grand public et pour l'instant, ce qui est le plus apparent du *Big Data*, c'est l'utilisation de données personnelles par de grandes entreprises commerciales, principalement pour des publicités ciblées. C'est le cas de *Google*, qui analyse les requêtes et les courriels des internautes pour mieux cibler sa publicité, ou d'*Amazon*, qui suggère aux usagers des livres à acheter.

Des armes nouvelles aux mains des dictatures... et des citoyens

Récemment, en particulier dans le cadre de l'affaire Edward Snowden aux États-Unis, la presse a souligné que divers gouvernements utilisent l'analyse de données privées à des niveaux surprenants. La principale raison invoquée pour ce type d'utilisations est la lutte contre le terrorisme. Mais le système PRISM de la NSA (*National Security Agency*, l'agence américaine de sécurité) et les systèmes équivalents des autres États sont aussi utilisés pour l'espionnage, notamment industriel. Ils peuvent l'être pour surveiller les opposants politiques. Le contrôle et l'analyse des données sont certainement les nouvelles armes parmi les plus inquiétantes des dictatures et des gouvernements totalitaires.

A contrario, la généralisation dans les pays démocratiques de l'*open data* (l'accès libre aux données du secteur public) devrait permettre aux « journalistes de données », et plus généralement aux citoyens concernés, de contrôler les actions de leurs gouvernants ainsi que celles des grandes entreprises. Les mêmes technologies rendent possibles la surveillance des personnes par l'État et celle de l'État par les citoyens. Les gouvernements de pays tels que les États-Unis, le Royaume-Uni, ou plus récemment la France se sont engagés dans le mouvement de l'ouverture des données. Cela ouvre la voie à de nouveaux services et permet aussi de mieux contrôler les actions des gouvernements et des grandes entreprises. On peut espérer que cela conduise à donner plus de responsabilité au citoyen et à refonder la démocratie.

Les technologies du *Big Data* semblent particulièrement adaptées pour prévoir des crises sanitaires, des problèmes environnementaux, des catastrophes naturelles, et pour y réagir. Elles devraient aider à résoudre les problèmes de santé, de transport, d'écologie, à lutter contre la pauvreté. Dans nombre de ces situations, il faut combiner des analyses de grandes masses de données stockées et des analyses en ligne sur un flux de données obtenues en temps réel. De telles combinaisons se retrouvent ainsi, par exemple, dans le suivi personnalisé de personnes en grande difficulté, de personnes très âgées, d'élèves en échec scolaire, etc. Il faut être capable d'analyser des évolutions sur de longues périodes, de prévenir si possible les problèmes, mais aussi savoir détecter l'urgence et réagir en temps réel à des crises.

Prenons l'exemple de la santé. Les données en rapport avec la santé d'un individu croissent sans cesse. Le cœur en est constitué par des informations telles que les examens médicaux, les diagnostics, les soins et les prises de médicaments. Mais on voit aussi se généraliser les données génomiques – la société californienne de biotechnologie *23andMe* propose ainsi aux individus le séquençage d'une partie importante de leur génome pour 99 dollars (moins de 80 euros). Il faut aussi prendre en compte les données sur la vie quotidienne de l'individu, son alimentation, son activité physique, son exposition à des pollutions particulières, etc. De plus en plus, de telles informations deviennent disponibles *via* des équipements de type téléphone intelligent ou *via* les réseaux sociaux.

Parallèlement, les chercheurs dans le domaine médical font un emploi croissant de l'analyse de données. Par exemple, ils espèrent découvrir des corrélations entre la prise de certaines combinaisons de médicaments et des pathologies particulières. Les patients, surtout, pourraient en profiter. Il s'agit d'abord de personnaliser les soins en adaptant les médicaments à chacun, en contrôlant les quantités administrées. On peut imaginer agir de manière préventive en proposant à chacun une hygiène de vie adaptée à ses risques, en accompagnant chaque personne dans sa nutrition, ses activités physiques, sa vie quotidienne. Mais on peut aussi espérer détecter les crises et aider à les gérer, par exemple en régulant la prise de médicaments.

Un problème : l'accès aux données personnelles

Le fait de disposer de toutes les données relatives à une personne ouvre ainsi considérablement le champ des possibles. Mais cela soulève le problème de l'accès aux données personnelles. Un assureur ou un employeur doit-il avoir accès à tout ou partie de l'information constituée par les données médicales et génomiques d'un client, ses données fiscales, ses achats, sa géolocalisation, ses courriels, ses échanges dans les réseaux sociaux ?

Ce sont les données personnelles du client. À ce titre, elles lui appartiennent, et il devrait être seul à pouvoir décider qui y a accès et comment elles sont utilisées. Mais ce n'est pas aussi simple. Par exemple, pour faire progresser la médecine, on devrait pouvoir faire des analyses sur les données médicales de tous. Il semble raisonnable que les résultats de ces statistiques soient publics. Mais il semble tout aussi clair que les données brutes, par exemple celles d'un hôpital, ne puissent pas elles-mêmes faire partie des données ouvertes, même anonymisées, puisqu'il serait impossible de garantir leur confidentialité. Il reste beaucoup à faire pour concilier ces différents aspects. ■

Hadoop pour gérer les données massives

Hadoop, logiciel libre de la fondation *Apache*, a été conçu en 2004 par l'Américain Doug Cutting. Ce logiciel est adapté à l'analyse de grandes masses de données. Il est fondé sur la technique *MapReduce*, que *Google* a utilisée pour son moteur de recherche.

Dans un calcul *MapReduce*, on commence par découper le problème en de nombreux sous-problèmes indépendants (étape *Map*) que l'on confie à des ordinateurs distincts. Ces machines résolvent les sous-problèmes et envoient leurs résultats à d'autres machines qui ont pour tâche de combiner les résultats (étape *Reduce*). L'objectif est de pouvoir travailler sur d'énormes volumes de données en parallélisant les calculs sur des grappes d'ordinateurs. *MapReduce* ne permet de traiter que des problèmes qui peuvent se décomposer en de multiples tâches parallèles.

Le logiciel de *MapReduce* le plus populaire est *Hadoop*, à la base des centres de données de géants du Web tels que *Amazon* et *Facebook*, et on le retrouve de plus en plus dans des offres de « Cloud computing ». Cependant, malgré ses atouts, *Hadoop* est un logiciel encore jeune et imparfait : il n'utilise pas un langage standardisé, ce qui rend sa gestion complexe, et ne réalise qu'un traitement différé, et non en temps réel, des données.

Les pétaoctets du LHC

L'accélérateur de particules LHC, au CERN, produit chaque seconde des milliards de milliards de collisions entre protons. Chacune engendre de l'ordre de un mégaoctet de données. Comment cette avalanche est-elle traitée ?

Faïrouz Malek

Installé au CERN (le Laboratoire européen de physique des particules), près de Genève, le LHC est l'accélérateur de particules le plus puissant au monde. À partir de novembre 2009, il a livré des données issues de collisions frontales entre protons à une énergie de 7 téraélectronvolts (7×10^{12} électronvolts, ou 7 TeV), puis de 8 TeV, jusqu'à son arrêt en mars 2013. Cet instrument géant et international redémarrera dans deux ans, le temps de le préparer à des collisions encore plus violentes, à environ 13 TeV.

Le LHC (*Large Hadron Collider*, ou Grand collisionneur de hadrons) s'accompagne de quatre gigantesques complexes de détection, nommés ALICE, ATLAS, CMS et LHCb, chacun ayant ses spécificités, qui enregistrent les traces des collisions. Lorsque le collisionneur est en fonctionnement, leurs détecteurs transmettent d'énormes volumes et flux de données mises sous forme numérique. Comme nous allons le voir, le stockage de ces données, leur traitement et leur analyse nécessitent une impressionnante infrastructure informatique, l'une des plus puissantes au monde.

L'enjeu du LHC et des équipements du même type est de tenter de percer quelques-uns des mystères de la matière et de l'origine de l'Univers. La traque du boson de Higgs, une particule attendue par les théoriciens depuis une quarantaine d'années et dont la découverte au LHC a été annoncée en juillet 2012, est l'objectif principal des deux expériences ATLAS et CMS. Voyons d'abord pourquoi cette traque s'apparente à la recherche d'une aiguille dans une botte de foin.

Dans le tunnel du LHC circulent en sens opposés deux faisceaux de protons accélérés à une vitesse proche de celle de la lumière. Chaque seconde, près de 40 millions de paquets regroupant chacun quelque 10^{11} protons entrent en collision. L'énergie de la collision proton-proton est si grande qu'elle peut se matérialiser en toutes sortes de particules, dont des bosons de Higgs.

La théorie indique qu'*a priori*, un seul boson de Higgs est produit en moyenne toutes les 10^{10} collisions. Par ailleurs, le boson de Higgs est instable et se désintègre très vite en deux bosons Z, lesquels se désintègrent à leur tour en leptons détectables (électrons ou muons), ou en deux photons, ou en un boson W^+ et un W^-, etc. Si les détecteurs étaient parfaits, ATLAS ou CMS enregistreraient un boson de Higgs toutes les 10^{13} collisions *via* sa désintégration en deux photons ou en deux Z. En réalité, le taux de détection est très faible. Ainsi, sur les deux

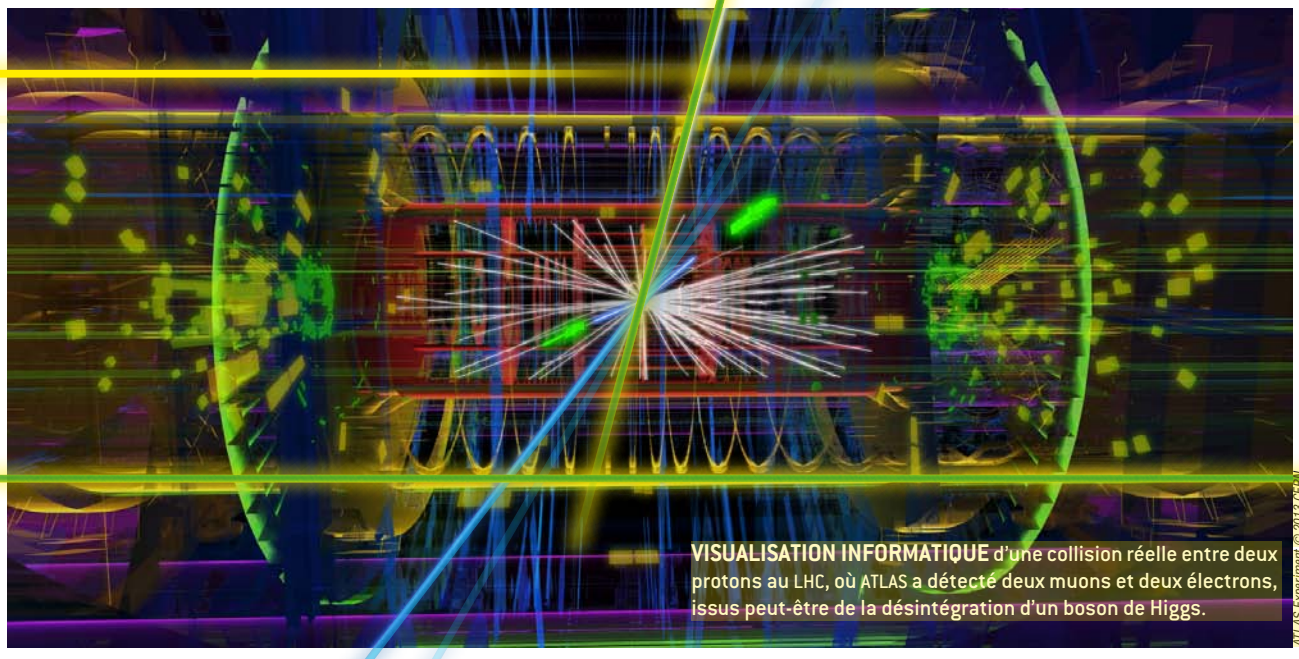
L'AUTEUR

Faïrouz MALEK est physicienne au Laboratoire de physique subatomique et de cosmologie (LPSC) du CNRS et de l'Université Joseph Fourier, à Grenoble. Elle participe à l'expérience ATLAS du LHC et est responsable scientifique du projet LCG-France, la partie française de la grille de calcul du LHC.

BIBLIOGRAPHIE

Dossier « La saga du boson de Higgs », *Pour la Science*, n° 419, pp. 22-35, septembre 2012.

F. Malek, **Le calcul scientifique des expériences LHC - Une grille de production mondiale**, *Reflets de la Physique*, n° 20, pp. 11-15, juillet-août 2010.



VISUALISATION INFORMATIQUE d'une collision réelle entre deux protons au LHC, où ATLAS a détecté deux muons et deux électrons, issus peut-être de la désintégration d'un boson de Higgs.

ATLAS Experiment © 2013 CERN

ans de prise de données (2011 et 2012) où des millions de particules ont été produites (dont environ 500 000 bosons de Higgs, en théorie), ATLAS n'a détecté que quelques centaines de bosons de Higgs se désintégrant en deux photons (contre quelques milliers si le détecteur avait été parfait).

Les flux de données que transmettent les détecteurs sont considérables. Les collisions engendrent en effet des milliers de particules secondaires, le passage d'une particule dans un détecteur se traduisant par des impulsions électroniques à partir desquelles les physiciens peuvent déterminer sa trajectoire et son identité. Les millions d'impulsions électroniques créées par chaque collision sont alors numérisées par des circuits intégrés aux détecteurs, avant d'être transmises par fibres optiques aux ordinateurs d'acquisition de données. Elles y sont regroupées sous formes de fichiers binaires de « données brutes ».

Chaque collision constitue, au niveau de chacune des quatre expériences, un événement dont la taille en données brutes est de l'ordre du mégaoctet. Un tri intelligent est effectué en moins d'une seconde pour ne garder que les événements potentiellement intéressants. Réalisé grâce aux circuits électroniques en amont puis à des grappes d'ordinateurs et des algorithmes spécialisés, ce tri réduit le taux d'acquisition à environ 200 événements intéressants par seconde. Ces données brutes sont enfin enregistrées à un taux équivalent à deux CD par seconde pour chacune des quatre expériences, sachant que la capacité d'un CD est de 700 mégaoctets.

Un gigaoctet par seconde

Les données brutes sont traitées afin d'extraire les informations « physiques » (les caractéristiques de chaque trajectoire, l'impulsion et l'identité de chaque particule détectée), puis de les analyser. Ces traitements réduisent la taille de l'information d'un facteur deux à dix. Ils sont en général répétés plusieurs fois par an sur l'ensemble des données acquises depuis le démarrage, pour tenir compte du développement continu des algorithmes de traitement et d'étalonnage du détecteur.

Chacun des quatre complexes de détection enregistre en un an 10^{10} collisions, chiffre comparable au nombre d'étoiles de la Voie lactée. Le traitement des données ainsi accumulées est un défi, en raison tant des flux (de l'ordre du gigaoctet par seconde) que des volumes (plusieurs dizaines de pétaoctets chaque année, un pétaoctet étant égal à 1 000 téraoctets, soit 10^{15} octets).

À tout instant, quelques milliers de chercheurs dans le monde sollicitent des ressources de calcul et de stockage afin d'analyser ces données. Pour faire face à ce défi, une « grille de calcul » a été mise en œuvre.

Cette grille met en commun les ressources de calcul et de stockage de nombreuses machines indépendantes, chaque processeur traitant un seul événement. Les événements étant indépendants les uns des autres, peu de données sont à transférer entre les processeurs. Le principal intérêt de ce système est qu'il suffit d'ajouter de nouvelles entités pour augmenter d'autant les capacités de calcul de la grille, laquelle peut traiter des pétaoctets de données. La grille du LHC s'appuie ainsi sur près de 200 centres de calcul répartis dans le monde entier. Ce réseau est hiérarchisé en quatre niveaux, selon la nature et la qualité des services fournis, connectés par des liaisons à haut débit.

Pour traiter l'ensemble des données enregistrées par les expériences du LHC, les besoins en capacité de calcul atteints en 2012 équivalaient à 500 000 ordinateurs de bureau et plus de 200 pétaoctets d'espace de stockage sur disques. Un gigantisme informatique à la mesure du gigantisme (27 kilomètres de diamètre...) du LHC lui-même ! ■