



INSTITUT DE FRANCE
Académie des sciences

Séance solennelle de l'Académie des sciences / 16 juin 2009
Réception des nouveaux Membres sous la coupole de l'Institut de France

La gestion de données distribuées
Serge Abiteboul

Si à ses débuts l'informatique servait à "calculer", on a vite compris que c'était aussi un outil fantastique pour gérer de l'information. Cela a conduit aux systèmes de gestion de bases de données relationnels, développées dans la seconde moitié du 20^e siècle, qui ont rendu possible le traitement de gros volumes de données. Cette technologie est le fruit d'une recherche fondamentale qui s'est développée en lien étroit avec une industrie florissante. Elle est fondée sur des bases mathématiques rigoureuses et a développé notamment des ponts entre la théorie des modèles finis en logique mathématiques et la théorie de la complexité en informatique. C'est dans ce cadre que j'ai débuté ma vie de chercheur. Par exemple, un de mes résultats, avec Victor Vianu, établit que deux logiques de point fixe ont exactement la même puissance d'expression si, et seulement si, les programmes en temps polynomial et ceux en espace polynomial expriment les mêmes fonctions, un problème toujours ouvert en théorie de la complexité.

On commençait à peine à maîtriser la gestion de gros volumes de données sur des serveurs centralisés, que le web débarquait. En une vingtaine d'années à peine, il a révolutionné nos modes de travail et notre vie sociale. Du catalogue de vente d'Amazon, aux photos de Flickr ou l'encyclopédie Wikipedia, les bases de données sont au cœur des systèmes du web, avec des problèmes d'échelle étonnants, comme les milliards de pages indexées par Google ou les centaines de millions d'utilisateurs de Facebook.

Le monde de la gestion de données a profondément changé au delà même de ces problèmes d'échelle. Il a changé d'abord par la nature de l'information que l'on manipule, des structures plus riches comme les arbres XML et surtout moins rigides, plus dynamiques. Il s'est transformé surtout parce que l'information est de moins en moins centralisée ; chacun à sa mesure peut contribuer à l'édition d'information sur le web, à leur diffusion, à leur exploitation.

La gestion de données distribuées a longtemps constitué un problème difficile à résoudre du fait de l'hétérogénéité des machines, des systèmes d'exploitation, des modèles de données, des langages utilisés pour les applications. Le web et ses standards ont changé la donne. On dispose en particulier maintenant d'un protocole de communication entre machines, les "services web", qui permet de transformer facilement des données personnelles en des ressources accessibles partout. Cela conduit en particulier aux systèmes P2P, *peer-to-peer*, pair-à-pair en français.

À la base, un système P2P est constitué de machines autonomes, parfois en très grand nombre, qui coopèrent pour réaliser une tâche. Ce sont des pairs, égaux en droits, tour à tour clients et serveurs pour d'autres pairs. Le P2P est surtout célèbre pour le téléchargement plus ou moins légal de musique ou de film. Mais ses applications vont bien au-delà. Ses avantages sont considérables. D'un point de vue technique, en disposant de nombreuses machines (de leurs processeurs, de leurs mémoires, de leurs disques), on peut offrir de meilleures performances et une meilleure

disponibilité. Par exemple, dans l'approche centralisée, plus une information est populaire, moins elle est accessible, car le serveur est vite saturé, ce qui est pour le moins paradoxal. En P2P, c'est l'inverse. Un client ayant téléchargé l'information devient serveur à son tour. Surtout, de tels systèmes permettent d'utiliser les ressources innombrables disponibles sur le réseau et peu utilisées, et de se libérer des serveurs commerciaux qui stockaient nos données. Il devient possible de reprendre le contrôle sur sa propre information.

Les bases de données relationnelles disposaient de bases mathématiques solides. Ma recherche actuelle participe au développement de bases semblables pour les données distribuées. Ces bases devraient permettre de mieux comprendre les systèmes que nous utilisons, de les rendre plus performants, de mieux les contrôler.

Que peut-on attendre de ce domaine ? Toujours plus d'innovation ! En déplaçant l'information vers le réseau, les avancées scientifiques et technologiques du web ont radicalement étendu l'univers du possible. Elles peuvent aider à résoudre nos problèmes les plus graves, comme celui du développement durable. Elles s'accompagnent bien sûr aussi d'écueils, et pour n'en citer qu'un, la fracture numérique qui prive une partie de l'humanité de l'accès à l'information. Nous devons apprendre à maîtriser leurs effets pervers mais sans en restreindre les possibilités extraordinaires.

Je voudrais maintenant conclure par une pensée pour les collègues et étudiants avec qui j'ai eu tant de plaisir à travailler et sans lesquels je ne serais pas ici. Je les remercie et je vous remercie pour votre attention.