Livre blanc IT@Intel

Intel IT

Meilleures pratiques IT
Business Intelligence
Juillet 2012

Exploration des Big Data pour optimiser la Business Intelligence

La capacité à extraire et analyser les Big Data permet aux grandes organisations de parvenir à une compréhension plus profonde et plus riche des modèles et tendances au sein de leur activité, les aidant ainsi à optimiser leur efficience opérationnelle et à générer des avantages concurrentiels dans les domaines de la fabrication, de la sécurité, du marketing et de l'informatique.

Moty Fania
Advanced Business Intelligence Solutions, Intel IT

John David Miller
Ingénieur principal, Intel IT Labs

Vue d'ensemble

Pour renforcer les capacités d'Intel en matière de Business Intelligence (BI), Intel IT déploie actuellement des systèmes et compétences dédiés à l'analyse des Big Data. Les Big Data sont essentiellement constituées de grands ensembles de données non structurées, pouvant représenter jusqu'à 90 % des données d'entreprise. La capacité à extraire et analyser les Big Data, quelle que soit leur forme, permet d'aboutir à une compréhension plus profonde et plus riche des modèles et tendances de notre activité. Cela nous permet d'optimiser l'efficacité opérationnelle et de générer des avantages concurrentiels dans les domaines de la fabrication, de la sécurité, du marketing et de l'informatique.

Le renforcement des compétences organisationnelles requises pour extraire et traiter les Big Data, dans le but de conduire des analyses prédictives et prescriptives, sera un facteur de performance clé à l'avenir. Cela permettra en effet à Intel de :

- Prendre de meilleures décisions
- Améliorer son agilité métier
- Accélérer le rythme de ses innovations
- Découvrir et exploiter de nouveaux marchés

Les progrès réalisés dans le domaine du traitement parallèle permettent aujourd'hui de prendre en charge les Big Data, tant et si bien que la pratique consistant à collecter et stocker des informations bien avant que leur valeur ne soit réellement comprise et exploitée est en train de devenir courante. Ces progrès permettent également d'aborder de nombreuses problématiques métier auparavant trop lourdes ou complexes à gérer.

Intel IT a déployé une plateforme Big Data en 2012, dans le cadre d'un partenariat étroit engagé avec les entités opérationnelles d'Intel, afin de conduire des essais de validation technique (*proof of concept*) pour démontrer sa capacité à fournir une BI exploitable par l'entreprise.

En 2012, nos projets Big Data ont notamment inclus :

- Détection des malwares
- Validation de conception des puces
- Information commerciale stratégique
- Système de recommandations

Intel n'en est encore qu'aux premiers stades de son programme de renforcement des capacités visant à optimiser la BI grâce aux Big Data. Nous anticipons cependant une croissance rapide de ces capacités dans les domaines suivants : recherche et développement, cybersécurité, ingénierie, fabrication, opérations, développement commercial et gestion des ressources humaines.

Table des matières

IT@INTEL

Le programme IT@Intel vise à mettre en relation les professionnels de l'informatique à travers le monde avec leurs homologues au sein de notre organisation, en partageant les enseignements tirés, les méthodologies et les stratégies. Notre objectif est simple : partager les meilleures pratiques d'Intel IT permettant de créer de la valeur métier et de générer des atouts concurrentiels. Retrouvez-nous dès aujourd'hui sur www.intel.com/IT ou contactez votre représentant Intel local si vous souhaitez en savoir plus.

LE DÉFI À RELEVER

Partout dans le monde, la quantité de données brutes est en croissance exponentielle, sous l'effet notamment de l'explosion du nombre des appareils connectés, services Internet, réseaux sociaux, caméras, capteurs et autres contenus générés par les utilisateurs. On estime par ailleurs que jusqu'à 90 % des données d'entreprise (documents, pages web et e-mails, notamment) ne sont pas structurées. Les logiciels de base de données courants sont totalement dépassés par le volume et la complexité de ces données. Cette situation appelle donc une nouvelle approche.

Inondés de données, mais assoiffés de connaissances

Selon le rapport « Big data: The next frontier for innovation, competition, and productivity », du McKinsey Global Institute, dans 15 des 17 secteurs d'activité recensés aux États-Unis, chaque entreprise conserve en moyenne plus de données que la Bibliothèque du Congrès américain.[1] Wal-Mart constitue à ce titre un bon exemple : le géant de la grande distribution gère plus d'un million de transactions client par heure. Cela génère des données qui sont importées dans des bases dont les volumes sont estimés à plus de 2,5 pétaoctets (Po), soit 167 fois la quantité d'informations contenue dans tous les livres de la Bibliothèque du Congrès.

La plupart des Big Data sont issues des milliards de transactions et autres éléments d'information que les entreprises telles qu'Intel enregistrent chaque jour sur leurs clients, leurs fournisseurs et leurs opérations. Autrefois considérées comme posant principalement un problème de stockage, les Big Data sont aujourd'hui reconnues comme un atout stratégique à part entière, véritable mine d'or permettant de faire émerger des éléments de compréhension exploitables dans tous les domaines d'activité de l'entreprise.

Les Big Data représentent pour nous un intérêt dans deux grands domaines :

- **Bases de données Big Data.** Ces bases contiennent des données structurées qui sont tout simplement trop volumineuses pour un système de gestion de base de données relationnelle (SGBDR) classique.
- **Analyse approfondie.** Permet de rechercher des réponses à des problèmes complexes et ne se prêtant pas à une résolution claire et définitive. Ces réponses ne sont en général pas directement codifiées dans les données source. Les outils de visualisation et d'analyse des Big Data permettent plutôt de faire émerger de précieux éléments de compréhension, au fil d'un processus d'affinement et d'abstraction.

Jusqu'à une période récente, la plupart des entreprises devaient se résoudre à essayer d'agréger les données pour conduire leurs analyses, ou à prélever des échantillons dont elles tentaient ensuite d'extraire un sens par extrapolation. Cela constitue en fait toujours le statu quo. Gartner prédit ainsi que « d'ici 2015, plus de 85 % des entreprises du classement Fortune 500 ne seront toujours pas parvenues à exploiter efficacement les Big Data pour en tirer un avantage concurrentiel. »[2] Les entreprises les plus en pointe ont toutefois d'ores et déjà déployé des grandes capacités d'analyse Big Data et pourraient bien en dégager des résultats substantiels. Selon Gartner, ces entreprises avancent très vite et modernisent leurs pratiques de Business Intelligence, d'extraction de données (Data Mining) et d'analyse décisionnelle en intégrant les nouveaux outils et compétences qui émergent à l'ère du Big Data.

Le professeur Eric Brynjolfsson, directeur du Center for Digital Business Research du MIT, a mené des recherches sur 179 grandes sociétés cotées en bourse et a constaté que les entreprises qui utilisent « un processus de décision fondé sur les données » étaient plus productives et plus rentables que leurs

[1] McKinsey Global Institute, mai 2011. www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[2] «Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond», communiqué de presse Gartner, 1er décembre 2011. www.gartner.com/it/page.jsp?id=1862714

[3] Experts MIT Sloan : Commentaires sur les enjeux du monde de l'entreprise d'aujourd'hui, 14 février 2012. www.mitsloanexperts.com/2012/02/15/erik-brynjolfsson-on-big-data-a-revolution-in-decision-making-improves-productivity

concurrents, à hauteur de 5 % environ. Il en conclut qu'« il existe beaucoup d'opportunités très faciles à saisir pour les entreprises capables d'exploiter les Big Data à leur avantage. »[3]

Un impératif : l'innovation

Les responsables informatiques d'aujourd'hui sont mis au défi de développer des systèmes permettant d'analyser les Big Data, pour aider leurs entreprises à prendre des décisions mieux informées. Les Big Data sont un domaine récent dont les intervenants et les meilleures pratiques ne sont pas encore bien établis. Les entreprises capables de proposer des services de formation, conseil et autres en matière de Big Data contribueront à combler cette lacune.

Pour notre part, cela fait environ deux ans que nous avons commencé à exploiter nos Big Data. Nos efforts dans ce domaine sont un élément clé d'une feuille de route globale qui vise à transformer l'activité d'Intel par des pratiques analytiques avancées.

LA SOLUTION

Intel IT conduit actuellement plusieurs études de validation technique, visant à démontrer l'efficacité des Big Data pour résoudre des problématiques métier à forte valeur.

Sur la base de nos recherches et des méthodes courantes de notre secteur, nous avons commencé en 2012 à déployer en interne une plateforme Big Data hybride et rentable. Elle devait en particulier répondre aux critères suivants :

- Entrepôt de données SQL à architecture de traitement massivement parallèle (MPP)
- Utilisation de Hadoop*, pour le traitement distribué des grands ensembles de données sur des clusters d'ordinateurs

Plateforme de gestion de base de données MPP

Les solutions classiques d'analyse décisionnelle utilisent des systèmes d'entrepôt de données conçus pour le traitement, et non l'analyse, des transactions en ligne. Ces systèmes sont construits avec des bases de données, serveurs et plateformes de stockage à usage généraliste, n'offrant pas la spécialisation nécessaire pour traiter des téraoctets (To) de données qui évoluent et se développent en permanence.

À l'inverse, les plateformes MPP d'aujourd'hui ne sont plus limitées à la programmation SQL, et il n'est pas rare qu'elles prennent en charge le développement dans des langages de programmation tels que Java*, C/C++ et R. Lorsqu'elles intègrent des fonctionnalités d'analyse riches et poussées, ainsi que des capacités d'exploration au sein même des bases de données, ces plateformes offrent toute la souplesse nécessaire pour mobiliser les compétences BI et SQL existantes et acquérir de nouvelles compétences, basées sur le langage de programmation le plus approprié compte tenu du but poursuivi.

Nous avons opté pour une solution basée sur un entrepôt de données tiers, doté d'une architecture massivement parallèle, avec une approche asymétrique, qui permet d'analyser de très gros volumes de données jusqu'à 100 fois plus rapidement qu'avec des systèmes traditionnels. Ces systèmes sont aujourd'hui proposés par un certain nombre de constructeurs.

Nous avons basé la sélection de notre solution sur plusieurs facteurs.

- Rentabilité (coût optimal compte tenu de la performance et des capacités)
- Simplicité et rapidité du délai de rentabilisation
- Évolutivité du stockage et de la performance (To à Po)
- Outils d'analyse avancés et prise en charge intégrale du langage de programmation statistique open source R
- Capacité d'intégration globale avec notre écosystème BI existant
- Interopérabilité avec un écosystème d'entreprise

Spécifiquement développé pour l'analyse, ce système associe une technologie de filtrage de données à des serveurs lames basés sur la famille de processeurs Intel® Xeon® E7 et des disques durs conventionnels, délivrant ainsi une excellente performance Big Data pour un coût et des besoins en maintenance limités. Des serveurs lames pourront être ajoutés au système, garantissant ainsi son évolutivité en termes de performance et de capacité. Chaque lame est connectée à plusieurs disques durs installés à proximité les uns des autres qui lisent les flux de données en parallèle, réduisant ainsi considérablement les latences d'accès par rapport aux solutions qui utilisent des systèmes de stockage séparés.

Notre processus de sélection comprenait notamment une étude des caractéristiques techniques proposées par onze fournisseurs d'entrepôts de données, suivie d'un appel d'offres formel auprès de cinq d'entre eux. Un fournisseur a finalement été choisi au terme des analyses conduites dans le cadre de l'appel d'offres et d'une évaluation technique.

Hadoop

Hadoop est une infrastructure open source dédiée au traitement des gros volumes de données. Au lieu d'utiliser un superordinateur unique, Hadoop coordonne le stockage et le calcul entre une multiplicité de serveurs qui agissent comme un cluster, chaque serveur fonctionnant avec un sous-ensemble des données.

Hadoop est un projet open source de haut niveau, développé par l'Apache Software Foundation. De nombreuses distributions commerciales sont également disponibles.

Hadoop opère comme un système d'exploitation pour le traitement distribué, en assurant deux services de base :

- **Système de fichiers distribué Hadoop.** Ce système de fichiers distribué fournit une architecture de stockage de type UNIX*, distribuée sur tous les nœuds du cluster Hadoop. Hadoop peut également utiliser d'autres systèmes de fichiers.
- **MapReduce.** Cette fonction de traitement distribué est au cœur de Hadoop. MapReduce coordonne chacun des serveurs au sein du cluster de façon à ce qu'il opère en parallèle pour une partie de la tâche de traitement globale.

Ce noyau est complété par de nombreuses applications, boîtes à outils et couches de données commerciales et open source, notamment :

- **Hive.** Langage SQL permettant d'interroger les données Hadoop
- **HBase.** Base de données de lecture/écriture à haute vitesse, orientée colonnes, capable de gérer des milliards de lignes et des millions de colonnes
- **Pig.** Environnement de gestion de scripts interactive pour le traitement des données
- **Mahout.** Bibliothèque d'apprentissage automatique offrant des algorithmes de groupement et des fonctions de filtrage collaboratif et de reconnaissances des similitudes

- **Sqoop.** Système d'échange par importation/exportation avec les bases de données SGBDR
- **Oozie.** Environnement de workflow pour la coordination des opérations complexes de traitement de données
- **Cassandra.** Base de données orientée documents

Hadoop intègre des fonctionnalités d'extrapolation linéaire. En doublant par exemple le nombre de machines sur un cluster, on pourra réduire de moitié le temps de traitement (ou traiter deux fois plus de données pour une période donnée).

Hadoop est écrit en Java et fonctionne sous Linux*. Les applications Hadoop sont généralement écrites en Java, mais d'autres langages peuvent également être utilisés. Certains outils Hadoop, notamment Hive et Pig, sont exécutés sur un ordinateur client et génèrent les programmes MapReduce à la volée.

Dans la mesure où Hadoop agrège le stockage de tous les serveurs sur son cluster, et où ces serveurs peuvent utiliser des disques durs conventionnels, le coût de stockage par To est très faible et le cluster peut prendre en charge des volumes de données se mesurant en pétaoctets. Hadoop permet donc de capturer et conserver, de façon rentable, des données qui auraient auparavant été supprimées. Cette solution permet également de capturer et stocker des données qui ne sont pas encore bien comprises, mais qui sont susceptibles de recéler de la valeur. Pour des domaines tels que l'analyse de texte, il a été démontré qu'il est préférable de disposer d'un maximum de données pour obtenir des résultats optimaux, même lorsqu'on utilise des algorithmes simples. Dans des domaines tels que la cyber-sécurité, les capacités massives de Hadoop permettent de conduire les analyses sur des périodes plus longues.

De manière générale, Hadoop et ses technologies connexes ne visent pas à remplacer les systèmes de traitement transactionnel en ligne ou autres systèmes SGBDR traditionnels. La force de Hadoop réside dans sa capacité à mener des traitements par lots sur de très gros volumes de données (mesurés en To ou Po).

Avantages des plateformes hybrides

En associant un entrepôt de données tiers, avec ses éléments d'architecture massivement parallèle et asymétrique, et Hadoop (voir Figure 1), nous avons pu mettre en place une plateforme Big Data rentable, hautement évolutive, et qui exploite au mieux les forces spécifiques de chacun de ses composants. Les composants sont colocalisés et connectés au sein d'un réseau et d'un chargeur de données à haut débit, permettant ainsi à la plateforme Big Data de transférer plus efficacement les portions de données entre les différents emplacements, au gré des besoins.

Renforcement des compétences et de l'expertise en matière de Big Data

L'un des plus grands défis posés par les Big Data est la nécessité de remédier à la pénurie de compétences d'expert. Selon le rapport McKinsey Global Institute cité plus tôt, les États-Unis pourraient à eux seuls, d'ici à 2018, se retrouver confrontés à une pénurie de 140 000 à 190 000 personnes dotées de compétences analytiques approfondies, et de 1,5 millions de managers et analystes possédant les connaissances requises pour exploiter les Big Data afin de prendre des décisions efficaces.

Si les compétences requises sur le volet scientifique (statistiques, mathématiques, apprentissage machine et analyse visuelle, notamment) sont indispensables, le savoir-faire nécessaire pour faire le rapprochement entre les données et la réalité métier de l'entreprise, puis pour transformer ces informations en retombées commerciales tangibles, l'est tout autant. Pour cela, les clients informatiques tels que les entités opérationnelles d'Intel doivent développer à l'interne des compétences poussées sur le plan de l'utilisation des Big Data, si elles veulent pouvoir véritablement exploiter cette ressource.

De nombreuses technologies Big Data, telles Hadoop, sont développées en open source par ces entreprises Internet pour traiter de gros volumes de données structurées et

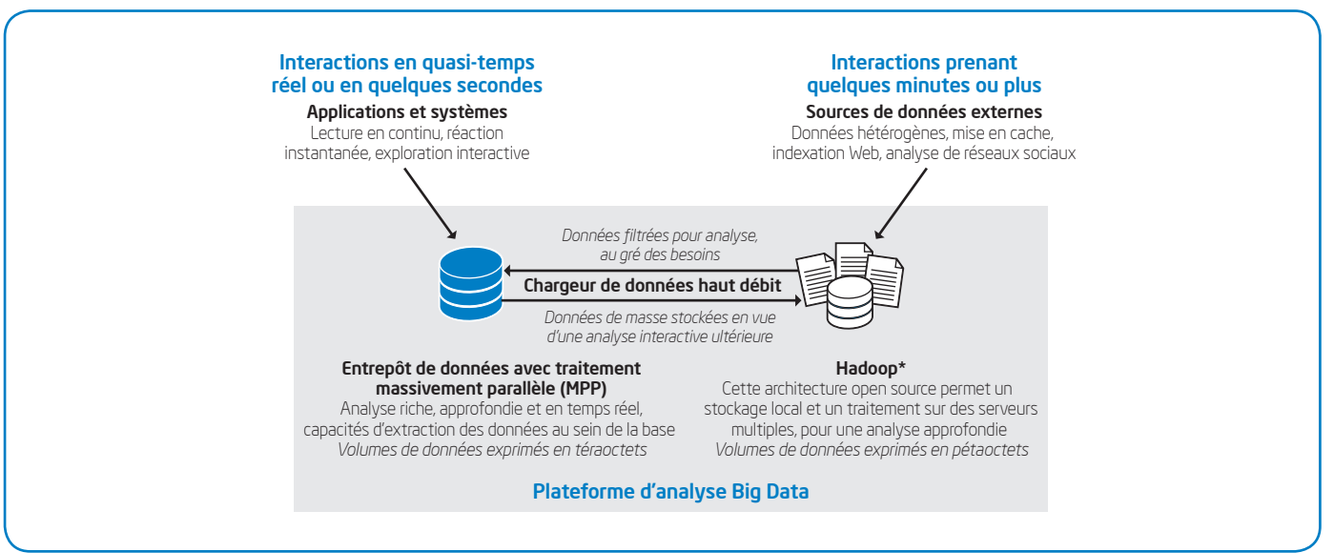

Figure 1. La plateforme Big Data d'Intel IT dédiée à la business intelligence est fondée sur l'association entre un entrepôt de données à architecture de traitement massivement parallèle (MPP) et des clusters de serveurs conventionnels fonctionnant sous Apache Hadoop*.

non structurées de manière rentable. Ces technologies gagnent rapidement en maturité, mais elles exigent actuellement des compétences techniques approfondies dans des domaines tels que Linux, le développement Java et le traitement distribué. Pour déployer leurs technologies Big Data, les entreprises devront également développer ces compétences, et bien d'autres encore.

COMBLER LES LACUNES SUR LES CONNAISSANCES ET LES COMPÉTENCES

L'acquisition de nouvelles compétences peut être plus difficile encore que la mise en œuvre d'une technologie. Intel IT et les entités opérationnelles Intel s'attachent à combler ce manque d'expertise Big Data en déployant des programmes de recherche et de formation, en embauchant des personnes disposant des compétences clés pour travailler sur les Big Data, en conduisant des essais et validations techniques (telles que celles mentionnées dans le présent article) et en mettant en œuvre des cas d'utilisation spécifiques.

En l'absence d'expérience directe, c'est en travaillant avec les données et en effectuant des analyses Big Data que l'on parviendra à acquérir les compétences et le savoir-faire requis. Nous avons la conviction que le résultat final, à savoir la capacité à exploiter les Big Data pour fournir une BI prédictive et prescriptive, justifie absolument les investissements consentis en matériel, logiciels, formation et temps de travail.

ÉTUDES DE VALIDATION TECHNIQUE

Intel conduit actuellement de nombreuses études de validation technique consacrées au Big Data. Nous présentons quatre de ces études dans cette section.

Détection des malwares

Les cyber-menaces continuent à s'aggraver, avec la sophistication croissante des auteurs de ces attaques et de leurs outils. L'objectif de la détection de sécurité est de découvrir les menaces à temps afin de permettre aux utilisateurs et aux responsables de la sécurité de prendre les mesures qui s'imposent.

Les fichiers de signature, qui étaient par le passé la méthode de défense la plus commune contre les logiciels malveillants, ont perdu en efficacité, peinant à suivre le rythme d'apparition des nouveaux malwares. Une approche plus performante consiste à conserver une longueur d'avance sur ces logiciels malveillants, en menant une analyse approfondie sur le mode d'action du malware et sa provenance, et en cherchant à prédire d'où les nouvelles attaques seront susceptibles de provenir à l'avenir. Ce travail approfondi de surveillance et de prédiction exige de mener une veille permanente sur l'activité des serveurs, afin de détecter toute anomalie, cela à tous les niveaux : système, réseau et application.

Les signes avant-coureurs permettant d'identifier de telles menaces sont souvent cachés au sein des divers logs de réseaux et serveurs (proxy, DNS, DHCP, VPN, etc.), qui recèlent des quantités potentiellement énormes de données. Les anomalies peuvent prendre une multiplicité de formes, s'agissant par exemple de signatures caractéristiques, ou encore de tendances générales permettant de soupçonner une activité ou un comportement suspect (par ex. : communication utilisant des URL identifiées comme douteuses ou inhabituelles). Le travail d'analyse nécessite une séquence d'étapes complexe, comprenant notamment la corrélation des données issues d'un grand nombre de

Défis et opportunités du Big Data

Le terme Big Data désigne, dans ses grandes lignes, les problématiques informationnelles sur lesquelles butent les approches traditionnelles, fondées sur les bases de données relationnelles, cela en raison des volumes de données ou de la grande diversité des sources et types de données. Cette diversité des données (texte, audio, vidéo, parcours de clics, logs, etc.) est l'autre grande caractéristique des Big Data ; c'est d'ailleurs pourquoi l'on peut parler de Big Data avec seulement quelques téraoctets de données non structurées.

Les fournisseurs de moteurs de recherche Internet, tels Google et Yahoo, ont été parmi les premiers à développer des outils Big Data, indispensables pour indexer le World Wide Web. D'autres entreprises Internet ont rapidement suivi, développant d'autres composantes dédiées à la gestion des commandes et recommandations, aux messages de type Facebook et à d'autres problèmes touchant Internet dans son ensemble. Aujourd'hui, les services informatiques des entreprises appliquent ces mêmes outils pour résoudre des problématiques métier à haute valeur ajoutée, qui étaient jusqu'à présent difficiles à analyser et résoudre.

L'application de ces nouvelles techniques n'est cependant pas toujours simple. Des difficultés considérables peuvent émerger sur les plans de l'intégration, du déploiement et de la maintenance de ces nouveaux outils, dont la plupart ne sont pas encore arrivés à maturité et nécessitent de nouvelles compétences informatiques sur Linux* et Java*. Le développement et l'optimisation d'une solution Big Data obligent à repenser le problème en termes d'infrastructures de traitement parallèle de type MapReduce (attention cependant, tous les problèmes ne se prêtent pas au traitement parallèle), ou encore d'indexation Web. La résolution de problèmes à grande échelle peut par ailleurs nécessiter une certaine relaxation de la sémantique ACID (atomicité, cohérence, isolation et durabilité), sur laquelle les programmeurs de base de données se reposent souvent, ainsi qu'un arbitrage entre faible latence et débit élevé.

Si les systèmes de données existants s'avèrent suffisants, il n'est pas forcément nécessaire de les changer. Les solutions Big Data peuvent en revanche être la réponse à des problèmes qui étaient jusqu'à présent hors de portée. Même si une justification claire n'a pas encore été établie, les entreprises peuvent tirer parti des faibles coûts associés au stockage des Big Data pour capturer et stocker la quasi-totalité des données générées par leur activité, pour en extraire plus tard la valeur latente.

sources, ainsi que l'établissement d'un niveau de référence permettant de définir une activité normale, ou encore la définition de motifs récurrents ou modèles d'utilisation permettant de détecter toute activité anormale.

Pour identifier suffisamment tôt ces anomalies, Intel déploie des technologies Big Data pour collecter les données brutes et non structurées et les structurer, puis appliquer des modèles statistiques (analyse prédictive, notamment) pour détecter toute tendance anormale dans l'activité.

Ces validations techniques doivent nous permettre de parvenir à une identification en temps réel de tels comportements, de sorte que les logiciels malveillants puissent être rapidement identifiés et contenus. La capacité à collecter et analyser des données amassées sur plusieurs mois, voire des années, nous permettra de mieux prédire les sources et la nature des menaces, et ainsi de déployer des mesures et systèmes de prévention plus efficaces.

Validation de conception des puces

La conception des puces nécessite une grande quantité de tests avant qu'un produit puisse être réalisé en silicium. Ces tests se poursuivent au cours des différentes phases d'implémentation du silicium, à l'aide de centaines de capteurs qui collectent des données à des taux d'échantillonnage très élevés (plusieurs milliers de fois par seconde). Ces tests très complets génèrent d'énormes quantités de données.

Dans cet essai de validation technique, Intel IT a étudié comment une plateforme Big Data pouvait être utilisée pour optimiser le processus de validation, en permettant l'analyse de milliards de données structurées et non structurées, l'objectif étant d'accélérer le processus de conception et de raccourcir les délais de mise en production (et à terme, les délais de mise sur le marché).

Un bon exemple d'un tel modèle d'utilisation est une pratique que nous appelons la couverture. Dans le domaine de la validation post-silicium, il n'existe pas de règles absolument claires permettant de déterminer précisément le moment où une puce est effectivement prête à être mise en production. D'un côté, la sortie commerciale d'une puce comportant des bugs peut sérieusement nuire à la réputation de l'entreprise. De l'autre, en revanche, des tests excessifs risquent de retarder la sortie commerciale de la puce, entraînant alors des pertes de ventes non réalisées se chiffrant en millions de dollars.

Le concept de couverture vise à éviter ces cas extrêmes. En collectant des données sur les états logiques et physiques dans lesquels le processeur a été testé (ou « couvert »), nous pouvons mieux comprendre la qualité des essais et des outils déployés pour ceux-ci, et ainsi déterminer si la puce est ou non prête à être mise en production.

L'analyse Big Data peut également être utile pour le processus de débogage, en regroupant et en triant automatiquement les défauts identifiés, ainsi qu'en permettant une analyse des causes profondes sur de très gros volumes de données, associés à tout l'historique de tests. En conduisant ainsi une analyse approfondie sur de grandes quantités de données collectées (et non pas seulement sur des échantillons), nous pouvons dresser une image plus précise des progrès réalisés à chaque étape et identifier les moyens d'améliorer et de rationaliser nos processus de conception. Cela permet également, à terme, d'améliorer le produit lui-même.

Information commerciale stratégique

Pour une entreprise comme Intel dont les produits se vendent sur les cinq continents et qui doit gérer une chaîne d'approvisionnement mondiale, il est primordial d'être en mesure d'anticiper l'évolution des conditions du marché et de faire des prévisions précises sur ce qui pourra arriver à une échéance d'un mois, de six mois, voire de cinq ou dix ans. Les grandes entreprises multinationales doivent trier d'énormes quantités de données, sur des éléments aussi divers que les tendances climatologiques, les données macroéconomiques, les forums de discussion, les sites d'actualité et les réseaux sociaux ou encore les wikis, tweets et autres blogs. Elles pourront, à partir de ces données, établir des projections précises, planifier leurs stratégies commerciales, évaluer les menaces de la concurrence, anticiper les changements de comportement des consommateurs, renforcer leurs chaînes d'approvisionnement, ou encore renforcer leurs plans de continuité d'activité.

Nous travaillons, dans le cadre de ces validations techniques, avec les entités opérationnelles Intel pour analyser des données issues d'une grande variété de sources disparates, avec les objectifs suivants :

- Améliorer nos prévisions sur les ventes potentielles dans les différentes régions du monde, affiner les niveaux de production et communiquer des prévisionnels plus précis à nos actionnaires
- Construire et tester des scénarios basés sur des événements potentiels d'envergure mondiale, afin de déterminer leur impact sur nos marchés, sur nos chaînes d'approvisionnement, et sur notre capacité à répondre à la demande et aux menaces concurrentielles
- Découvrir de nouveaux utilisateurs et de nouveaux modes d'utilisation de nos produits

Système de recommandations

Croulant sous la croissance exponentielle des contenus, les utilisateurs ont bien souvent besoin d'aide pour trouver les informations qui correspondent le mieux à leurs demandes et à leurs intérêts. Cela induit une demande accrue pour les services basés sur la recommandation au sein du groupe Intel, pour des applications internes comme externes. Les systèmes de recommandation, analogues à ceux proposés par Amazon et Netflix, aident les utilisateurs en réduisant les temps de recherche et de navigation et en permettant des résultats plus personnalisés et ciblés. Cela permet d'améliorer la productivité, la crédibilité et l'expérience globale de l'utilisateur.

L'implémentation d'un système de recommandations évolutif nécessite une expertise en matière d'analyse prédictive et de Big Data, dans la mesure où cela implique l'exécution d'algorithmes complexes et très exigeants en ressources sur des gros volumes de données historiques.

Cette validation technique portait sur un moteur de recommandation générique et réutilisable, doté d'une architecture à deux couches (en ligne et hors ligne) recouvrant notre plateforme Big Data. La composante hors ligne est un processus discontinu de traitement par lots, qui exécute le cœur de l'algorithme de recommandation. Cela garantit la capacité de nos modèles à prendre en charge le traitement des Big Data dans le cadre d'un environnement évolutif. La composante en ligne opère comme une couche de service, permettant de répondre à chaque demande de service. Elle charge les calculs intermédiaires correspondants, effectués lors de la phase hors ligne, puis effectue la dernière étape de l'algorithme, à savoir la génération de la recommandation elle-même. Elle applique par ailleurs une logique de configuration en fonction du contexte, en ajustant la recommandation finale au contexte de la requête.

L'évolutivité de la solution est assurée par la mise en œuvre du cœur des algorithmes avec

Mahout. Mahout est une bibliothèque open source dédiée à l'exploration de données, écrite en Java par-dessus Hadoop. Elle tire parti de l'architecture Hadoop en exécutant des tâches parallèles au sein d'un cluster d'équipements conventionnels, dans un environnement non partagé (*shared-nothing environment*). Tous les résultats intermédiaires sont écrits sur la plateforme SGBDR MPP, permettant ainsi la récupération rapide de la composante en ligne.

Le déploiement de ce service de recommandation sera un facteur de facilitation clé qui permettra de fournir un contenu personnalisé, dans une optique « juste-à-temps ». Cela nous permettra de renforcer la productivité des employés lorsqu'ils utilisent des applications Intel internes. Cela nous permettra également d'aider nos clients externes à mieux choisir nos produits, contribuant ainsi à notre chiffre d'affaires. Nous pourrons ensuite appliquer l'expérience et les connaissances acquises au fil de ces analyses prédictives complexes à de grands volumes de données, de façon à approfondir le développement de ces solutions à l'avenir.

CONCLUSION

Intel IT adopte une approche systématique, en intégrant l'analyse Big Data à l'ensemble de ses efforts BI. Cette approche a été engagée avec plusieurs travaux de validation technique en 2012. En développant ses capacités à extraire et analyser les Big Data, Intel veut faire évoluer ses fonctions BI, qui doivent passer de l'analyse descriptive à l'analyse prédictive et prescriptive, ce qui permettra une compréhension plus riche et plus profonde des modèles et tendances de son activité.

Nous avons désormais achevé la première étape, qui consistait à concevoir et construire une plateforme Big Data associant un entrepôt de données tiers à Hadoop, une architecture open source dédiée au traitement de très gros volumes sur plusieurs serveurs. Cette solution nous permet d'effectuer un MPP sur des données structurées et de procéder au traitement distribué de grandes quantités de données sur des serveurs standard. Nous renforçons par ailleurs en interne nos compétences, notre expertise et notre sophistication en matière de Big Data, au sein à la fois de notre équipe BI et des entités opérationnelles.

Une fois achevées les validations techniques, Intel mettra sa plateforme Big Data en production et la mobilisera pour résoudre des problématiques métier à forte valeur ajoutée et ainsi offrir une nouvelle efficience opérationnelle, accroître les sources de revenus existantes et en créer de nouvelles. Notre programme d'analyse Big Data va continuer à se développer au cours des prochaines années, fournissant ainsi à Intel une BI véritablement exploitable et qui permettra de générer de nouveaux avantages concurrentiels, notamment dans les domaines de la fabrication, de la sécurité, du marketing et de l'informatique.

POUR PLUS D'INFORMATIONS

Rendez-vous sur www.intel.com/it pour consulter nos livres blancs sur des sujets connexes :

- «Roadmap for Transforming Intel's Business with Advanced Analytics»

CONTRIBUTEURS

Jessica Brindle, planification stratégique BI, Intel IT

ACRONYMES

BI	Business Intelligence
MPP	*massively parallel processing* (traitement massivement parallèle)
Po	pétaoctet
SGBDR	Système de gestion de base de données relationnelle
To	téraoctet

Pour en savoir plus sur les, meilleures pratiques d'Intel IT, rendez-vous sur www.intel.com/it.

 Imprimé aux États-Unis ♻ Pensez à recycler le papier 0712/WWES/KC/PDF 327474-001US