The Big Data - Same Humans Problem

Alexandros Labrinidis

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA

labrinid@cs.pitt.edu

1. INTRODUCTION

Big data is transforming all aspects of the human experience, be it everyday life, scientific exploration and discovery, medicine, business, law, journalism, and decision-making at all levels of government.

On the one hand, big data is primarily driven by computing technology (i.e., computing power, data storage capacity, network capacity, etc.) becoming better and cheaper. This trend is captured by *Moore's law*[1] (i.e., the observation that, over the history of computing hardware, the number of transistors in a integrated circuit doubles approximately every two years) and by *Bezos' law*[2] (i.e., the observation that, over the history of the cloud, a unit of computing power price is reduced by 50% approximately every three years).

As expected, these advances in computing technology lead to exponential increases in the size of data generated (through simulations) and processed. However, they also translate to exponential increases in data collected by scientific instruments, be it centralized massive instruments such as the Large Hadron Collider[3] and the Large Synoptic Survey Telescope[4], or by great numbers of small-but-now-more-affordable instruments, such as those used for next-generation sequencing [1], or by even greater numbers of personal mobile devices and tiny sensors.

On the other hand, the *thirst* for data is becoming the norm both from a consumer point of view (e.g., businesses want to collect as much data as possible for their customers) and also from a producer point of view (e.g., people increasingly feel the urge to share more and more details of their lives on social networks), leading to an exponential increase in user-contributed content.

Despite the increases in computing technology and availability/demand for data in the last few decades, the performance of one critical component in the data processing pipeline has remained roughly the same. Namely, the ability of humans to process data has **not** changed significantly in the last few decades!

We refer to this disparity as:

the big data - same humans problem.

Therefore, when talking about the scalability aspect of big data, we need to make the distinction between two different types:

- *scalability from a systems point of view* – i.e., traditional scalability/performance measures, such as response time, throughput, scale-up, scale-out, etc., and
- *scalability from a human's point of view* – i.e., how well the system is making sure that human users do not get lost in a sea of data.

For example, assume a monitoring application that can process incoming data lightning-fast, let's say 1,000,000 inputs per second (so it can be considered scalable from a systems point of view). However, if all the alerts from this monitoring application (let's say a mere 1 per second) are simply output on a screen without any prioritization or classification, the application will clearly not exhibit good scalability from a human's point of view.

We believe that scalability from a human's point of view will soon become a requirement for successful big data systems and applications. As such, there are plenty of relevant research topics in the areas of summarization, ranking, recommendations, outlier detection, personalization, classification, visualization (e.g., visual analytics), and crowd-sourcing, to name a few. By crowd-sourcing, we are not only referring to techniques to farm out (small) tasks to cheap online labor, but also techniques to make sure that the "scraps" of effort performed by others, are fully utilized when they are very relevant to one's own work.

Finally, it is important to also consider the entire big data processing pipeline [2], in order to consider end-to-end performance [3], but also to enable cross-layer optimizations, that could lead to better scalability, from both a systems and a human's point of view.

Acknowledgments: This work was funded in part by NSF award OIA-1028162 and NSF Career award IIS-0746696.

[1] http://en.wikipedia.org/wiki/Moore's_law

[2] http://blog.appzero.com/blog/futureofcloud

[3] http://home.web.cern.ch/topics/large-hadron-collider

[4] http://www.lsst.org/lsst/

This article is published under a Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits distribution and reproduction in any medium as well as allowing derivative works, provided that you attribute the original work to the author(s) and CIDR 2015.

7th Biennial Conference on Innovative Data Systems Research (CIDR '15)
January 4-7, 2015, Asilomar, California, USA.

2. REFERENCES

[1] E. R. Mardis, "Next-generation dna sequencing methods," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 387–402, 2008.

[2] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big Data and Its Technical Challenges," *Commun. ACM*, July 2014.

[3] D. Abadi *et al.*, "The Beckman Report on Database Research." http://beckman.cs.wisc.edu, July 2014.