Originally published in French (07/11/2017), this post is a translation (mainly) made with the help of: www.DeepL.com/Translator

BIG DATA: A NEW FORM OF COLLECTIVE INTELLIGENCE (3A)

Mini-Series of Posts #3 : From search engines to Big Data. And, what about conflict resolution?

Post #3A – Improvement of search engines and birth of Hadoop

The Internet and the Web have quickly entered into habits and households. This breakthrough results from high technical ingenuity, find out how to respond effectively to the ever-increasing user requests was a complex enigma. The performance of search engines needed to be improved to make them capable of quickly and automatically analyzing millions of web pages. What a challenge! At this stage, no Big Data concerns, in its current sense of massive data analysis. Nevertheless, the genesis of Big Data lies in this initial need to improve search engine efficiency; it is this link, named Hadoop, that we will discover in this mini-series of Posts #3. The following themes will be tackled:

- Birth of Hadoop
- Hadoop - a distributed architecture
- Hadoop - softwares: MapReduce and the Hadoop Distributed File System
- Hadoop - an evolution, an ecosystem
- Hadoop and Big Data
- Hadoop and conflict resolution

Hadoop, in essence, it is...

Hadoop is the implementation of a conceptual breakthrough with traditional technologies[i]. Hadoop's main idea is to "*cut a large volume of data into small chunks distributed to each node. The task is thus "parallelized" between many machines, all of them contributing to the result. Such tools make scaling-up possible in order to perform calculations on terabytes or petabytes*[ii] *of data (…).Hadoop knows how to handle breakdowns. If a "calculation portion" went wrong, Hadoop detects it and asks another machine to carry out again the incorrect calculation only. In the worst case, frequent breakdowns slow down a calculation, but they don't stop it from going to the end*" [iii]. Hadoop is, therefore, softwares that rely on a distributed architecture to function: a cluster of computers.

Hadoop is also the story of an evolution: how a solution, initially intended only to improve search engines, has become an ecosystem of massive data processing and analysis tools, widely adopted by companies in the digital age, beyond social networking companies. *"Every time you see the expression Big Data, Hadoop is somehow involved"* did I read[iv]. Hadoop has evolved, developed into an ecosystem. The initial concepts that make up his DNA, nevertheless, remain valid. *"The key to understanding technology is not in the technology itself, but in the knowledge of the context in which it was created. (…) Principles do not change over time. The new Hadoop is just the improvement of the original one"*[v].

Grasping these principles is not an easy process, it is difficult, technical. To analyze the transformation of the legal ecosystem by the Digital - vocation of my Justice 2.0 Laboratory[vi] - requires this effort. These innovations represent not only major technological changes, but above all paradigm shifts in terms of computer architectures and programming. Yet, these innovations underpin the whole of our society,now digital. Some questions inevitably arise : do they carry with them the seeds of societal transformations in general, and in conflict resolution methods in particular? Presumably... Because, as the sociologist Jean-Pierre BONAFE-SCHMITT points out, there is a link between justice and society, a model of society generates a corresponding mode of regulation[vii]. What are these changes? What about conflict resolution in the digital age? Beginning of answers in this mini-series of Posts #3.

I have tried to limit the technical developments to those seemingly essential in the perspective of this twofold objective: to understand the concept of Big Data and to clarify the impact of these innovations on the conflict resolution methods. To those who would find this mini-series of Posts #3 too technical, I understand you, but the driving force behind the exciting evolution of our society is technical. For those who would like to learn more about theses technical aspects, the clear and educational book written (in French) by Juvenal CHOKOGOUE: *Hadoop: become operational in the world of Big Data*[viii], on which this mini-series of Posts #3 is based, will hopefully nourish your curiosity.

Improving search engines - a major challenge

Google search engine has become the reference to the point of giving a verb : to google (googliser or google in French), meaning according to the very official Larousse Dictionary :*"Search information (especially about someone) on the Internet using the Google search engine"*[ix]*.*Google is certainly not the only search engine available to Internet users, there are others ones[x]. I mention it here because of the essential role it played in improving search engines efficiency. Today, it must respond in near real time to the millions of searches carried out every minute[xi] around the world, by indexing all the web pages that make up the Internet and searching within each of them for the requested words. Search engines and patience do not get along very well ; the challenge is considerable since it is estimated that after 5 seconds without a reply, a user will notice the request failure, and move on[xii].

Two names: *Doug Cutting*, *Hadoop* and a logo: a yellow elephant, symbolize the won challenge of improving search engines performance. However, Hadoop, the solution proposed by Doug Cutting, crystallizes the influence of several projects stemming from various sources. Doug Cutting and Mike Cafarella launched a project called Nutch to design an open source search engine. In June 2003, they presented an operational demonstration version of a search engine containing 100 million documents[xiii]. For comparison, *"the number of websites in the world today is estimated at over 1 billion with a growth of 5.1% (estimate made by Netcraft)"*[xiv]. They were not the only ones working on this challenge. Google also discerned at a very early stage the need for efficient management of large volumes of data related to Internet users' requests, even if at the time those amounted a smaller volume than today. The Californian firm was moreover among the first to identify that the data processing issue could no longer be addressed by applying the solutions of the past.[xv]. In addition to their volume, as the Internet developed, data of new formats - unstructured data - were generated. Processing them turned out to be beyond the scope of traditional statistics: a text, a photo, a video, music, cannot be stored in a traditional database like, for example, companies' operational data[xvi].

Doug Cutting was inspired by two Google publications displaying the Mountain View company's advances in terms of distributed data processing and storage. The first one, *The Google File System*[xvii], dated October 2003, described a distributed file protocol, a design that namely increases the processing power of Web searches[xviii]. Google's abovementionned awareness of the need to take a different approach from what had prevailed hitherto unequivocally stood out from this 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 Google's abovementionned awareness of the need to take a different approach from what has prevailed hitherto stands out unequivocally from this publication"* [xix]. The following year, in December 2004, Google published another article, *MapReduce: Simplified Data Processing on Large Clusters*[xx], explaining *"how to optimize processing and calculation methods in the context of massive data"*[xxi]. Google had, in essence, developed a new conceptual approach: i) to distribute the data storage (Google File System) and ii) to parallelize the processing of this data on several nodes of a computer cluster[xxii] (MapReduce)[xxiii].

Doug Cutting had in the meantime joined Yahoo! bringing with him the Nutch project. The beginning of 2006 saw the birth of Hadoop, the result of the many aforementioned influences.

The Nutch project was divided: the search engines kept the name Nutch, while the distributed calculation and processing became Hadoop[xxiv], named after Doug Cutting's[xxv] son's yellow stuffed elephant[xxvi]**.**

MapReduce is a conceptual approach that must be implemented to be used; this is what we owe to Doug Cutting. He implemented (executed) in JAVA programming language [xxvii] the algorithmic model - the way to program - MapReduce and the distributed file system of Google which became *Hadoop Distributed System File* (HDSF)[xxviii].

The design-implementation duo could be illustrated by a parallel with the construction field: the design refers to the plan of a building designed by an architect, whereas the implementation refers to the realization of the project by a general contractor. The plan is not usable per se. The move from plan to construction has transformed a concept (the building on plan) into something usable (an actual building).

The story goes on, Hadoop will develop to become nowadays a true ecosystem.

At this stage, already many questions

Before moving ahead with the evolution of Hadoop, some questions need to be answered first: What is Hadoop, originally? What is so ingenious about it, that this invention has contributed to changing our world? What are its characteristics, its main ideas?

To answer these questions, I will proceed in two steps: first, Hadoop - an architecture and then, Hadoop - softwares. This division reflects the two aspects of a computer system: the architectural aspect (hardware, physical part) and the software aspect (software, virtual aspect), mirrors of the double paradigm shift imposed by the Digital. Due to their volume and format (unstructured data), traditional approaches to data processing were no longer sufficient; it was necessary to innovate and find a new approach. In short, it was time to move beyond traditional (structured) data environments[xxix], to shift paradigms[xxx] in terms of computer architecture and programming[xxxi].

See you soon for Post #3B : *Hadoop - a distributed architecture, a new infrastructural paradigm*!

Anne-Sylvie Weinmann
www.medialien.ch

References & 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] For more information on *bytes* and their dimensions, please refer to the Post #2 (http://www.medialien.ch/blog-fr170.html). As a reminder: 1 terabyte (TB) =10^{12} bytes, 1 petabyte (PB) = 10^{15} bytes. More concretely, *"all books ever written require only a few hundred terabytes of plain text (without images). But the amount of data produced by CERN's particle collider in a minute ranges in the order of a hundred petabytes"*. ABITEBOUL Serge, PEUGEOT Valérie, *Terra Data*, Paris, Le Pommier, 2017, p. 27 (https://www.editions-lepommier.fr/terra-data).

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

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

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

[vi] I invite you to see my website : 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] Translation of *Googliser*, http://www.larousse.fr/dictionnaires/francais/googliser/10910928#sBX0TjCTSj7Ullsg.99. In English : *Google (verb)*, https://en.wikipedia.org/wiki/Google_(verb)

[x] *Liste de moteurs de recherche*, https://fr.wikipedia.org/wiki/Liste_de_moteurs_de_recherche. In this Post, references and quotes from Wikipedia are usuallly from its version in French, translated in English for the purpose of this Post

[xi] Post #2 presented the exponential evolution of these figures in 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] I will look at the question of thedata format in Post #4 "Big in the sense of Big Data".

[xvii] *Excerpt of 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] *Excerpt of GHEMAWAT Sanjay; GOBIOFF Howard; LEUNG Shun-Tak, op. cit.*

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

[xxvi] Interview with Doug Cutting on the origins of Hadoop's name and logo, 11/08/2015 (http://itsocial.fr/format/articles-decideurs/big-data-hadoop-interview-exclusive-de-doug-cutting-son-createur-video/).

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

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

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

[xxx] Meaning, in particular, *"a representation of the world, a way of seeing things, a coherent model of the world based on a defined basis (disciplinary matrix, theoretical model, current of thought)"*. (*Paradigme*, https://fr.wikipedia.org/wiki/Paradigme).

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