LE BIG DATA : UNE NOUVELLE FORME D'INTELLIGENCE COLLECTIVE (3A)

Mini-Série de Billets #3 : Des moteurs de recherche au Big Data. Et la résolution des conflits ?

Billet #3A – Amélioration des moteurs de recherche et naissance d'Hadoop

Internet et la Toile sont rapidement entrés dans les usages et dans les foyers. Cette percée a nécessité le déploiement d'ingéniosités techniques pour répondre de manière efficace aux requêtes toujours croissantes des utilisateurs. La performance des moteurs de recherche devait être améliorée pour les rendre capables d'analyser rapidement et de manière automatisée des millions de pages web. Quel défi ! A ce stade, nulle préoccupation Big Data, dans son sens actuel d'analyse de données massives. En revanche, la genèse du Big Data se trouve dans cette nécessité initiale d'améliorer l'efficacité des moteurs de recherche ; c'est ce lien qui porte le nom d'*Hadoop* que nous découvrirons dans cette mini-série de Billets #3. Les thèmes abordés seront les suivants :

- Naissance d'Hadoop
- Hadoop – une architecture distribuée
- Hadoop – des logiciels : le MapReduce et le Hadoop Distributed File System
- Hadoop – une évolution, un écosystème
- Hadoop et le Big Data
- Hadoop et la résolution des conflits

Hadoop en substance c'est…

Hadoop c'est l'implémentation d'une rupture conceptuelle avec les technologies classiques[i]. L'idée principale d'Hadoop consiste à « découper un gros volume de données en des petits bouts distribués à chacun des nœuds. La tâche est ainsi « parallélisée » entre de nombreuses machines, qui toutes contribuent au résultat. C'est ce genre d'outils qui permet de passer à l'échelle pour réaliser des calculs sur des téraoctets ou de pétaoctets[ii] de données. (…) Hadoop sait gérer les pannes. Si un « bout de calcul » s'est mal passé, Hadoop le détecte et demande

à une autre machine de reprendre uniquement le calcul erroné. Dans le pire des cas, des pannes fréquentes ralentissent un calcul, mais elles ne l'empêchent pas d'aller au bout »[iii]. Hadoop c'est donc des logiciels qui s'appuient sur une architecture distribuée pour fonctionner : un cluster d'ordinateurs.

Hadoop c'est également l'histoire d'une évolution : comment une solution initialement destinée uniquement à améliorer des moteurs de recherche est devenue un véritable écosystème d'outils de traitement et d'analyses de données massives adoptés largement par les entreprises de l'ère numérique, au-delà des entreprises de réseaux sociaux. « Chaque fois que vous verrez l'expression Big Data, Hadoop est impliqué d'une manière ou d'une autre » ai-je lu[iv]. Hadoop a évolué, s'est développé en un écosystème. Néanmoins, les concepts initiaux qui constituent son ADN demeurent valables. « La clé de la compréhension de la technologie ne se situe pas dans la technologie elle-même, mais dans la connaissance du contexte dans lequel elle a été créée. (…) Les principes ne changent pas dans le temps. Le nouvel Hadoop est juste l'amélioration de l'ancien »[v].

Comprendre ces principes n'est pas choses aisée, c'est difficile, technique. Analyser la transformation de l'écosystème juridique par le Numérique, vocation de mon Laboratoire Justice 2.0[vi], nécessite cet effort. Ces innovations représentent non seulement des changements technologiques majeurs, mais surtout des changements de paradigmes en termes d'architectures informatiques et de programmation. Or, ces innovations sous-tendent l'entier de notre société désormais numérique. Alors, questions : portent-elles en elles les germes de transformations sociétales en général, et dans les modes de résolution des conflits en particulier? Vraisemblablement…. Car comme le relève le sociologue Jean-Pierre BONAFE-SCHMITT, il existe un lien entre justice et société, un modèle de société engendre un mode de régulation correspondant[vii]. Quels sont ces changements ? Quid de la résolution des conflits à l'ère du Numérique ? Début de réponses dans cette mini-série de Billets #3.

Je me suis efforcée de limiter les développements techniques à ceux qui m'ont semblé être essentiels par rapport à ce double objectif : cerner la notion de Big Data et clarifier l'impact de ces innovations sur les modes résolution des conflits. A ceux qui trouveraient cette mini-série de Billets #3 trop technique, je vous comprends, mais le moteur de l'évolution passionnante de notre société est technique. A ceux qui souhaiteraient en apprendre davantage, l'ouvrage clair et pédagogique de Juvénal CHOKOGOUE *Hadoop : devenez opérationnel dans le monde du Big Data*[viii], sur lequel s'appuie cette mini-série Billets #3 nourrira je l'espère votre curiosité.

Améliorer les moteurs de recherche – un défi de taille

Google - moteur de recherche devenu la référence au point de donner un verbe : *googliser,* signifiant selon le très officiel Dictionnaire Larousse : « Rechercher des informations (en particulier sur quelqu'un) sur Internet en utilisant le moteur de recherche Google. (On dit aussi *googler*.) »[ix]. Google ne constitue certes pas l'unique moteur de recherche à disposition des internautes, il en existe d'autres[x]. Je l'évoque ici en raison de son rôle essentiel dans l'amélioration de l'efficacité des moteurs de recherche. Il lui incombe aujourd'hui de répondre en temps quasi réel aux millions de recherches effectuées chaque minute[xi] de par le monde, en indexant l'ensemble des pages web qui forment Internet et en recherchant à l'intérieur de

chacune d'elles les mots demandés. Moteurs de recherche et patience ne faisant pas bon ménage, le défi est de taille puisqu'on estime que passées 5 secondes sans réponse, un utilisateur va constater l'échec de sa requête, et poursuivre son chemin[xii].

Deux noms : *Doug Cutting*, *Hadoop* et un logo : un éléphant jaune, symbolisent la résolution du défi d'améliorer la performance des moteurs de recherche. Hadoop, solution trouvée par Doug Cutting cristallise toutefois l'influence de plusieurs projets émanant de sources diverses. Doug Cutting et Mike Cafarella initient un projet dénommé Nutch qui porte sur la conception d'un moteur de recherche open source. En juin 2003, ils présentent une version opérationnelle de démonstration d'un moteur de recherche regroupant 100 millions de documents[xiii]. Pour comparaison, « le nombre de sites internet dans le monde aujourd'hui est estimé à plus de 1 milliard avec une croissance de 5.1% (estimation faite par Netcraft) »[xiv]. Ils n'étaient pas les seuls à plancher sur ce défi. Google a également ressenti très tôt cette nécessité d'une gestion efficace des gros volumes de données liés aux requêtes des internautes, même si à l'époque elles étaient d'un moindre volume que de nos jours. La firme californienne a en outre été parmi les premiers à percevoir que la question du traitement des données ne pouvait plus se faire en appliquant les solutions du passé[xv]. En effet, avec le développement d'Internet, outre leur volume, des données de formats nouveaux – des données non-structurées - étaient générées. Les traiter excédait le cadre des statistiques traditionnelles : un texte, une photo, une vidéo, de la musique, ne peuvent en effet être rangés dans une base de données traditionnelle à l'instar, par exemple, des données opérationnelles des entreprises[xvi].

Doug Cutting a été inspiré par deux publications de Google qui présentent les avancées de la société de Mountain View en termes de traitement et de stockage de données distribués. La première, *The Google File System*[xvii], qui date d'octobre 2003, décrit un protocole de fichiers distribué, dispositif permettant notamment d'augmenter la puissance de traitement des recherches sur le Web[xviii]. La conscience de Google, évoquée ci-dessus, de la nécessité d'adopter une approche différente de ce qui avait prévalu jusqu'alors, ressort sans équivoque de cette publication : « (…) our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points »[xix]. L'année suivante, en décembre 2004, Google publie un autre article, *MapReduce: Simplified Data Processing on Large Clusters*[xx], qui explique « comment optimiser les modes de traitement et de calcul dans le contexte des données massives »[xxi]. En substance, Google a développé une nouvelle approche conceptuelle : i) distribuer le stockage des données (Google File System) et ii) paralléliser le traitement de ces données sur plusieurs nœuds d'une grappe de calcul (un cluster d'ordinateurs[xxii]) (MapReduce)[xxiii].

Dans l'intervalle, Doug Cutting a rejoint Yahoo ! amenant avec lui le projet Nutch. Le début de l'année 2006, voit la naissance de Hadoop, fruit de ces multiples influences évoqués ci-avant. Le projet Nutch a été divisé : les moteurs de recherche gardèrent le nom Nutch tandis que le calcul et le traitement distribués devinrent Hadoop[xxiv], du nom de l'éléphant en peluche jaune du fils[xxv] de Doug Cutting[xxvi].

Le MapReduce consiste en une approche conceptuelle qui pour être utilisée doit être implémentée ; c'est ce que l'on doit à Doug Cutting. Il a implémenté (exécuté) en langage de programmation JAVA[xxvii] le modèle algorithmique – la façon de programmer - MapReduce et le système de fichiers distribué de Google devenu *Hadoop Distributed System File* (HDSF)[xxviii]. Le couple conception-implémentation pourrait être illustré par un parallèle avec le domaine de la construction : la conception renvoie au plan d'un bâtiment dessiné par un architecte alors que l'implémentation se réfère à la réalisation du projet par un entrepreneur général. Le plan n'est pas utilisable en soi. Le passage du plan à la construction a transformé un concept (le bâtiment sur plan) en quelque chose d'utilisable (un bâtiment effectif).

L'histoire continue, Hadoop va se développer pour devenir de nos jours un véritable écosystème.

A ce stade, déjà beaucoup de questions

Avant de poursuivre sur l'évolution d'Hadoop, des questions demandent une réponse : Initialement, Hadoop, de quoi s'agit-il en termes simples ? Qu'est-ce que cette invention possède-t-elle de si astucieux qu'elle a participé à changer notre monde ? Quelles sont ses caractéristiques, ses idées principales ?

Pour y répondre, je procéderai en deux temps : d'abord, Hadoop – une architecture et ensuite, Hadoop – des logiciels. Cette division reflète les deux aspects d'un système informatique : l'aspect architectural (hardware, partie physique) et l'aspect logiciel (software, aspect virtuel), miroirs du double changement de paradigmes imposés par le Numérique. En raison de leur volume et de leur format (données non structurées), les approches classiques de traitement des données n'étaient plus suffisantes, il fallait innover, trouver une nouvelle approche. En bref, le temps était venu de dépasser les environnements de données traditionnels (structurées)[xxix], de changer de paradigmes[xxx] en termes d'architecture informatique et de programmation[xxxi].

A bientôt pour le Billet #3B *: Hadoop – une architecture distribuée, un nouveau paradigme infrastructure*l !

Anne-Sylvie Weinmann

Nota bene :

Tous les billets de cette série seront publiés sur LinkedIn mais également disponibles en format pdf avec les références détaillées sur le blog de mon site (www.medialien.ch).

An English version of this series of posts - Big Data: a new form of collective intelligence - will follow in a while.

Références & Notes :

[i] CHOKOGOUE Juvénal, *Hadoop : devenez opérationnel dans le monde du Big Data*, St-Herblain, ENI, 2017, p. 19 (https://m.editions-eni.fr/livre/hadoop-devenez-operationnel-dans-le-monde-du-big-data-9782409007613#).

[ii] Pour plus d'informations sur les octets et leurs dimensions, veuillez-vous référer au Billet #2 (http://www.medialien.ch/blog-fr170.html). Pour mémoire : un téraoctet (To) = 10^{12} octets, un pétaoctet (Po) = 10^{15} octets. Plus concrètement, « tous les livres jamais écrits ne demandent que quelques centaines de téraoctets en texte brut (sans image). Mais la quantité de données produites par le collisionneur de particules du CERN en une minute est de l'ordre d'une centaine de pétaoctets » (ABITEBOUL Serge, PEUGEOT Valérie, *Terra Data*, Paris, Le Pommier, 2017, p. 27 (https://www.editions-lepommier.fr/terra-data).

[iii] ABITEBOUL Serge, PEUGEOT Valérie, *op. cit*, pp. 71-72.

[iv] CHOKOGOUE Juvénal, op. cit., p. 24.

[v] CHOKOGOUE Juvénal, op. cit., p. 147.

[vi] Je vous invite à voir mon site : www.medialien.ch

[vii] BONAFE-SCHMITT Jean-Pierre, *La médiation : une justice douce.* Paris, Syros-Alternatives, 1992, p. 180.

[viii] CHOKOGOUE Juvénal, *Hadoop : devenez opérationnel dans le monde du Big Data*, St-Herblain, ENI, 2017, p. 19 (https://m.editions-eni.fr/livre/hadoop-devenez-operationnel-dans-le-monde-du-big-data-9782409007613#).

[ix] *Googliser*, http://www.larousse.fr/dictionnaires/francais/googliser/10910928#sBX0TjCTSj7Ullsg.99

[x] *Liste de moteurs de recherche*, https://fr.wikipedia.org/wiki/Liste_de_moteurs_de_recherche

[xi] Le Billet #2 rapportait l'évolution exponentielle de ces chiffres en millions. (http://www.medialien.ch/blog-fr170.html).

[xii] CHOKOGOUE Juvénal, *op. cit.*, p. 20.

[xiii] *Nutch*, https://fr.wikipedia.org/wiki/Nutch

[xiv] CHOKOGOUE Juvénal, *op. cit.*, p. 20.

[xv] CHOKOGOUE Juvénal, *op. cit.*, pp. 20 et 97.

[xvi] Je me pencherai dans le Billet #4 « Big au sens de Big Data » sur la question du format des données.

[xvii] *Extrait de GHEMAWAT Sanjay; GOBIOFF Howard; LEUNG Shun-Tak, The Google File System,* 10/2003 : « We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients. While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points. The file system has successfully met our storage needs. It is widely deployed within Google as the storage platform for the generation and processing of data used by our service as well as research and development efforts that require large data sets. The largest cluster to date provides hundreds of terabytes of storage across thousands of disks on over a thousand machines, and it is concurrently accessed by hundreds of clients. In this paper, we present file system interface extensions designed to support distributed applications, discuss many aspects of our design, and report measurements from both micro-benchmarks and real world use. » *(*https://research.google.com/archive/gfs.html).

[xviii] BABINET Gilles, *Big Data : penser l'homme et le monde autrement*, Paris, Le Passeur, 2015, p. 27 (http://www.eyrolles.com/Informatique/Livre/big-data-penser-l-homme-et-le-monde-autrement-9782368904923).

[xix] *Extrait de GHEMAWAT Sanjay; GOBIOFF Howard; LEUNG Shun-Tak, op. cit.*

[xx] *Extrait de* DEAN Jeffrey, GHEMAWAT Sanjay, *MapReduce: Simplified Data Processing on Large Clusters*, 12/2004 : « MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. Our implementation of MapReduce runs on a large cluster of

commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day. » (https://research.google.com/archive/mapreduce.html).

[xxi] BABINET Gilles*, op. cit.*, p. 27.

[xxii] *Grappe de serveurs*, https://fr.wikipedia.org/wiki/Grappe_de_serveurs

[xxiii] CHOKOGOUE Juvénal, op. cit., p. 21.

[xxiv] *Hadoop*, https://fr.wikipedia.org/wiki/Hadoop

[xxv] Interview de Doug Cutting sur les origines du nom et du logo d'Hadoop, 11/08/2015 (http://itsocial.fr/format/articles-decideurs/big-data-hadoop-interview-exclusive-de-doug-cutting-son-createur-video/).

[xxvi] *Hadoop – Tout savoir sur la principale plate-forme big data*, Le Big Data 07/02/2017 (http://www.lebigdata.fr/hadoop).

[xxvii] *Java*, https://fr.wikipedia.org/wiki/Java_(langage)

[xxviii] CHOKOGOUE Juvénal, *op. cit.*, pp. 23,74 et 115.

[xxix] BABINET Gilles, *op. cit.*, p. 28.

[xxx] Signifie, notamment, « une représentation du monde, une manière de voir les choses, un modèle cohérent du monde qui repose sur un fondement défini (matrice disciplinaire, modèle théorique, courant de pensée) » (*Paradigme*, https://fr.wikipedia.org/wiki/Paradigme).

[xxxi] CHOKOGOUE Juvénal, *op. cit.*, pp. 26, 74 et 97.