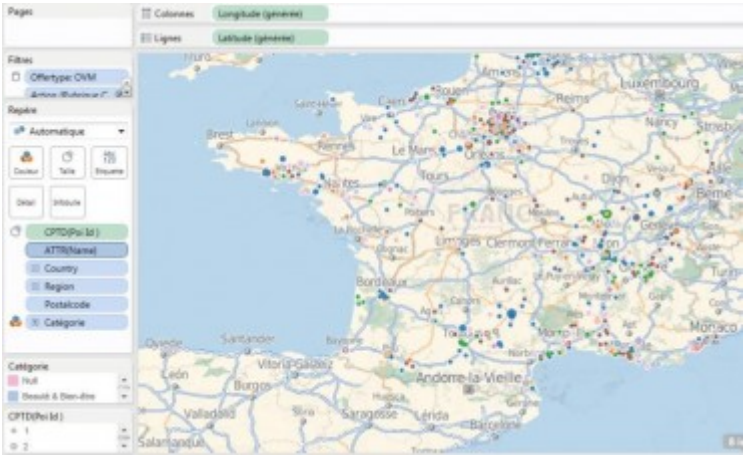
Big Data : Mappy accélère son cluster Hadoop… sans acheter de serveurs

C'est l'histoire d'une application Big Data confrontée à une crise de croissance. Spécialiste des cartes et itinéraires sur le web, Mappy (10 millions de visiteurs par mois, à 45 % sur mobile) s'est lancé dans une stratégie d'analyse de données à grande échelle voici environ 3 ans. « *A l'origine, on travaillait sur l'analyse de la performance de nos serveurs, avec une base SQL Server* », raconte **Cyril Morcrette**, le directeur technique de Mappy. Rapidement, la société bascule sur Hadoop, pour réduire les temps de calcul de batch quotidiens, qui passent alors d'environ 22 heures sur une partie des données… à moins de 2 heures sur l'intégralité du jeu de données.



La plate-forme mise en place génère de nouveaux appétits dans l'entreprise, notamment au sein de l'équipe marketing. « D*e plus en plus, nous avons notamment dû faire face à des questions portant sur les points d'intérêt (monuments, transports, hôtels, restaurants, commerces, etc.) : le nombre d'affichage, d'ouverture de fiches, de clics… Et à des demandes permettant d'assurer le suivi au jour le jour de ces indicateurs* », précise **Nicolas Korchia**, responsable BI du site Internet (à gauche ci-dessus). Après avoir rempli un rôle technique (prévoir l'évolution de l'infrastructure), le cluster Hadoop est donc mobilisé pour mieux connaître le comportement des utilisateurs. « *Car, sur ces points d'intérêt, nous avons des services payants, via les offres de notre maison-mère SoLocal, pour lesquels nous réalisons du reporting* », précise Cyril Morcrette.

Les temps de réponse se dégradent dans Tableau

Le cluster Hadoop ingurgite donc de plus en plus de données. Actuellement, il grossit de **150 Go par jour**. « *A force, nous sommes arrivés à des volumétries importantes* », observe le directeur technique. Un embonpoint qui met à mal l'architecture pensée au départ pour des volumes plus modestes. « *Nous consommons les données dans Spark In-Memory et assurons leur visualisation dans Tableau Software, avec des jointures avec les dimensions hiérarchisées assurées dans Hive*, détaille Nicolas Korchia. *Or, chaque interaction d'un utilisateur dans Tableau génère une dizaine de requêtes SQL. Avec des tables de 10 ou 100 millions de lignes, les temps de réponse étaient corrects. Mais ils se sont dégradés très vite quand nous avons dépassé les 500 000 millions de lignes*. »

L'équipe technique est alors confrontée à des temps de réponse d'environ 20 secondes en moyenne. Un manque de réactivité qui pénalise les utilisateurs de la solution de dataviz « *On a découvert les limites de Tableau sur les grosses volumétries* », résume Cyril Morcrette. Une des solutions aurait pu consister à muscler le cluster Hadoop composé de 6 serveurs (comportant chacun 24 cœurs de processeurs). « *Cela nous aurait amené à aller jusqu'à 20, voire 50 nœuds en fonction de la*

volumétrie anticipée », dit Nicolas Korchia. La solution viendra finalement d'une alternative, une **couche d'indexation** développée par un des freelances employés par Mappy, **Florent Voignier** (à droite sur la photo en haut de page).

Un moteur d'indexation développé par un freelance

« *A partir du cas métier de Mappy, il a monté une solution sur ses heures perdues et nous l'a présentée* », précise Cyril Morcrette. Sans modifier l'architecture, l'ajout de ce moteur d'indexation, baptisé Indexima, a **ramené les temps de réponse à environ 0,1 seconde**. 32 Go de mémoire ont été dédiés au moteur, fonctionnant comme une application Yarn (Yet Another Ressource Negociator, assurant la gestion des ressources de Hadoop depuis la version 2.0). « *Avec 17 milliards d'affichage de points d'intérêt, sur 2,7 milliards de lignes, les requêtes SQL sont exécutées en moins de 100 ms. Les utilisateurs ne s'en rendent plus compte. Et, à ce jour, on ne voit pas de limite à l'outil* », dit Nicolas Korchia. Indexima est **en production chez Mappy depuis 6 mois**.

Lancé officiellement lors du salon Big Data Paris (qui se tenait les 7 et 8 mars au Palais des Congrès), la solution Indexima, portée par l'entreprise du même nom, combine différentes techniques. « *Elle fonctionne entre 100 et 1 000 fois plus vite, principalement grâce au cumul de trois techniques : des index multi-dimensionnels In-Memory, de la pré-agrégation et du stockage sur disque orienté colonnes* », résume Florent Voignier, qui s'exprimait lors d'une conférence sur le salon parisien. Le développeur envisage désormais de passer sa technologie en Open Source, en y greffant une licence payante pour son moteur de création d'index. Une nouvelle accueillie très positivement par Mappy, « *d'autant qu'il reste des éléments à stabiliser dans Indexima* », note Cyril Morcrette.

« Le mur d'après »

Ce nouveau greffon ajouté au SI du spécialiste du Web (qui emploie quelque 80 personnes, auxquelles s'ajoute une vingtaine de prestataires) illustre le besoin des entreprises de faire perpétuellement évoluer leurs architectures Big Data. « *A chaque fois qu'on finit une évolution, on se heurte au mur d'après* », résume Cyril Morcrette. En deux ans, l'architecture Hadoop du spécialiste de la cartographie a connu **4 virages majeurs** : le passage de Python à MapReduce pour la préparation de données, le choix de Tableau Software pour la dataviz, le remplacement des cubes par Spark In-Memory et, enfin, l'arrivée d'Indexima. « *Comme toutes les entreprises, on se cherche encore beaucoup sur le sujet de la donnée* », dit Cyril Morcrette.

L'équipe chargée de la donnée chez Mappy (4 développeurs, 2 analystes BI, 2 personnes du marketing auxquels s'ajoute une équipe de préparation des données) est aujourd'hui confrontée à des contraintes réseau, liées aux choix d'infrastructure passés. « *Le système HDFS fonctionne sur des baies Isilon. Le cluster échange tellement avec HDFS qu'on frôle la saturation du réseau* », résume Nicolas Korchia. Mais le plus gros défi réside dans les **nouveaux champs d'exploitation de la donnée**. « *Nous allons travailler sur la géolocalisation des utilisateurs, afin de personnaliser l'application grâce au Machine Learning* », dévoile Cyril Morcrette. Objectif par exemple : reconnaître automatiquement quand un utilisateur s'apprête à faire son trajet travail-domicile. « *Cela nécessite de nouvelles évolutions de compétences, au niveau technique bien sûr, mais aussi au plan marketing* », ajoute-t-il.

A lire aussi :

LinkedIn place en Open Source son outil Big Data, WhereHows

La Poste Courrier préposée à transformer le Big Data en or

Données inutiles ou non classées : quand le Big Data coûte bonbon