Towards a Framework for Executives and Decision Makers to Discriminate Big Data Projects for International Development

Driss Kettani

Ph.D., Al Akhawayn University in Ifrane, Morocco

Abstract— *In the context of International Development, Big Data Projects are often ill defined with a clear mix of terminology with trendy words such as Apps, GIS, Open Data, etc. often leading to overestimated budgets, unnecessary computing resources and unrealistic social outcomes. In this paper, we propose a framework for Executives and Decision Makers to allow them to clearly discriminate Big Data Projects regardless to the technical details related to this new Technology. The focus in our framework is rather on the broader Context, Objectives and Intended Outcomes of Projects.*

Keywords— Big Data, International Development, Decision Making Framework.

I. INTRODUCTION

I am currently involved in an interesting Project titled 'Harnessing the Economic Power of Data in the Middle East and North Africa', funded by the International Development Research Centre of Canada (www.idrc.ca). The goal of this project is to explore the Potential of Big Data Driven Innovations to improve Data Collection, Management, and Usability, particularly in the areas of Entrepreneurship and Youth Employment in the Middle East and North Africa Region. When I started consulting literature on Big Data, I was "hurt" by the vagueness of the different definitions and the confusion people have around the meaning and applications of Big Data, especially within the Social Science and Business Science Communities. In most international development projects where big data is highlighted, the motivations are not clear and not convincing at all. The Big Data Component of these projects is ill defined, with a clear mix of terminology with trendy words such as Apps, GIS, e-Business, Open Data, etc., often leading to overestimated budgets, unnecessary computing resources investments and unrealistic social outcomes.

This unhealthy situation calls for more conceptualization efforts from the research community to come out with a Model that allows the distinction between what is Big Data and what is not. Particularly, this Model targets Executives and Decision Makers in order to provide them with a conceptual framework that does not focus on specific technology enablers and/or associated features but, rather, on the broader Context, Objectives and Intended Outcomes of Project.

II. BIG DATA IN THE LITERATURE

The term Big Data has been used here and there to convey different kind of concepts including: huge quantities of data, social media analytics, real-time data, etc. One of the most accepted/cited definitions is that included in **META/Gartner** Report (Gartner 2001) that proposes a threefold definition encompassing three characterizing "Vs":

- **Volume (V_1):** refers to the quantity of generated and/or stored data. This quantity determines the value and potential insight of the application, and whether it can considered as big data or not. Generally, a threshold of several Petabytes (Millions of Billion Characters!) is required.
- **Velocity (V_2):** refers to the increasing speed at which data is created, processed, stored and analyzed by Database Systems.
- **Variety (V_3):** means that we can't anticipate the input data format and we need to expect rather unstructured/semi-structured input such as text, photo, audio, video, sensors, etc.

This "initial" definition of Big Data has been extended afterword to include two other important Vs (Gartner 2012), (IBM 2012):

- **Value (V_4):** business value that gives organizations a competitive advantage, due to the enhancement of their decision making process.
- **Variability (V_5):**to stress on the huge diversity of data sources (Sensors, Radars, RFID,GSM, IoT, etc.).

Oracle(Oracle 2012) avoids employing any Vs in offering a definition. Instead, Oracle contends that big data is the derivation of value from traditional relational database driven business decision making, augmented with new sources of unstructured data. Such new sources include blogs, social media, sensor networks, image data and other forms of data which vary in size, structure, format and other factors. Oracle, therefore asserts a definition which is one of inclusion of additional data sources to enhance existing operations. Oracle definition is also focused upon infrastructure, including: NoSQL, Hadoop, HDFS, R and relational databases.

For **Microsoft** (Microsoft 2013), Big Data is "the term increasingly used to describe the process of applying serious computing power - the latest in machine learning and artificial intelligence - to massive and often highly complex sets of information". This definition clearly indicates that big data requires the application of significant computer processing power and particular technologies including machine learning and artificial intelligence.

Intel(Intel 2012) links big data to organizations "generating a median of 45 terabytes (TB) of data daily". Intel describes big data through the experiences of its business partners and suggests that the organizations which were surveyed deal extensively with unstructured data and place an emphasis on performing analytics over their data. Intel asserts that the most common data type involved in analytics is business transactions stored in relational databases, followed by documents, email, sensor data, blogs and social media.

It is clear from these various definitions that Big Data is mostly associated to a technical trend and a technological evolution rather than clear business and/or societal needs. To the best of my knowledge, in the context of development, there isn't a single application where Big Data is necessarily needed to succeed the purpose and goals of the intended project.

III. A FRAMEWORK FOR EXECUTIVES AND DECISION MAKERS TO DISCRIMINATE BIG DATA APPLICATIONS

Before we introduce our framework, it is important to explain Big Data in simplerterms for Executives and Decision Makers to "better" understand the concept and its surrounding issues.Let's consider the "conventional" Data Models and their associated manipulation tools (DBMS Systems) and see how and why they are being challenged today.In fact, a number of well-known current DBMS Systems (including, Oracle, Access, DB2, FileMaker, Ingres, etc.) have been released in the beginning/middle of the eighties, i.e., much before:

- the Internet Democratization,
- the Telecommunication Revolution,
- the enabling of public use of localization devices, and other data capture devices (RFID, etc.).

Hence, these DBMS were implemented based on a number of fundamental hypotheses related to the technology status at that time, which definitely do not stand today, including:

- The Quantity and Speed of Data Generation/Update (Volume and Velocity in the literature) which is a simple consequence of the considerable growth of Data Sources, including human and "electronic" Stakeholders. On the human side, computers and/or similar equipment have been democratized and are being massively used for divers purposes (leisure, business, professional) leading to the generation of Terabytes of Data per minutes, mainly from social media, games and leisure Applications. On the Electronic stakeholders' side, Data generated by Sensors, Radars, RC, RFID, GPS, Cars, Trains, fixed and Mobile phones, etc. is simply gigantic! One needs to narrow the scope of desirable devices for a specific application to approximate the quantity of the potential useful data. And, further to the Quantity of generated data which is easily "overcomable" given the quick and considerable progress in data storage devices, the rate/ratio of Data production/update (massively, continuously, instantly and from everywhere!) is still impossible to entirely control through conventional DBMS Systems.
- The Nature of Data (Variety in the literature) DBMS systems work fine (and efficiently!) when the Data they process is well structured and organized, at the conceptual, logical and physical levels, following a specific Conceptual Model/Type. Some of the popular database models include relational models, hierarchical models, flat file models, object oriented models, entity relationship models and network models. Depending on the model in use, a database model can include entities, their relationships, data flow, tables and more. For example, within a hierarchal database mode, the data model organizes data in the form of a tree-like structure having parent and child segments. One noticeable problem that DBMS struggle to face today is the massive flow of UNSTRCTURED Data from different

devices and sources; any time, any language and any location! This new kind of Data has no particular format and/or structure nor does it have a predefined (or at least anticipated) size! All of the sudden, the DBMS industry found itself facing a strange kind of Data that could not be represented using any of the usual and known Data Structures! Due to the Social and Business opportunity this new kind of Data represent, considerable efforts are being made to process/exploit them the way we do with "normal" Data, but the path is still very long.

- The Sources of Data(Variability in the literature) which was basically supposed to be the Keyboard/Mouse (through data entry clerks). Although, DBMS evolved, with time, to include some extended input devices such as scanners, video cameras and security authentication devices (Face recognition, fingerprints, etc.), they became more and more "uncomfortable" with the "unexpected" advent of new input tools including GPS, RFID, Intelligent devices (TVs, Phones, domotics, etc.).
- The Business value of generated data(Veracity/Value in the literature):in conventional DBMS, the data that is stored is necessarily associated to a business value in the sense that, at any time, we are able to know what is true, what is false, what is fuzzy and what is "nothing"! In the way Big Data is generated it is difficult to categorize data with respect to its associated business value, which makes its processing almost impossible aside anecdotic manipulations and interpretations.
- The typical definition of Big Data using Vs (Volume, Velocity, Variety, etc.)is certainly good for Engineers and People who have a technical background on the topic and who conceptualize the meaning of these terms and their serious impact on the global Design and Implementation of a System and the enabling technologies to be used. Otherwise, a Petabyte is just a quantity, Variety of Data (Types) isgood for diversity and Velocity simply the process of data updates and accumulations! For Executives and Decision Makers, many questions arise when it comes to precisely identifying the importance and role of each specific V in "transforming" a conventional application into a Big Data Application (BDA), noticeably:
- Giving the increasing number of Vs (from 3 to 5 to 7 to 9 sometimes!) what are the Vs that are must and those that are not?
- Shall an application fulfill all Vs to qualify as a BDA, or simply fulfill some or any of them?
- Shall we consider all Vs of equal importance or not? If not, what is the respective weight of each V and why?
- What are the valid combinations of Vs and what are those that are not or "less" valid?
- Etc.

As far as we know, there is no answer to these crucial questions in the literature, and this cruelly contrasts with their importance in the decision making process, especially in the context of International Development Projects. There is a need for a more conceptual framework that allows the distinction between what is Big Data Application and what is not based on the broader Context, Objectives and Intended Outcomes of Project.

As a starting point for this framework, Decision Makers must distinguish between 2 clear different types of Big Data Applications:

Big Data Processors (BDP): these are systems whose main goal is to search, browse and explore Big Data Sets, for various purposes. Typical examples of this category are search engines (including Google), Analytics' tools, matching systems, people profiling, business targeting, etc. In this case the main discriminants are the size of the data to be processed by the system and its Veracity/Truth Value. The Data could be structured, semi-structured or unstructured; depending on how/why the Datasets have been produced. The two most significant challenges in this category are related to: (a). Datasets that are both Big and Divers; and (b). the estimated rate of pertinent data which determines the business opportunity of specific dataset.

Big Data Collectors (BDC): these are systems whose main goal is to gather data coming from different sources such as Sensors, GSM, Intelligent Tags, Antennas, Radars, and any connected device that transmits (with or without explicit solicitation) data packets of any type (tablets, mobile phones, computers, etc.). Typical examples of this category are Social Networks (possibly the hugest ever!), surveillance systems, RFID systems, intelligent home appliances, etc. The main

discriminants in this BDA category are related to the Sources/Devices that are being used to collect/generate the Data, including Speed, Type and Business Value of collected data.

Aside their significant difference in terms of technologies and functional goals, the distinction between these two Big Data Applications (BDA) categories is also due to the fact that BDC systems are prerequisite and prepare the ground for BDP to properly work. This means that without BDCs, BDPs would be meaningless/useless. All Commercial and Open Datasets have been built thanks to BDCs, before they get, sometimes, shaped and customized, to better fit specific clients' needs. It is quite common that BDC and BDP cooperate within the same business environment but their "Sequencialization" should mandatory take place (before running any BDP program), using batch processing techniques. For instance, Google (and most WWW search engine) first fetch web resources (through Google bot Tool) to discover new pages it did not visit beforehand. It then updates its huge table of contents (technically called index) that maps between users search entries and the available WWW Resources. When any user uses Google to search, the Search will be based on the last updated Index, independently of the work being done by Googlebot to fetch new resources. This perfectly illustrates the distinction between BDC and BDP Systems and their respective importance.

Now that we presented the two categories of BDA and their respective discriminant(s) and features, it is easier to state whether an Application X is BDA (or not) by classifying first as BDP or BDC, and then justify the necessary discriminants to fully validate it.

If we take the example of a GIS(Geographical Information System) application, is it a BDA or not? From a user point of view, a GIS is primarily a system that browses, searches, analyzes geographical maps, and provides/visualizes the results. Hence, if the extent of the map (spatial scope: building, quarter, city, country, earth region, entire earth, universe, etc.) and/or the details of the information the GIS provides (3D true color maps, diversity and comprehensiveness of thematic layers, photos, videos, comments, etc.) require a Petabytes' Database then definitely this GIS is a BDA. Remember that a Petabyte is 1000Terabytes and a Terabyte is a 1000Gigabytes which is equivalent to a BILLION characters! This is extremely huge as a size of Data and, as far as I know, only Google Earth (with its 20 Petabytes of geographical data as of Dec 2015) and some Military and/or Security agencies of powerful countries of the world do qualify as BDA!

Let's consider as another example, the case of the big brothers of retail and/or e-commerce such us as e-Bay, Target, Walmart, Amazon and the like, that are reputed in the open literature to be Big Data operators. These companies have a huge stock (hundreds of millions of items every day, updated on a real time basis) with RFID tagging, a huge number of clients all over the world, and a huge logistic (thousands of trucks all over north America, fully connected and updated on a real time basis). If we make a simple calculation to estimate the size of the data needed to handle such companies, we would realize that it (the size) certainly can reach the Petabyte range, but what is more data consuming (and more relevant to BDA) is not related to the managing of conventional operations/transactions but, rather, the "Intelligent" Management side. This includes the anticipation of clients' behavior/needs, the analysis of clients' comments on social media, clients profiling, market targeting, etc. These operations require a huge quantity of semi structured and unstructured data and "particular" algorithms to be processed; hence the need of Big Data Approach and Technologies. The reasoning we applied for the big brother of retail also applies to banking and payment systems, with some contextualization indeed.

New Social Media are Big Data Applications simple because they daily generate a huge quantity of data that is mainly unstructured. Most of this data is saved in the local infrastructure of the companies (Facebook, Twitter and the like) and then processed to produce commercial datasets for marketing, profiling, targeting, analytics, etc.For e-Health Systems, they are sometimes considered as BDA because data generated by medical analysis (and that must be handled afterword!) is both massive and unstructured. Each patient file contains further to the personal data and textual indications of Doctors, dozens/hundreds of picture, images and videos related to the medical evidences such as high resolution and true color Radiographies, Scanner images, RMI videos, and many others. E-Government Systems also qualify as BDA if they are seamless and handle physical administration evidence related to citizens and other stakeholders.

IV. CONCLUSION

We discussed in this paper the need of a more conceptual framework for Executives and Decision Makers to deal with Big Data Project in the context of International Development. The existing definitions of Big Data mostly target technicians and engineers and do not provide the necessary 'conceptual picture' to support the Decision Making process. We proposed a framework that is based on a Categorization of Big Data Application and some associated discriminants that are simple to

identify. Such a framework is badly needed in our current context where there is mushrooming of Big Data Projects for International Development with a clear mix of terminology and fuzzy justifications of Big Data.

ACKNOWLEDGEMENTS

This is to express my appreciation and gratitude to the International Development Research Centre (IDRC) of Canada for funding this Project. Without such assistance this Research could not have been achieved. I extend also my appreciation to my friends Dr Nasser Assem, Dr.Raed Sharif and Dr.Nagla Rizk for their insightful thoughts on the matter.

REFERENCES

[1] (Hanna 2010): Nagui Hanna, "e-Transformation: Enabling New Development Strategies", Springer Verlag, 2010.

[2] (Kettani and Moulin 2014): Driss Kettani and Bernard Moulin, "E-Government for Good Governance: a Practical Guideline for Decision Makers and Practitioners", Anthem Press, WPC/London, UK, Hardback | ISBN 9780857281258, June 2014.

[3] (Gartner 2001): L. Douglas. 3rd Data Management: Controlling Data Volume, Velocity and Variety, Gartner, 2001.

[4] (Gartner 2012): M. A. Beyer and D. Laney, "The Importance of Big Data: A Definition", Stamford, CT: Gartner, 2012.

[5] (H. J. Watson 2014): Hugh Joseph Watson, "Tutorial on Big Data Analytics: Concepts, Technologies and Applications", Communication of the Association for Information Systems, Vol. 34, Article 36, 2014.

[6] (IBM 2012): IBM, "What is Big Data? - Bringing Big Data to the Enterprise", http://www-01.ibm.com/software/data/bigdata/, July 2013.

[7] (Intel 2012): "Peer Research on Big Data Analysis", http:/ /www.intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html.

[8] (Mike2.0 2012): MIKE2.0, "Big Data Definition - the open source methodology for Information Development", http://mike2.openmethodology.org/wiki/BigData Definition.

[9] (Oracle 2012): J. P. Dijcks, "Oracle: Big data for the enterprise", Oracle White Paper, 2012.

[10] (V. Barneveld et Al. 2012): Angela van Barneveld, Kimberly E. Arnold, and John P. Campbell, "Analytics in Higher Education: Establishing a Common Language", ELI Paper 1: EDUCASE 2012.