

La mer des données

Jérôme Fenoglio

Ils veulent tout faire eux-mêmes. C'est la marque de fabrique des physiciens et aussi, parfois, leur défaut. Ancien collaborateur de Carlo Rubbia au CERN, Jean-Pierre Revol se souvient avoir trouvé le Prix Nobel de 1984 « un fer à souder à la main », à la fin d'une journée où il avait aussi bien pu écrire des algorithmes ou diriger la construction de son détecteur. Aujourd'hui, les tâches sont plus spécialisées, mais le principe demeure. Aucun domaine de la construction du collisionneur de particules, le LHC, n'a échappé aux scientifiques, qui donnent parfois l'impression d'avoir confié comme à regret des tâches aux entreprises sous-traitantes ou aux ingénieurs.

Ils ont même annexé de nouvelles activités. Ils se sont avancés dans les vastes territoires, très mal cartographiés en langue française, de la *computer science*, la science des ordinateurs. « C'est un métier nouveau pour nous. C'est la première fois que nous nous occupons à ce point de l'architecture de systèmes, de la distribution et du stockage des données », dit Faïrouz Malek, physicienne algérienne (Laboratoire de Grenoble, CNRS), responsable de la part française de ces activités informatiques.

Même le temps de transport de l'information, le long des torrents de fils qui sortent des millions de capteurs de l'appareil, compte

Les physiciens des particules se sont toujours comportés en producteurs boulimiques de données. « Les premières idées de circuits d'ordinateurs leur sont venues parce qu'ils en avaient marre de devoir crier simultanément, de chaque côté d'un détecteur, quand ils voyaient l'étincelle qui indiquait le passage d'une particule », raconte Jean-Pierre Revol, aujourd'hui membre de la collaboration d'Alice (l'un des quatre détecteurs de collisions de particules qui composent le LHC). Ils ont toujours eu aussi un rapport étroit avec les statistiques, ne serait-ce que parce que les comportements des particules qu'ils observent, régis par les lois de la physique quantique, ne peuvent être décrits que par des probabilités. Ce qui est nouveau, avec le LHC, c'est le changement d'échelle, une augmentation exponentielle du nombre de données à trier, conserver, dépouiller.

Pourquoi? Parce que dans cette discipline scientifique, la notion même de découverte n'a rien à voir avec les représentations que s'en font les profanes. Parce que les détecteurs du LHC n'apercevront pas, un beau matin, une particule inconnue. Parce que les physiciens ne tomberont pas immédiatement d'accord pour annoncer qu'ils ont enfin trouvé le boson de Higgs, l'objet de leur quête depuis tant d'années.

En réalité, personne ne verra vraiment ce boson, si tant est qu'il existe. Les physiciens ne pourront que déduire sa présence de la détection d'autres particules qui naissent, d'après les théories, de sa propre désintégration. Le problème, c'est que ces configurations particulières de photons, d'électrons, de muons, qui peuvent succéder à la matérialisation éphémère du boson convoité, sont produites à chaque seconde en quantité par les collisions du LHC.

Pour identifier les très rares événements qui pourraient signaler un boson de

Higgs, il faut d'abord connaître le nombre précis de ceux qui leur ressemblent mais ne sont pas causés par l'apparition de la particule. Cette foule d'individus anodins, les physiciens l'appellent « bruit de fond ». Cette toute petite quantité d'événements surnuméraires, ils la nomment « signal ».

A force d'accumuler des collisions, ces événements isolés s'ajoutent les uns aux autres. Sur les graphiques, le signal émerge de plus en plus nettement du bruit de fond. Quand la bosse qui se forme apparaît cinq fois plus élevée que la courbe des événements normaux, la découverte peut être officialisée. Ce surnombre, c'est là démonstration de la présence d'une nouvelle particule. C'est pour cela que les physiciens disent qu'ils « extraient » un résultat. C'est pour cela qu'ils ont besoin d'accumuler, pendant des années, des milliards et des milliards de collisions. C'est pour cela que le LHC les transforme en explorateurs, obligés, pour atteindre leur nouveau monde, de naviguer sur des océans de données.

Comment ne pas se laisser submerger? Comment s'orienter dans ces immensités? Comment conserver en sécurité les stocks d'informations à consommer au cours du voyage vers la découverte? Les physiciens ont été les premiers à devoir répondre à ces questions, qui commencent à se poser à tous les secteurs de nos sociétés développées. Comme pour le Web à l'époque, ce sont eux qui ont ouvert les nouvelles routes en imaginant, il y a plus de dix ans, des solutions pour le LHC.

D'abord, il a fallu se résoudre à une hécatombe. Aucun ordinateur actuel ne peut garder la mémoire de toutes les traces laissées par les collisions produites. Les physiciens ont donc choisi de ne conserver que les bruits de fond qui peuvent déboucher sur le signal d'une découverte. Tout le reste, soit 99 % des événements, jugés sans intérêt, part à la poubelle. Ce tri à la source est opéré par des circuits placés au plus près des détecteurs. Parce que même le temps de transport de l'information, le long des

torrents de fils qui sortent des millions de capteurs de l'appareil, compte. En quelques milliardièmes de seconde, ces circuits doivent décider si la particule qui vient de traverser le détecteur mérite d'être enregistrée, avant qu'une autre se présente. *« C'est une des parties les plus cruciales des expériences »,* dit Yves Sirois (Ecole polytechnique, CNRS). *Parce que nous savons que nous ne reverrons jamais les événements que nous avons choisi de ne pas garder.* »

Mais même avec ce tamis, les quatre détecteurs produiront, à plein régime, 15 pétaoctets par an. Ces millions de milliards d'octets représentent une pile de CD-ROM haute de 20 km. Pour la traiter, le CERN a imaginé la « grille », un réseau étendu au monde entier de centres de données qui mettent en commun leurs capacités de stockage et leur puissance de calcul. Grâce à ce système, tout membre d'une collaboration, où qu'il se trouve, peut accéder, quelques heures après les collisions, aux événements qui l'intéressent et se lancer dans ses propres analyses.

Car au bout du compte, la découverte sort toujours des tamis que les physiciens

plongent dans ces flux qui traversent la planète. Le meilleur extracteur de particules inédites sera toujours celui qui a l'intuition de l'endroit où il faut chercher, et qui s'est doté des bons outils. Cette nécessité est à l'origine d'une manie des collaborations : l'écriture du « code ».

Ce code, c'est l'ensemble des algorithmes, des ordres donnés aux ordinateurs pour qu'ils effectuent la bonne sélection dans les données, qu'ils reconstituent correctement la trajectoire des particules dans les détecteurs, et qu'ils compensent les défauts des capteurs ou éliminent les sources d'erreur. Chacun en rédige son bout qui vient compléter un immense ensemble de plusieurs millions de lignes, le corpus de programmation de chaque collaboration. Il est écrit dans la langue des programmeurs, barbare pour les non initiés, mais dans laquelle les meilleurs spécialistes affirment reconnaître le style de chaque rédacteur. *« C'est une activité qui comporte une dimension technique et un aspect presque littéraire »,* dit Federico Carminati, responsable du calcul dans la collaboration d'Alice.

Est-ce à cause de cet aspect esthétique

que le code provoque les affrontements les plus vifs au CERN ? Il nourrit des fiertés d'auteurs, qui peuvent dégénérer en âpres rivalités lorsque deux équipes écrivent deux morceaux de programmes différents mais pour le même usage, et qu'aucune d'elles ne veut renoncer à son œuvre de plusieurs années. Il a déjà déclenché ses batailles d'Hernani, notamment lorsque le CERN a décidé d'imposer, au début des années 1990, un cadre unifié pour les futures analyses du LHC, cette fois rédigé par des ingénieurs informaticiens.

« On m'a traité de tyrannosaure, on en est presque venu aux mains », se souvient René Brun (CERN), physicien et auteur de programmes qui font référence dans sa discipline, à qui l'on avait ordonné de ne plus s'occuper de rien. Les programmeurs avaient voulu faire entrer le code touffu, composé d'ajouts successifs et de collages plus ou moins orthodoxes, dans le moule d'un savoir-faire standard. Mais dès les premiers tests, personne ne put maîtriser la version « professionnelle ». Les collaborations sont revenues aux versions de René Brun, et nombre de physiciens se sont sentis confortés dans l'idée qu'eux seuls savent ce qui est bon pour leur science.

Le code a aussi ses petites mains : des jeunes physiciens qui maîtrisent souvent bien mieux que leurs aînés le nouveau langage de programmation. Le peuple des collaborations est ainsi composé de scribes en début de carrière, venus des meilleures universités. S'ils ne persévèrent pas dans la physique des particules, ils pourront toujours se recaser facilement, dans le secteur financier notamment, qui, avalanche de données oblige, a de plus en plus besoin de virtuoses de la statistique.

Ces geeks de tous les pays ont aussi importé au CERN les coutumes de leur temps. Nombre d'entre eux racontent leur vie et leurs travaux dans des blogs. Quitte parfois à rendre publics les petits secrets de leurs collaborations ou à alimenter des rumeurs sur les avancées des recherches. C'est un autre effet imprévu du Web : les grandes découvertes du LHC ont toutes les chances d'être divulguées par ces nouveaux chemins, et non par des publications scientifiques. Les organisations, si collectivisées, vont devoir composer avec ces forces centrifuges, ces nouveaux moyens, pour les physiciens, de mettre en valeur leur « signal » dans le « bruit de fond » de leurs immenses collaborations. ☺