Guide Pratique :

Les éléments clés d'un projet de Big Data réussi

La mise en place de projets de type Big Data est sur la feuille de route stratégique de nombreuses entreprises. Mais pourquoi et comment mener un projet Big Data ? Didier Kirszenberg de HPE fait le point avec nous sur cette thématique.

Le Big Data est l'évolution du décisionnel traditionnel. Son sujet est donc la valorisation de la donnée et non débat de simple stockage de gros volumes.

Le Big Data se différentie d'un système traditionnel selon 3 axes :

Là où le décisionnel classique définit qui fait quoi le Big Data mets en avant la analyse mathématique de la donnée avec la possibilité de faire des regroupement plus fin, d'identifier des comportements atypiques voir de définir des modèles prédictifs.

Le décisionnel historique ne savait traiter que de la donnée structurée la ou les technologies issue du Big Data savent donner du sens à n'importe quelle source de donnée. Cela peut couvrir des log comme ceux des site web pour suivre une navigation, jusqu'à de la donnée de vidéosurveillance.

Enfin là où le décisionnel historique collectait la donnée du service ou au mieux de l'entreprise, le Big Data recommande une approche ouverte pouvant inclure les informations publiées sur de site web externe ou la mise en commun de la donnée de l'entreprise avec celle d'autres sociétés aux métiers complémentaires.

En fait on peut résumer pour un dirigeant le sujet du Big Data à la question suivante : « Quelles sont les sources de données disponibles à l'intérieur ou à l'extérieure de l'entreprise et puis je leur donner de la valeur pour augmenter mon efficacité ».

Désigne des traitements de données massives, dont le volume est tellement important que cela impose l'utilisation d'outils de nouvelle génération. Voire l'application de nouvelles approches.

Didier Kirszenberg, responsable des architectures Massive Data chez HPE France, nous propose de faire le point sur les éléments clés qui vous permettront de mettre en place un projet de Big Data couronné de succès.

Ne pas se tromper de projet

Il faut prendre le sujet par le bon bout. *« Bien plus que sur le Cloud, un projet Big Data est avant tout un sujet métier et fonctionnel. Il est donc essentiel de savoir ce que l'entreprise veut valoriser au travers de ce projet. Il convient de déterminer ce que l'on veut faire ressortir des données. »*

Bref, avant d'aller vers un projet Big Data, il faut se poser les bonnes questions et bien comprendre les spécificités du sujet. L'entreprise passe en effet ici d'une logique de décisionnel / reporting propre aux data warehouses vers du traitement data centric.

Fonctionner en mode itératif

Un data lake ne doit pas être un fourre-tout, et les informations doivent être choisies avec soin. Chaque use case viens avec potentiellement une nouvelle source de donnée et peuvent à ce titre ouvrir de nouvelle opportunités de croisement d'information. *« Le data lake, c'est un cheminement »*, philosophe notre interlocuteur.

L'autre raison au mode itératif est que dans un grand nombre de cas on ne peut pas donner de ROI à priori sur un projet Big Data. Ceci est lié au fait que l'on fait souvent de l'exploratoire pour trouver des modèles mathématique sur la donnée et qu'il est possible de ne pas en trouver (ce qui peut aussi être lié à la qualité et la précision de la collecte). Dans ce contexte, mieux vaut opter

pour un mode itératif. *« Il est préférable de monter de petits projets, montrer qu'ils permettent de dégager de la valeur, puis passer à l'étape suivante, les projets dégageant du profit vont au final couvrir les petits investissement de ceux qui n'ont pas été poursuivi. »* Pas besoin de penser immédiatement à la mise en place d'une infrastructure Big Data massive, même si certains projets risquent parfois de collecter beaucoup de données et d'imposer des contraintes fortes en matière de matériel.

« Un projet Big Data, c'est de l'exploratoire : plus les métiers vont développer le goût pour la valorisation de la donnée, plus l'infrastructure va grossir pour répondre à ces besoins. »

S'assurer de la qualité de la donnée

« L'enjeu principal dans ce type de logique, c'est la qualité de la donnée. Surtout quand on se dirige vers un data lake (du stockage de données Big Data, NDLR), explique notre expert. *Le Big Data demande énormément de travail pour qualifier la donnée. Ce n'est pas seulement un sujet technique, mais aussi métier. »*

La qualité de donnée dans le Big Data est différente de la qualité de donnée du décisionnel historique. Historiquement la qualité de donnée est surtout sur la synchronisation des sources et on écarte ce qui est atypique. Dans le Big Data il faut que la donnée non structuré capturé ai été validé d'un point de vue métier sur sa validité. Par ailleurs comme le Big Data se propose d'analyser les signaux faible les donnée anormale ne doivent plus être excluent mais demande au contraire un approfondissement pour savoir si leur caractère anormal est liè à du comportement ou à un problème dans la collecte.

L'échange se doit d'être permanent entre la technique et les métiers. *« Le data scientist va se charger de créer ce lien entre métiers et technique »*, explique Didier Kirszenberg.

Pour favoriser ces échanges, méthodes agiles et la logique organisationnelle des initiatives de type DevOps sont des atouts clés.

En résumé, il convient de bien définir les attentes côté métier et l'exploratoire côté technique, le niveau de qualité de la donnée étant clef dans la faisabilité du projet.

Choisir les bons outils

Hadoop est souvent cité comme base pour une solution de Big Data. Oui, mais quoi exactement dans Hadoop ? *« Hadoop est écosystème de modules : il faut déterminer lesquels utiliser. La DSI doit sélectionner les bons outils et comprendre la nature et les spécificités de chacun. »* Une infrastructure de Big Data est tout sauf une solution mono produit et mono usage. Il ne faut donc pas négliger des offres tierces, comme les bases en colonne très adaptées à certains traitements analytiques. Certains de ces outils conservant une interface SQL les DSI seront en mesure de réexploiter un savoir-faire existant.

Quid du Big Data en mode Cloud ? *« Il y a un évident risque de perte de contrôle des données. La plupart des fournisseurs ont des clés de facturation multiples, largement sous-estimées par les utilisateurs. L'analytique suppose beaucoup d'accès aux données. Un simple changement dans ces accès peut faire exploser la facture. »* Méfiance donc, même si ce mode de déploiement n'est pas à écarter systématiquement.

Une infra Big Data n'est pas une infra Cloud !

Il est important de comprendre qu'une infrastructure de Big Data n'est pas identique à une infrastructure de Cloud Computing. Hormis leur taille et l'utilisation de technologies non propriétaires, les deux ne partagent rien. Elles sont mêmes à bien des égards opposées.

Un Cloud se charge d'opérer une multitude de services sur un nombre le plus limité possible de systèmes, à raison de plusieurs machines virtuelles par machine physique. *« Une infrastructure Big Data c'est l'inverse, avec un même service fonctionnant simultanément sur plusieurs machines physiques »*, explique Didier Kirszenberg. Nous nous trouvons ici face à une architecture massivement parallèle, où un traitement pourra s'accaparer tout ou partie du cluster. Nous allons donc employer des techniques issues du monde du calcul de haute performance (HPC pour High-Performance Computing).

« Chez HPE, nous avons adapté des produits et des méthodes issus de 20 années d'expérience HPC au monde du Big Data. » Si les acteurs du Cloud sont habitués à travailler sur de telles volumétries de serveurs, mieux vaut faire appel à un spécialiste du HPC, car le mode opératoire est ici différent. Par exemple, les outils d'orchestration se doivent de réallouer des pôles de serveurs, et non des morceaux de serveurs. Un outillage adapté doit donc être utilisé.

Des contraintes techniques fortes

« *Avec une architecture data centric, le risque d'engorgement du réseau est réel »*, alerte Didier Kirszenberg. L'approche classique Nord-Sud, qui trace le chemin le plus court entre le serveur et les utilisateurs, se doit ici de laisser place à une stratégie Est-Ouest, afin de rapprocher les serveurs entre eux, leur permettant ainsi de faire front face à l'afflux de données.

Le stockage de la donnée doit également se faire au plus près de la machine, les déploiment actuels en Big Data utilisent des serveurs assurant à la fois calcul et stockage. Avec l'ajout de plus en plus de module spécifique les clients ayant de gros volumes et une grand diversité d'usage peuvent voir une limite à cette approche car chaque workload peut demander une ressource spécifique. *« Sinon il faudrait opter pour des serveurs blindés en tout (processeur, mémoire, stockage, gpu, etc.) »*, ironise notre expert.

HPE préconise une approche qui permet dans un deuxième temps de besculer vers une architecture asymétrique, qui (re)discocie compute et storage, mais en conservant ces deux éléments sur des serveurs X86 (il faut absolument rester sur le standard de l'open source car il évolue vite) dans le même rack qui devient alors une sorte de super serveur modulaire, capable de croître en fonction des besoins, par ajout de nouvelles unités de stockage, de calcul classique, voire de calcul massivement parallèle, via des GPU.

Du métier à l'infra… jamais l'inverse

Nous l'avons vu, un projet Big Data est avant tout un projet métier. Si les utilisateurs n'ont aucune idée de ce qu'ils pourront tirer de leurs données, mieux vaut ne pas démarrer la mise en place d'un tel projet !

Une fois les besoins définis, la technique devra sélectionner les bons outils. Il conviendra également de qualifier les données avec soin. L'infrastructure, hors fonds de roulement lié au stockage de la donnée elle-même, sera dimensionnée en fonction des tâches opérées sur les données.

Avec une infrastructure de type data centric, c'est en effet le volume et la complexité des traitements appliqués aux données par les utilisateurs qui va définir la taille du cluster à mettre en place par la DSI.