Spécial Big Data : 3 – A la recherche des 'Data scientists'

C'est certainement la face cachée du Big Data, voire son talon d'Achille. Il existe bien un enjeu majeur au-delà du stockage de données hétérogènes dans des environnements dispersés (cf. précédents articles 1 et 2) : comment accéder aux bonnes compétences pour exploiter statistiquement la donnée, lui apporter plus de valeur et en faire un outil de prise de décision ?

Cette difficulté provient du positionnement original du Big Data analytique, qui se définit entre technologies de stockage et analyses statistiques. Deux approches qui ne sont naturellement pas compatibles, car elles proviennent de deux mondes faisant appel à des compétences différentes. Le professionnel IT, plus spécifiquement le spécialiste du stockage, n'est pas un statisticien – et inversement. Même s'ils peuvent afficher, parfois, un parcours commun à travers les mathématiques. D'ailleurs, chacun d'entre eux dépend d'un métier différent dans l'entreprise.

La communauté du SQL sur les bases structurées

On pourra objecter que l'analytique n'est pas chose nouvelle dans le monde des IT. Et que les outils de **BI** (***Business Intelligence***) sont très répandus. Il faut cependant différencier ces deux approches, comme déjà évoqué.

La BI s'appuie principalement sur l'analyse de la donnée structurée. Cette donnée s'exprime sous la forme de tableaux, en lignes et colonnes, où la ligne correspond, généralement, à une fiche, et la colonne à un champ (par exemple une fiche individu est composée des champs 'nom', 'prénom', etc.). D'où la notion de structure. L'outil d'interrogation de ces bases de données, généralement associées aux métiers de gestion de l'entreprise (fichiers comptables, salariés, clients, etc.), utilise un langage commun de requêtage, le **SQL** (*Structured Query Language*).

Les outils de BI reposent donc d'une part sur les bases de données structurées de l'entreprise, qualifiées de SQL ; et d'autre part sur un langage d'interrogation de ces bases, qui leur a donné son nom pour les qualifier puisqu'il s'agit du langage SQL. Cette base est quasiment la même pour tous les acteurs du marché, qu'ils se concentrent sur la base de données ou sur son interrogation.

Le SQL est ainsi commun à toutes les plateformes de gestion, donc largement pratiqué. La réalisation d'analyses sur ces environnements se révèle être une démarche relativement maîtrisée qui s'exprime à travers des compétences répandues.

L'émergence du NoSQL sur les bases non structurées

Le Big Data est une réponse à la multiplication des données qu'accumule l'entreprise ou auxquelles elle peut se donner accès. Certaines de ces données sont qualifiées de « non structurées », c'est-à-dire qu'elles n'entrent pas dans le moule SQL et ne peuvent être organisées en colonnes et en champs.

Citons les mails, les fichiers bureautiques, les images, les vidéos, les tweets, etc. Le modèle SQL d'interrogation des bases de données ne s'applique pas ici… même si de plus en plus d'éditeurs tentent de proposer des outils de requêtage NoSQL qui se rapprochent de ce modèle. Concrètement, cela signifie que la compétence traditionnelle des spécialistes des bases de données en matière d'analyse et de BI ne s'applique pas sur la très grande majorité des données évoluant dans le Big Data !

Il y a **deux problématiques** liées à la création d'analytiques sur les données en Big Data:

- La première provient de l'environnement des données. Les compétences SQL ne permettent pas d'exploiter des données non structurées. A commencer par la gestion de ces données à travers l'environnement Hadoop et le système de fichiers HDFS.
- La seconde est liée à l'absence de support du langage SQL par les bases de données non structurées. L'interrogation des bases Big Data pour réaliser de l'analytique doit donc être créée via des langages supportés par la plateforme Hadoop (en particulier Java) et les bases NoSQL.

De la difficulté à cumuler les compétences

Comment ces problématiques se traduisent-elles sur le terrain ? Aujourd'hui, la réalisation d'analytiques sur le Big Data nécessite des compétences que ne possèdent pas tous les développeurs d'outils de BI ou les utilisateurs de tableurs de type Excel ! Ces compétences doivent porter sur le pilotage des données Big Data en environnements non structurés et l'application de statistiques. Nous sommes donc là en présence de deux métiers qui s'ignorent : l'informaticien et le statisticien.

Pourtant un métier pourrait émerger, dont la promesse est de réunir ces deux compétences : le ***Data scientist***. La fonction a un nom, il faut maintenant trouver la perle rare qui saura réunir ces compétences. Comme son nom l'indique, ce nouveau métier réunit les mondes de la donnée, donc du stockage en Big Data et du traitement de la donnée, et celui des sciences – mathématiques et statistiques…

Ne le cherchez pas dans les cursus de formation universitaires et de formation des ingénieurs, ces deux mondes ne cohabitent pas… ou tout du moins pas encore. Certaines universités américaines proposent déjà des formations de Data Scientist, les écoles européennes et françaises suivront. Attendons encore 2 à 3 ans avant que sortent des moules les premiers ingénieurs *Data scientists* diplômés.

En attendant, le focus des entreprises qui lancent des solutions de Big Data analytique porte principalement sur le déploiement des infrastructures de stockage et de compilation des données. Autour des technologies Hadoop.

Quelques intégrateurs et sociétés de services ont commencé à compenser l'absence de *Data scientist* par la réunion d'équipes pluridisciplinaires réunissant des spécialistes des IT et des statisticiens. Un mariage délicat pour le moment, surtout concentré sur la création des premiers PoC (*Proof of concept*). Les entreprises les plus avancées dans le Big Data analytique sont rares, mais existent, à l'image d'IBM. Le recrutement de scientifiques et statisticiens par une société dont la

culture est à la fois orientée IT et R&D leur offre un avantage unique, celui de disposer de *Data scientists*.

Si la perle rare du Big Data analytique, le Data Scientist, existe, encore faut-il la trouver… ou faire appel au bon partenaire.

_____

Ref.: **DOSSIER** :

Spécial Big Data : 1 – Que recouvrent au juste le Big data et la gestion de données?

Spécial Mobilité : 2 – Tous ces enjeux qui pèsent sur la DSI

Spécial Cloud Computing : 1 – L'intérêt des architectures ouvertes

Spécial Réseaux sociaux professionnels: 1- L'effet consumérisation

___