> qui alimente les applications de décision ou d'apprentissage. Qu'il s'agisse d'algorithmes simples ou complexes, déterministes ou prédictifs, opérant sur des données exactes ou bruitées, leur compréhension est indispensable pour entrevoir en quoi consiste l'exploitation des *big data*.

Occultés par les succès récents de l'apprentissage machine et autres programmes de jeux, ces algorithmes enfouis sont donnés pour acquis, avec un coût nul pour leur mise en œuvre et leur exécution. Nous verrons que ce n'est pas le cas, un travail permanent de recherche et d'ingénierie est nécessaire pour les adapter aux nouveaux contextes applicatifs, les étendre, les optimiser et en inventer de nouveaux.

Il y a autant de processus de gestion et de valorisation des données que de raisons de valoriser ces données. On y retrouve cependant de grandes activités génériques (*voir l'encadré page ci-contre*), sur chacune desquelles existent des verrous importants.

Pour réaliser ces grandes activités, il est important de spécifier les tâches constitutives en fonction des applications visées. La consultation intensive des données n'a pas les mêmes besoins qu'un traitement des flux de données en temps réel. Une application d'aide à la décision n'a pas les mêmes besoins qu'une application de prédiction du panier d'un consommateur ou d'aide au diagnostic d'une tumeur. La nature de l'application et des tâches algorithmiques qui en découlent doivent être mises en perspective avec le modèle de calcul le plus adapté.

DEUX MODÈLES DE CALCUL

La gestion des données n'est pas constituée d'algorithmes *ad hoc* réalisés au coup par coup, dans un langage donné, en fonction des applications. C'est une science qui a ses fondements théoriques, ses propres objets d'étude, ses propres abstractions et ses propres méthodes. Les technologies produites ont pour vocation de réduire l'effort nécessaire à la gestion des données, de contrôler l'accès à ces données et de faciliter l'évolution des systèmes sans réécriture de leur code. Pour simplifier, il y a deux modèles de calcul sur les grands volumes de données: celui dédié à la recherche d'information et celui adapté à l'analyse de données. Même si les deux approches partagent de nombreux concepts et techniques et peuvent se combiner, elles diffèrent d'un point de vue sémantique.

Dans le premier cas, le sens est porté par la requête de l'utilisateur; si la requête est pertinente, les données délivrées seront bien conformes aux critères de recherche spécifiés. Par exemple, «rechercher les références de produits vendus en quantité supérieure à 200 dans chacun des magasins de la région parisienne durant la période de Noël», est une requête complexe, mais qui définit clairement le calcul à faire et les critères de sélection de données. Ainsi spécifié, le système de gestion de données se chargera de générer le code nécessaire à l'exécution de cette requête. D'autres requêtes sont plus approximatives, leurs critères étant spécifiés de façon lâche. C'est le cas des recherches sur le Web. Ainsi, «rechercher les magasins parisiens pratiquant les prix les moins chers pour le chocolat et le champagne durant la période de Noël» est une requête floue. Elle impose un traitement approché du critère «moins cher» obligeant à présenter les résultats dans un certain ordre.

Dans le second cas, la question du sens est plus complexe, car elle interroge à la fois la méthode utilisée pour l'analyse des données et l'interprétation par l'utilisateur des résultats produits. Pour une requête d'analyse de l'évolution des cours de bourse des entreprises du CAC40 ou de prédiction des ventes de smartphones par région et par période de l'année, les résultats ne sont pas interprétables sans connaissance du métier et des modèles de prédiction sous-jacents.

Ce problème peut devenir plus aigu dans des algorithmes encore moins transparents. Par exemple, dans un système de reconnaissance fondé sur l'apprentissage, la confusion entre un chat et un caniche fait sourire, mais celle d'un honnête citoyen avec un criminel peut être dramatique. Les algorithmes sous-jacents à ces systèmes posent des défis à la société.

Il y a autant de processus de valorisation des données que de raisons de valoriser ces données

D'un point de vue conceptuel, un modèle de calcul est défini par les structures de données qu'il reconnaît et les opérations qu'il offre sur ces structures de données. Les structures de données peuvent être des tables (ou matrices), des arbres, des graphes ou des séquences finies ou infinies. Les opérations sont associées à chaque structure: les opérations de tri, projection, produit et union sur