INTERVIEW **CROISÉE**

Philippe NIEUWBOURG
Analyste indépendant spécialisé en informatique décisionnelle, journaliste et auteur, fondateur de decideo.fr

Franck LEONARD
Senior Solution Consultant, spécialisé dans les projets analytiques et décisionnels au sein de TIBCO Software.

Le data lake en question

Positionnement, maturité, scénarios d'usage, solutions requises. Deux experts confrontent leurs points de vue.

TIBCO®

Le Data Lake en une phrase

Le data lake désigne un référentiel au sein duquel des données de nature variée (fichiers CSV, XML, JSON, logs, emails, PDF...) sont stockées en volume et dans leur état brut sans qu'un schéma d'organisation préalable ne leur soit imposé.

- Le data lake est-il un concurrent du data warehouse ?
- Pourquoi connaît-il un tel intérêt ?
- Quels sont ses champs d'application de prédilection ?
- À quoi ressemble une solution d'analyse adaptée au data lake ?

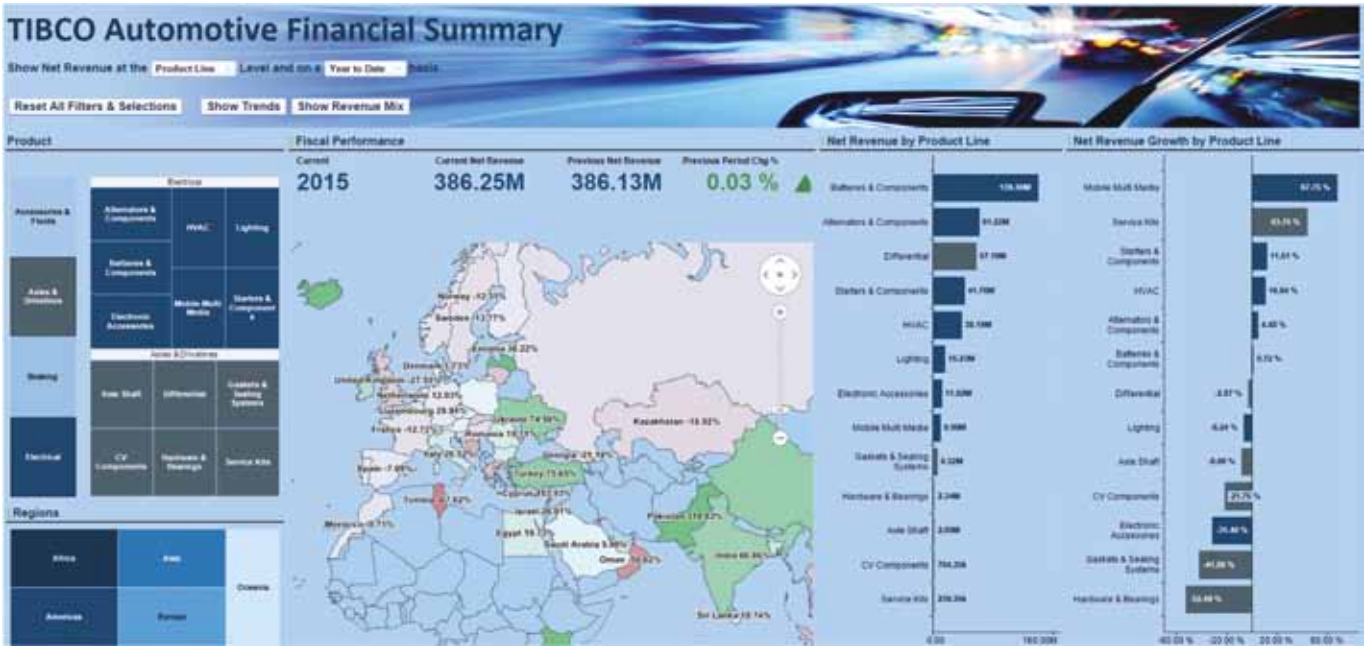
Réponses à deux voix avec Philippe Nieuwbourg et Franck Léonard.

Commençons par mettre de côté les malentendus : le data lake est souvent opposé au data warehouse. À raison ?

PhN Pas vraiment. Le data lake n'a clairement pas vocation à remplacer le data warehouse. Pour expliquer la différence, j'aime bien filer la métaphore du magasin Ikea™. Généralement, quand je m'y rends, je sais que je veux acheter une chaise d'un modèle précis, vu préalablement dans le catalogue, et je sais dans quelle allée la trouver. Le magasin est organisé pour ce type de recherche. En revanche, si j'arrive chez Ikea™ et que je demande à voir tous les produits blancs, ma question restera sans réponse, car le magasin n'est pas organisé pour une recherche par couleur. Le magasin Ikea est donc conçu comme un data warehouse, il est structuré pour des requêtes que l'on connaît à l'avance.

Au contraire d'un data lake, destiné à accueillir des données sans organisation prédestinée. L'un et l'autre sont donc conçus pour des usages bien distincts. Techniquement, je peux mettre ma comptabilité dans un data lake mais, en termes d'usage, c'est… ridicule.

FL Data warehouse et data lake vont en effet cohabiter encore longtemps, de la même manière que la business intelligence traditionnelle, avec ses reportings et tableaux, reste pertinente aux côtés d'une analyse plus visuelle et interactive. De nombreuses entreprises – prenez simplement le cas des banques – ont besoin de stocker et de manipuler de grands volumes de données structurées qui ne sont pas près de quitter le data warehouse.



TIBCO Spotfire propose l'usage de stratégies d'accès aux données adapté aux environnements big data et data lake et synthétise visuellement le résultat : l'utilisateur dispose de la même agilité quelle que soit la source de données.

Pourquoi alors cette effervescence autour du data lake ? Faut-il l'expliquer par l'émergence de nouvelles technologies ?

PhN Hadoop, la technologie la plus associée au data lake, est âgée de 15 ans, on ne peut donc pas dire qu'elle soit très nouvelle. La raison de cet intérêt est plutôt à chercher du côté des usages. Les analyses de données externes se sont fortement accrues. Or, ces données viennent des réseaux sociaux ou de l'internet des objets (IoT). Des sources dont on contrôle beaucoup moins le format des données. Le data lake est un bon candidat pour analyser ces nouveaux flux.

FL Les usages expliquent clairement l'actualité des data lakes. Philippe évoque l'IoT. On peut aussi ajouter par exemple le développement des DMP (Data Management Platform), qui entendent agréger des données de sources très variées pour réussir à reconstituer le parcours omnicanal des clients et prospects. Donc, oui, le développement de nouveaux usages légitime le recours au data lake, mais ce n'est pas le seul facteur. N'oublions pas la question du coût. Avec un data lake, le coût de possession de la donnée est 15 à 20 fois inférieur à celui du data warehouse. En outre, il est aisé avec ce genre de solution de commencer petit pour croître à son rythme, cela grâce à l'élasticité du cloud notamment.

PhN C'est vrai, le stockage avec le data lake ne coûte pas grand chose. Je crois aussi qu'il favorise un comportement très humain – la procrastination – qui renvoie à plus tard l'effort de structuration et d'organisation. Il faut se méfier de cette tendance. Ce n'est pas parce que je peux tout stocker en vrac que je dois le faire…

À vous écouter, on a le sentiment que le data lake cristallise une querelle entre les anciens et les modernes…

PhN Oui, c'est vrai, cette opposition existe. Ceux de l'ancienne école, les statisticiens, sont perturbés par ces data lakes avec lesquels, en revanche, les data scientists sont pleinement à l'aise. Je suis convaincu que pour faire mûrir les projets il faut sortir de cette querelle et éviter les extrêmes, c'est-à-dire la surcharge du data lake avec ces données qui n'ont rien à y faire.

FL Pouvoir stocker de la donnée, sans savoir exactement ce que l'on en fera, juste dans la perspective de demandes futures que l'on ne connaît pas encore, cela change tout. Il faut par exemple repenser le modèle budgétaire du stockage des données. Ne le nions pas, une petite révolution organisationnelle et culturelle se dessine. Mais je suis d'accord avec Philippe : un stockage non raisonné dans le data lake peut créer des problèmes de gouvernance de la donnée, avec des coûts cachés possibles en retour. Voilà pourquoi j'ai la conviction que nous allons gagner en rigueur dans l'alimentation du data lake.

Aujourd'hui, concrètement, quelle est la réalité des projets ? Où en sont les entreprises ?

PhN Pour l'heure, on voit surtout des projets de data lake pour stocker les données, moins pour les analyser. Ce n'est pas surprenant, c'est le début logique de l'histoire. En outre, la collecte des données n'est pas aussi simple qu'on le croit. Nous manquons encore de carburant. Des problèmes légaux se posent aussi et peuvent freiner ces projets, notamment en Europe, où il est difficile de démarrer même un simple PoC (Proof of Concept) sans avoir vérifié préalablement la conformité réglementaire.

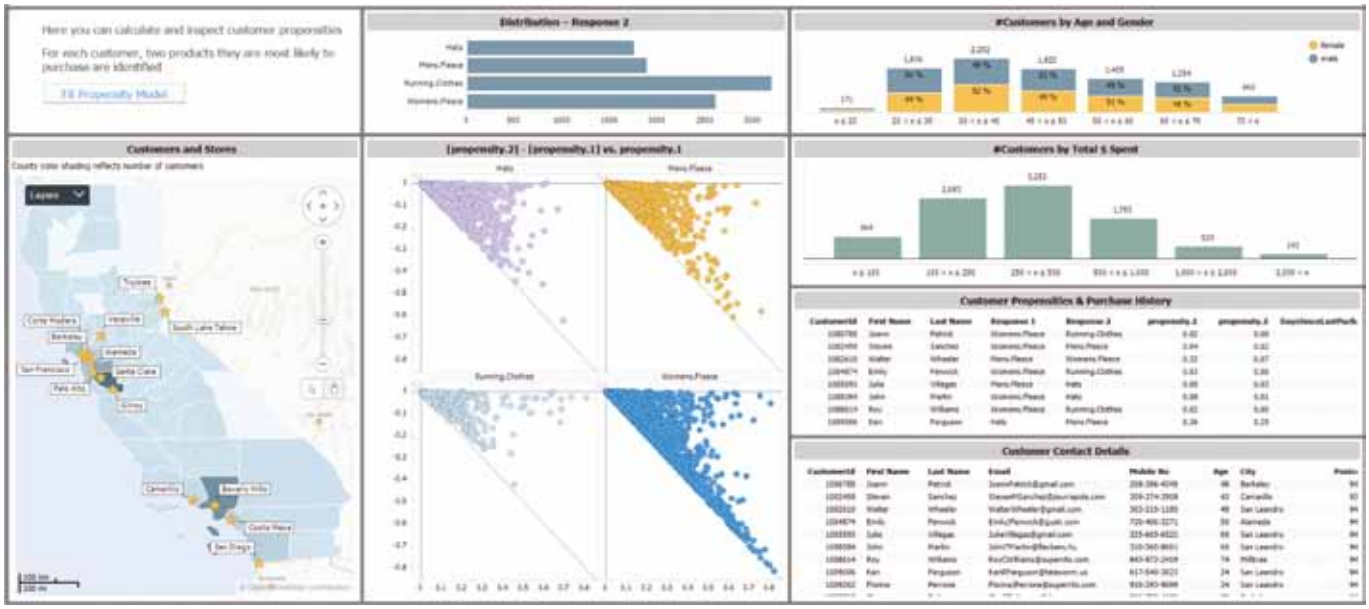
FL Les données ne sont en effet pas toujours aussi disponibles qu'on le souhaiterait, mais les cas d'usage se multiplient, et dans des univers métiers très variés. Des acteurs pétroliers réduisent quotidiennement leurs coûts en analysant en temps réel les informations émises par leurs équipements de pompage, les banquiers répondent aux besoins de traçabilité demandés par une norme telle

que BCBS 239 en comparant à grande échelle les versions de leurs données, les services marketing engrangent davantage de données pour des analyses plus détaillées de leurs campagnes… Ce que l'on voit émerger, c'est un vrai besoin d'analyse et de modélisation d'événements en temps réel.

Pour analyser les données stockées dans ces data lakes, de nouvelles solutions s'imposent ?

PhN Oui, clairement. Les outils traditionnels de business intelligence sont conçus pour travailler sur des univers structurés et se montrent peu adaptés aux data lakes. Pour ces derniers, il faut des outils plus agiles qui se prêtent à l'interactif et à l'itératif pour créer un vrai dialogue entre les utilisateurs et les données. C'est là qu'une solution telle que TIBCO Spotfire se démarque, dans l'autonomie donnée aux utilisateurs.

FL Créer un pont entre le data lake et les utilisateurs, c'est la meilleure manière de résumer la vocation de TIBCO Spotfire. Avec des fonctions comme le data wrangling, Spotfire permet aux utilisateurs d'explorer, de nettoyer et d'organiser les données eux-mêmes. Cette exploration des données suit un processus simple et très visuel. Et avec la datavisualisation, sujet sur lequel Spotfire est fort apprécié, notamment grâce à ses capacités de représentation des données sur des cartes très interactives, nous combinons l'avantage du self-service avec la puissance du data lake.



Malgré l'augmentation du volume de données à analyser, il faut prendre des décisions de plus en plus rapidement : TIBCO Spotfire répond facilement à ce paradoxe en associant stratégie d'accès aux données, logique calculatoire et visualisation en libre service.

À quoi pourrait ressembler la suite de l'histoire, ou du moins une évolution souhaitable du data lake ?

PhN Ce que l'on peut souhaiter, c'est que les projets privilégient un usage raisonné du data lake, lui attribuent un rôle adapté à sa nature et optimisent son usage en conséquence. Cela pourrait passer par le fait de structurer, même a minima, les informations versées dans le data lake, afin de gagner du temps ensuite.

FL Nous n'en sommes qu'au tout début de l'histoire. Avec la croissance de l'IoT, les données froides et chaudes (temps réel) vont être de plus en plus mélangées, il va donc falloir repenser les modèles d'analyse mais aussi optimiser en conséquence l'accès aux informations. C'est ce que nous faisons d'ores et déjà avec TIBCO Spotfire Data Catalog qui, le machine learning aidant, intègre des capacités innovantes d'apprentissage automatique pour rendre la navigation dans les données d'entreprise aussi simple que la recherche de contenus sur le Web. Quant au data lake, lui aussi va encore connaître de fortes évolutions.

EN SAVOIR PLUS

TIBCO® Spotfire®

http://spotfire.tibco.com/fr/

À propos **des experts**

philippe.nieuwbourg@decideo.com

Philippe Nieuwbourg

Analyste indépendant, Créateur de Decideo

Analyste indépendant, journaliste et auteur, Philippe Nieuwbourg cumule 20 ans d'expérience dans les domaines de l'informatique décisionnelle et de l'analyse de données. Fondateur du site decideo.fr, il analyse au quotidien les tendances et actualités de l'intelligence d'affaires. Philippe est également membre du BBBT (Boulder BI Brain Trust), un groupe informel d'analystes indépendants créé à l'initiative de Claudia Imhoff.

Chargé de cours dans plusieurs universités et écoles en France et au Canada, il a été parmi les premiers à mettre en lumière la valeur créée par la combinaison de la visualisation de données (datavis) et de la mise en récit (storytelling), au travers d'une nouvelle approche, la mise en récit des données (data storytelling).

fleonard@tibco.com

Franck Leonard

Senior Solutions Consultant, TIBCO Software

Franck Leonard travaille depuis 1991 dans l'édition de logiciels. Il a connu différents services et rempli plusieurs missions depuis la réalisation de projet jusqu'à la gestion de data center. Afin de favoriser l'efficience et le pragmatisme, il a activement participé à la démocratisation de solutions cloud, d'ordonnancement et de distribution logicielle.

Depuis 2 ans, il a rejoint le monde analytique. Il intervient dans de nombreux projets de data visualisation, de statistique descriptive et d'analyse prédictive notamment dans le cadre d'initiatives big data pour le compte de TIBCO Software.

TIBCO Software France
25 rue Balzac
75008 Paris
+33 1 44 51 45 88
www.tibco.com

TIBCO Software accompagne les entreprises dans leur transformation digitale en interconnectant les différents éléments de leurs systèmes d'information et en augmentant l'intelligence des décisions basées sur la donnée. Cette combinaison garantit des réponses plus rapides, de meilleures décisions et des actions plus pertinentes. Depuis 20 ans, TIBCO met à la disposition des solutions et services innovants qui s'adressent aux opérationnels, aux développeurs ainsi qu'aux data scientists. Des milliers de clients de par le monde se démarquent en faisant confiance à TIBCO pour concevoir des business modèles innovants et offrir des expériences clients convaincantes. Pour en savoir plus www.tibco.com.