets de données brutes.



Les SBA regroupent les données provenant de diverses sources dont les applications métiers, les machines de tri de courriers, les matériels de codage-vidéo scalable et permettent d'avoir une vue globale sur les flux de courriers pour :

- Une détection et une correction des évènements exceptionnels en temps et en heure (Analyse du Service de Qualité).
- Une prévision et une optimisation des flux de traitement et distribution.
- Des analyses multidimensionnelles d'optimisation du planning stratégique

Cette capacité à regrouper et manipuler les données offre aussi à La Poste, le moyen de développer de nouveaux services premium pour ses clients tels que les boîtes aux lettres sécurisées et virtuelles pour pouvoir recevoir et stocker d'importants documents, une messagerie SMS "push" pour les livraisons, un service Track & Trace pour les lettres, la livraison de courriers concrets par le biais d'emails (mail2email) et, pour les clients commerciaux concernés par de larges volumes, la gestion complète de leur organisation de courrier.

Pour cette application, cette entreprise aurait pu avoir recours à une base de données NoSQL et un moteur de recherche. Cependant, le moteur CloudView en lui-même a réussi à satisfaire ses besoins de capture, de traitement, de stockage et d'accès aux données.

Système Multicanal d'Informations Clients

Cette application CloudView propose une solution élégante qui évite d'avoir recours à un processus complexe et coûteux :

- En présentant un accès unique en quasi temps réel aux données, gérées au sein de 10 grandes bases de données qui permet d'avoir une vue globale sur les clients et les opportunités.
- En tirant profit de la couche unifiée de données pour permettre les interactions entre clients à travers tout type de canal : chat, SMS, Web call back, téléphone, email, messagerie instantanée et interactions en face à face dans les agences postales.

En choisissant la stratégie SBA, cette entreprise a déployé, en seulement 90 jours, la première version opérationnelle de son Système Multicanal d'Informations Clients et ce, sans impact sur les systèmes existants.

D. Et Bien d'Autres Encore...

Vous trouverez, ci-dessous, d'autres profils d'entreprises ayant fait le choix de CloudView pour mieux gérer leur Big Data.

Rightmove

Rightmove est un site Web anglais spécialisé dans l'immobilier et attirant plus de 29 millions de visiteurs par mois. Rightmove a choisi de passer son système de recherche et d'accès de sa base Oracle sur CloudView. En seulement 3 mois, Rightmove a :

- Considérablement améliorer son expérience utilisateur
- Mettre à 9 le nombre de CPU à la place de 30 CPU Oracle
- Divisé le coût des 100 requêtes par 6
- Supporté un maximum de 400 requêtes par seconde
- Atteint un niveau de disponibilité de 99,99%

> *"EXALEAD CloudView™ a permis un développement rapide de la fonctionnalité de recherche avancée, tout en réduisant le coût de la recherche de 83%."*
>
> Peter Brooks-Johnson, Rightmove Product Director

Un acteur international des télécommunications

L'une des plus grandes entreprises de télécommunications au monde (avec plus de 200 millions d'utilisateurs) a choisi CloudView pour jouer un rôle pivot dans la modernisation de ses systèmes de support technique. Avec CloudView, les services de maintenance des réseaux ont une vue d'ensemble sur les différentes informations relatives au support et au service, dont :

- Les données CRM (nom du client, adresse, section de marché, type de client, etc.)
- Informations de provisions (type d'équipements, longueur de câbles, obstacles aux lignes)
- Données de surveillance de réseaux (statut, performance, etc.)
- Données relatives aux contrats (options, durée, termes, etc.)
- Informations techniques (historique d'interventions, problèmes encourus par les techniciens sur place, rendez-vous en attente)

Entreprise de Produits de Grande Consommation

CloudView est en voie d'intégrer une nouvelle plate-forme mobile de collaboration pour une entreprise de Produits de Grande Consommation qui s'apprête à fournir à 35 000 utilisateurs un accès intelligent à 175 Téraoctets de données brutes (avec 3 millions de nouveaux objets ajoutés au quotidien).

BnF

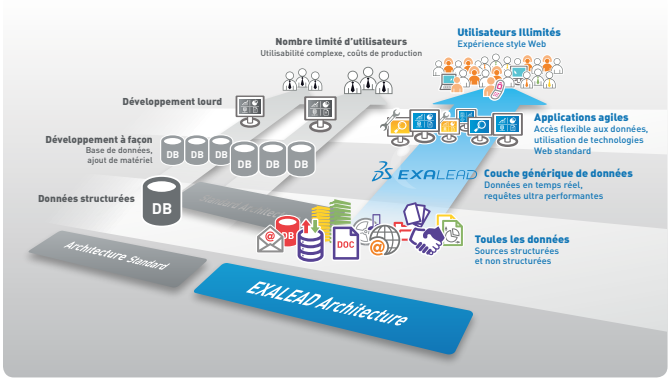
CloudView est une pierre angulaire de l'infrastructure de l'accès de "Gallica", la bibliothèque numérique de la Bibliothèque nationale de France (ou BnF). Au jour d'aujourd'hui, la BnF a numérisé plus d'un million d'oeuvres dont des livres, des cartes, des manuscrits, des images, des journaux, des partitions et enregistrements musicaux.

Ministère de la Défense des Etats-Unis

CloudView est utilisé pour opérer un search Web vertical et privé, centré sur les informations et problèmes à caractère environnemental, pour le Ministère de la Défense américain. Les données proviennent de serveurs publics (tels que Google, CloudView, Yahoo ! etc.), et de contenus de type Deep Web issus des bases de données et applications gouvernementales, scientifiques, industrielles et commerciales.

Le Directory Content Enhancer EXALEAD

Le Directory Content Enhancer EXALEAD est un outil permettant aux éditeurs d'annuaires online d'exploiter les ressources illimitées du Web pour valider, enrichir et étendre leurs propres contenus et ainsi enrichir l'expérience utilisateur.



POURQUOI CLOUDVIEW ?

Simplicité d'utilisation, agilité et performance font d'EXALEAD CloudView™ la plateforme de recherche idéale pour les environnements Big Data. Elle offre en effet :

- **Une Performance Big Data**
  Développée pour le Web et les entreprises, CloudView permet des traitements sémantiques avancés sur des milliards d'enregistrements tout en garantissant des temps de réponse instantanés pour des milliers d'utilisateurs simultanés.
- **Une Connectivité Big Data**
  CloudView fournit le Web crawler le plus avancé et les connecteurs les plus perfectionnés aux sources typées Big Data telles que les files d'attente de données, les unités centrales, les registres NoSQL (Hadoop HDFS), les entrepôts de données, les plates-formes BI et les réseaux sociaux.
- **L'Analytique Big Data**
  Les capacités de calcul et de faceting de CloudView sont les plus sophistiquées du marché. La plate-forme permet des calculs complexes d'ensembles et de clusters numériques, géophysiques et virtuels. Les outils embarqués de visualisation, la navigation multi-dimensionnelle, le NLP donnent à chacun le moyen d'effectuer des analyses riches exploratoires et opérationnelles du Big Data. Ce, sans avoir besoin d'être formé ou de demander de l'aide à des experts IT ou autre.
- **Le Développement d'Applications Métiers pour le Big Data**
  Enfin, CloudView est la seule plate-forme fournissant un cadre de développement ultra simplifié avec, par exemple, un outil de MashUp Builder permettant à l'aide du "glisser-déposer" de construire des applications métiers agiles sur des sources de Big Data en un temps record.

Pour en savoir plus sur le rôle que joue CloudView pour aider les entreprises à saisir les opportunités du Big Data, nous vous invitons à nous contacter dès aujourd'hui afin que nous puissions vous faire une démonstration de certaines de nos nombreuses SBA. Vous pouvez aussi nous contacter si vous souhaitez effectuer des pilotes sur vos données.

END NOTES

1. La traduction plus littérale du grec est "Donnez-moi un point d'appui et je soulèverai le monde," mais la variation utilisée ici est plus appliquée aux challenges et aux opportunités du Big Data.
2. Pour une analyse de l'histoire du Big Data, vous pouvez consulter les comptes-rendus de ces 35 dernières années, fournis par les conférences annuelles "Very Large Data Base"(VLDB) Endowment : www.vldb.org.
3. Chris Anderson, Wired Magazine, N° 16.07, "The Petabyte Age: Because More Isn't Just More — More Is Different,"Juin 2008.
4. New SQL est un terme utilisé par l'analyste de 451 Group Matthew Aslett dans un blog post du 6 Avril 2011 pour designer un groupe récent de bases de données relationnelles SQL hautement performante. Majoritairement open source, elle inclut les moteurs de stockage MySQL (ScaleDB, Tokutek), des appareils avec logiciel et materiel intégrés (Clustrix, ScalArc, Schooner) et des bases de données qui utilisent des technologies de fragmentation transparente (ScaleBase, CodeFutures). Nous appliquons ce terme à un ensemble plus étendu de technologies et incluons de nombreux systèmes commerciaux.
5. Gantz J and Reinsel D., "The Digital Universe Decade – Are You Ready?" IDC, Mai 2010, Sponsorisé par EMC Corporation.
6. TheInfoPro Inc. Storage Study, Wave 9, Avril 2007.
7. Scott Spangler, IBM Almaden Services Research, "A Smarter Process for Sensing the Information Space," Octobre 2010.
8. McKinsey Global Institute, "Big data: The next frontier for innovation, competition, and productivity," Mai 2011.
9. Galen Gruman, "Tapping into the power of Big Data," Numéro 3, Technology Forecast (Making sense of Big Data), PriceWaterhouseCoopers, 2010.
10. Constance Hays, "What Wal-Mart Knows About Customers' Habits," The New York Times, 14 Novembre 2004.
11. The Economist, "A different game: Information is transforming traditional businesses," 25 Février 2010.
12. Alon Halevy, Peter Norvig, and Fernando Pereira (Google), "The Unreasonable Effectiveness of Data," IEEE Intelligent Systems, Numéro 2, Mars/Avril 2009.
13. Jeremy Ginsberg, et al, "Detecting influenza epidemics using search engine query data," Nature, v457, Février 2009. Voir aussi www.google.org/flutrends/.
14. Ibid, The Economist.
15. Giuseppe DeCandia, et al, Amazon.com, "Dynamo: Amazon's Highly Available Key-value Store," ACM SOSP'07, octobre 14–17, 2007.
16. Une partie du contenu de la section NoSQL est extraite du livre "Search-Based Applications: At the Confluence of Search and Database Technologies," rédigé par Gregory Grefenstette et Laura Wilber, Morgan & Claypool Publishers, Decembre 2010. Pour plus d'information sur les systèmes NoSQL, voir aussi www.nosql-database. org .
17. Jeff Jonas, "Sensemaking on Streams," Jeff Jonas Blog, 14 février, 2011.
18. Site Web AWS : "AWS Import/Export - Selecting Your Storage Device," aws.amazon.com/importexport/.

Pour des produits hors pair

CATIA
Le produit virtuel

GEOVIA
La planète virtuelle

SOLIDWORKS
La conception 3D

EXALEAD
L'intelligence de l'information

SIMULIA
La simulation réaliste

NETVIBES
Les outils de veille personnalisés

DELMIA
Fabrication et Production numériques

3DSWYM
L'innovation sociale

ENOVIA
L'innovation collaborative

3DVIA
La communication 3D

A propos d'EXALEAD

EXALEAD est un fournisseur de logiciels de recherche et d'accès à l'information en entreprise et sur le Web. Sa solution EXALEAD CloudView™, plate-forme logicielle innovante, offre la possibilité de concevoir simplement une nouvelle gamme d'applications : les Search-Based Applications (SBA).

EXALEAD est une société du Groupe Dassault Systèmes.

EXALEAD EMEA
Dassault Systèmes
10 place de la madeleine
75008 Paris
France

Visit us at
3DS.COM/EXALEAD
