CAS CLIENT

Neopost s'appuie sur HDP pour assurer le bon fonctionnement de son parc machines

Interview d'Hugues le Bars,
Chief Data Officer chez Neopost

HORTONWORKS®

Dans cet Interview, Hugues Le Bars, Chief Data Officer chez Neopost (leader mondial des solutions de communication d'entreprise) nous fait part de son expérience de la création d'un data lake avec la plateforme Hortonworks HDP. Méthodologie, facteurs de succès, mise en œuvre dans un contexte de maintenance prédictive du parc machines de Neopost… Au-delà de son approche terrain, Hugues Le Bars expose sa vision des enjeux du big data à l'échelle de l'entreprise.

Pouvez-vous nous exposer le principal intérêt de la plateforme HDP pour l'entreprise ?

HDP offre une version d'Hadoop adaptée à l'exploitation dans un contexte business. Les apports se matérialisent surtout au travers de la notion de data lake, qui offre la possibilité technique de rassembler toutes les données de l'entreprise, dans des volumes considérables, sans se soucier du format (structuré ou hétérogène) et sans passer par de multiples copies.

En fait, Hadoop évite les problèmes de réplication des données par projet dans l'entreprise. Nous pouvons stocker les données une seule fois au même endroit (dans le « data lake »). Ensuite, toutes les opérations de traitement (BI, prédictive analytics, exploration de données…) viennent se greffer au-dessus pour exploiter ce data lake. L'avantage est que l'on a la même source de vérité pour tout le monde.

La solution Hortonworks vous permet donc de préserver la qualité de vos données ?

C'est effectivement un objectif essentiel. HDP nous permet de réduire l'intervention humaine qui caractérise les approches traditionnelles en silos. Dans ces approches traditionnelles, tout projet qui implique d'importants volumes d'informations amène à faire de nombreuses copies de bases de données ou de sources de données. Or, toute copie altère la donnée et entraîne un problème de qualité.

De plus, la création d'une base de données implique l'élaboration d'un schéma qui tronque la réalité et altère la donnée dès le premier stockage, avant-même que l'information commence à être exploitée.

La qualité des données est donc un sujet majeur dans le contexte du big data. Il faut bien comprendre une chose : lorsque l'on centralise toutes les données sur un data lake, on centralise également les problèmes associés à ces données ! Au final, la mise en place d'un HDP est très révélatrice de l'état de votre système décisionnel. Il faut être prêt à voir la réalité en face. Je pense que la notion de data quality – qui peut paraître abstraite – correspond à des challenges majeurs que les responsables informatiques doivent nécessairement vivre pour en mesurer l'importance.

Comment cette problématique d'altération de la donnée se traduit-elle sur le terrain ?

Prenons l'exemple d'un projet qui va impliquer plusieurs corps de métiers : le marketing, la finance, et les ventes. Pour mener des études et réaliser des calculs, chacune de ces organisations métiers aura besoin des données issues du CRM.

Le Marketing, les ventes, la finance vont chacun faire une copie de la source de données et par la suite faire des extractions de cette source dans un Excel. Les résultats seront à nouveau copiés dans un Excel et réinjectés dans un système de BI. Autant dire qu'entre la source brute et le résultat final, la donnée sera fortement altérée…

> **« Comment envisager des analyses pertinentes si les données qui servent de base à votre raisonnement ne correspondent pas à la réalité ? »**

En pratique, le problème vient du fait que chacun fonctionne selon ses propres règles métiers et élabore des schémas d'utilisation de la donnée de manière isolée. Tout le monde fait parfaitement son travail à partir des solutions existantes : le seul problème est que ces solutions ne parlent pas le même langage ! Les résultats sont donc faussés dès le départ puisque personne ne travaille avec les mêmes informations.

Avec l'architecture HDP, toutes les données sont stockées au même endroit, les règles métiers sont énoncées une seule fois, et implémentées dans un datalake. On obtient donc une vue unique, consistante et cohérente de la donnée.

Pouvez-vous nous donner un exemple d'utilisation de HDP dans le contexte de Neopost ?

HDP nous a permis de passer d'un mode curatif à un mode préventif dans le processus d'amélioration de la satisfaction client, la qualité de nos interventions terrains ainsi que le planning de ces interventions. L'impact sur les opérations a été très fort.

Si je devais résumer les apports d'HDP chez Neopost, je dirais que la solution nous permet :

- De faire converger la donnée et permettre en totale ubiquité de bénéficier d'un outil qui va envoyer une restitution du business et du parc machine à l'identique pour tous, où que l'on soit dans le monde.

- D'offrir un outil d'exploration de la donnée à des fins d'analyse, par exemple pour anticiper le remplacement des machines d'affranchissement confiées à nos clients, ce qui représente une information capitale à la fois pour les commerciaux, les services de maintenance ou les acheteurs.

Quelle visibilité aviez-vous sur votre parc machine avant l'implémentation de HDP ?

Pendant des années, Neopost a fonctionné avec de l'ETL et des BDD structurées. Les collaborateurs en charge de la qualité agrégeaient ces données et les visualisaient. Mais cette visualisation était pauvre et longue… Nous n'avions pas une idée précise et fiable de l'état de notre parc machine.

Avec Hadoop, nos équipes ont gagné en instantanéité et surtout en précision. Tout cela a commencé par une totale remise en question de l'organisation de NeoPost : en matière de décisionnel, tout ce qui avait été réalisé jusqu'à l'implémentation de HDP ne correspondait pas à la réalité. La première solution de BI en place s'appuyait sur des données structurées qui tronquaient la donnée en elle-même. Donc nous n'avions pas une restitution fidèle du business et du parc machine. De plus, l'importante volumétrie de données impactait les temps de traitement. Il fallait accélérer les process et mettre le business en position d'être prêt à décider. C'est ce que nous avons pu réaliser avec les équipes d'Hortonworks.

> **« Avant l'implémentation de HDP, toutes les actions que nous avions pu entreprendre en matière de décisionnel ne correspondaient pas à la réalité »**

Quels types d'analyses avez-vous pu mener sur votre parc de machines grâce à l'architecture HDP ?

L'analyse des données remontées par les machines à affranchir nous permet de détecter des comportements ponctuels anormaux, que nous appelons des signaux faibles.

Chez Neopost, chacune des machines à affranchir dispose d'un composant hardware sécurisé qui implémente un protocole de cryptage afin de sécuriser les transactions financières avec les serveurs postaux.

Par ailleurs, les machines embarquent des horloges afin que toutes les transactions soient chiffrées et horodatées. Il suffit que ces horloges subissent un dérèglement de quelques minutes pour que les machines à affranchir cessent de fonctionner. Lorsque l'on sait que le parc de Neopost comprend des milliers de machines, on peut imaginer l'ampleur du problème en termes business… Que fait le client qui utilise la machine en question ? Est-il en vacances ? Son entreprise est-elle en faillite ? Sa machine est-elle en panne ou a-t-il simplement cessé de l'utiliser ? Pour quelles raisons ? Autant de questions auxquelles nous avons pu répondre en identifiant des signaux annonciateurs de ce dérèglement.

Grâce à l'analyse des données, nous avons pu remarquer que les machines atteignent un point de rupture lorsqu'elles se dérèglent au-delà de la demi-heure. C'est alors qu'elles cessent de fonctionner. Bien évidemment, un tel dérèglement ne se fait pas du jour au lendemain, mais progressivement, de minute en minute. Le retard qui se forme au fur et à mesure transparaît dans les données de fonctionnement qui sont envoyées par les machines. L'architecture Hadoop nous permet de le détecter et de prendre des mesures au plus tôt. Désormais, nous sommes en mesure de déterminer plusieurs mois à l'avance les machines qui vont devoir subir une intervention. D'un mode curatif, nous passons à un mode préventif.

> **« A l'échelle de la supply chain, HDP nous permet d'expliquer certains dysfonctionnements sur notre parc de machines tout en améliorant le produit »**

Sans HDP, ces analyses ne seraient pas envisageables ?

Avec une organisation en silos, cela serait difficilement envisageable, et à un coût prohibitif. Si vous voulez être en mesure d'analyser efficacement les données et d'élaborer des hypothèses à valider ou invalider, vous devez réfléchir à un nouveau modèle pour vos données. Cela peut s'appréhender au travers d'un autre exemple…

Imaginez que nous souhaitions analyser de nouveaux comportements ou de nouveaux signaux en provenance de nos machines. En présence d'un schéma de données structurées, nous serions contraints d'aller demander en bout de chaîne aux équipes IT de modifier le schéma des tables. En fait, dans une architecture traditionnelle reposant sur des bases SQL, vous devez savoir à l'avance les données que vous allez recevoir afin de mettre en place un schéma. Ce n'est pas le cas dans une approche non structurée. Grâce à l'architecture HDP, nous pouvons stocker en brut tout ce que nous recevons. C'est ensuite que nous pouvons explorer ces données, sans aucune altération de leur qualité. Pour résumer, nous passons de l'ETL (Extract-Transform-Load) à l'ELT (Extract-Load-Transform). Auparavant, nous transformions la donnée avant de la charger. Avec HDP, nous la chargeons et la transformons après. C'est un changement majeur.

> **« Avec HDP, nous passons de l'ETL vers l'ELT »**

Comment s'est déroulée l'implémentation de la plateforme HDP ?

L'implémentation puis l'utilisation d'un data lake de type HDP est une invitation forte à changer sa façon de penser. Chez Neopost, nous avons créé un espace de type startup interne, une organisation extrêmement véloce en capacité d'actionner HDP et le cloud computing avec une approche de type LeanStartUp. Il faut bien comprendre que la mise en place d'un HDP conduit à remettre en question toute l'informatique décisionnelle de l'entreprise. Il faut donc s'inscrire dans une logique de conduite du changement.

L'approche de Neopost a été d'agir en mode accompagnement. Nous avons rassemblé des gens de la R&D, de l'IT, du business ainsi qu'un tiers intervenant sur le projet afin de s'attaquer directement à un problème business, en allant droit à l'essentiel. Notre premier chantier a été la mise à disposition de différents datasets dans un lac de données afin de lutter contre le churn, à savoir le phénomène de diminution de la base clients.

L'implémentation de HDP a été réalisée dans un premier temps sur le cloud avec une version d'Hortonworks. Par la suite, nous avons adopté la solution sur un cluster on premise. En termes de planning, nous sommes partis sur des délais courts, 3-4 semaines, une petite équipe et pas de plan préétabli. Pour bien démarrer un projet Hadoop, je pense qu'il faut un gros problème à résoudre et quelques personnes pour travailler dessus…

La structure doit être petite et hyper agile car travailler sur une architecture HDP implique de tester, comprendre, échouer, recommencer en permanence… Rien n'est véritablement prévisible et nous apprenons au fur et à mesure que nous avançons. J'insiste sur une chose : le budget est secondaire, car très faible. En fait, quand on démarre avec HDP, on se retrouve face à un paradoxe : ce qui compte ne se compte pas bien, et ce qui se compte bien ne compte pas du tout.

> **« Au démarrage d'un projet HDP, ce qui compte ne se compte pas bien, et ce qui se compte bien ne compte pas du tout. »**

Pourquoi avoir opté pour Hortonworks plutôt qu'une autre plateforme ?

Hortonworks propose la version d'Hadoop la plus proche de la mouture open source dans sa maintenance et son utilisation. La version est directement exploitable dans un environnement business. Mais ce qui a vraiment fait la différence est l'accompagnement dont nous avons pu bénéficier.

Hortonworks s'est d'emblée attaché à nous accompagner afin de démarrer de manière extrêmement pragmatique et incrémentale. Nous avons commencé par une étape de «business understanding». Ensuite seulement, nous avons entamé les phases d'identification des sources (Data understanding) à utiliser, leur extraction puis leur exploitation.

Quels sont les pièges à éviter dans un projet Hadoop ?

Mettre en place un hadoop, c'est accéder à la donnée et être en capacité de l'analyser, de la comprendre et de l'exploiter. On sait que les informations représentent du pouvoir. Lorsqu'une partie de l'organisation met en place un lac de données, il y a de fortes probabilités pour qu'une autre partie de cette même organisation souhaite en faire autant. Attention à ce cas de figure !

Quand on fabrique un lac de données, on rassemble tous les silos de données de l'entreprise au même endroit. On fait en quelque-sorte un lac de silos. Ce qui est tout à fait normal. En revanche, si plusieurs départements de l'entreprise décident de mettre en place leur Hadoop, un nouveau problème voit le jour… Il faut bien faire attention de ne pas fabriquer des silos de lacs.

> **« Un lac de données c'est un lac de silos. Attention aux silos de lacs qui risquent de naître au fil de l'adoption au sein de votre organisation».**

A propos d'Hugues Le Bars…

Hugues Le Bars est ingénieur en génie logiciel (ETGL) et détenteur d'un DEA en aérothermochimie (Université de Rouen). Il occupe actuellement le poste de Chief Data Officer chez Neopost, leader mondial des solutions d'envoi de courrier et de communication pour les entreprises.

Acteur transversal pluridisciplinaire, (Marketing, opérations, finance, technologie, conduite du changement) il s'attache à promouvoir des attitudes et des organisations « data driven ».

Conférencier International, il a précédemment occupé les postes de développeur logiciel chez Thalès Naval Combat Systems, conseiller chez Onesixty2 et European Technical Leader chez Yahoo ! pendant 11 ans, période au cours de laquelle il a contribué au lancement de Yahoo ! Finance Europe. Dans son travail quotidien, Hugues porte une attention spécifique à la croissance à grande vélocité et aux « capaciteurs » (enablers) des écosystèmes Big Data et Cloud Computing.

HUGUES LE BARS
Chief Data Officer

NEOPOST