Enhancing Capacity to Govern through Big Data

-Invited Paper-

Amita Singh[1]

Abstract

Most South Asian countries tend to treat Information and Communication Technologies (ICT) as a onetime adoption. Their institutions which govern the advancement of technology are relatively slower as compared to their neighborhood East Asian and Pacific countries. South Asian countries have spent a hefty sum on e-governance projects and invested heavily in ICT infrastructures. They have been fast to adopt ICTs and create cyber cities to expand business and marketing hubs so much so that ICT applications have brought a ‘data tsunami’. It is here that these countries suffer a phenomenal lack of trained personnel for reordering data and finding in it a key to growth. If governments do not simultaneously generate capacity to reorder, select and classify this uncontrollable flow of data, the most likely consequence would be derailment of GDP promotion efforts. South Asian countries need skilled personnel to analyze this almost arbitrary and wild communicational parameters of social media, marketing and commercial sites. Data needs to be analyzed, grafted and cleaned before it is stored in ICT storage spaces within each country. In terms of traditional public administration this is equivalent to storing file-information systematically in accordance to its subject, relevance and priority, subsequently discarding the waste unmindfully stuffed in office cupboards and storehouses. South Asian ICT infrastructure is likely to become an office which has unclassified and unmarked files littered all over its spaces to an extent that it becomes too overwhelming and gargantuan for managers to seek any information out of it. Most institutions such as legislatures, Judiciary and Election Commission to name a few encounter extreme challenges in their achievement graph.

1. Introduction

The impact of ICT on the functioning of governance institutions has come to a stage where some immediate and comprehensive steps should be taken. More than two and a half quintillion of data is produced every day in the world and 90 percent of all data today has been produced in the last two years. This indicates that

[1] Amita Singh is Professor at the Centre for the Study of Law and Governance (CSLG), Jawaharlal Nehru University, India. She is also the current Secretary General of the Network of Asia Pacific Schools and Institutes on Public Administration and Governance (NAPSIPAG). Email: amita@mail.jnu.ac.in

governance is likely to get buried or become irrelevant under the load of data. This directs attention towards the problem of organizing data. Big Data (BD) suggests that even digitization of information has reached its saturation point and is now to be stored through higher analytical skills in governance. These special skills are required for use in identifying content as well as their analytical relevance which could be used later or whenever required. Google's Eric Schmidt writes, "Our propensity for selective memory allows us to adopt new habits quickly and forget the ways we did things before."(p.8). BD enables us to keep track and simplify the crowding of scattered data which is creating a 'data tsunami' with the communication companies now. Ignoring this challenge may bring serious hurdles for public policy and for governance. It is also indicated that countries which delay attending to this problem may have to spend large sum of capital on retrofitting through the help of 'Big Data Analytics' from USA and despite that are likely to lose important information. If this capacity to manage Big Data is enhanced then many policies would become self-reflective, participative and relatively more inclusive since access and content simplicity which is the key to BD would enlighten citizens as well as governments. For example, the Human Resource programmes use BD to match positions to existing employees. One big problem in organizations is that the employees' profiles generally do not match the positions they get posted in due to their self-descriptions. HR departments scour through social media profiles, blogs and online conversations across the internet where talents and special skills are discovered for organizational requirements. BD helps to find out all details about the employee to post him/her where best suited. The Big Data expert from the IBM Company, Jeff Jones says, "*You need to let data speak to you"* and this is possible only when the unstructured data is converted into structured data.

2. Indispensable 'Big Data' for Public Institutions

For many reasons, Big Data is becoming an unavoidable fact of governance in present times. Governance being an overlapping team work between public, private and non-state philanthropic enterprises, organizations need to find better ways to tap into the wealth of information hidden in this explosion of data around them to improve their competitiveness, efficiency, insight, profitability and more (Eaton et al 2012). The realm of BD as Eaton and his group of IBM experts suggest is the analysis of all data (structured, semi-structured and un-structured) so that quick access to relevant information becomes easier for everyone. As Big Data experts have revolved around many 'Vs', it would be interesting to look into some of them here.

The volume of data being created every day is breaking through the storage spaces. In 2003 it was 14 trillions in a day which required five exabytes of space. This volume was produced in two hours in 2011 and 10 minutes in 2013. For an average service to 100 million customers, Customer Service Providers would need 50 terabytes of location data daily. If stored for 100 days it would need five petabytes as almost five billion records are created in a day. In 2010 in US records, the most popular service provider company AT&T had 193 trillion Customer Data Records (CDR) in its database.

The velocity of the data is also increasing. The global mobility data is growing at 78 percent of a compounded growth rate. Cisco Visual Networking Index (VNI-2013-2018), an ongoing initiative to track and forecast the impact of visual networking applications found that, 'Traffic' from wireless and mobile devices will exceed traffic from wired devices by 2016. By 2016, wired devices will account for 46 percent of IP traffic, while Wi-Fi and mobile devices will account for 54 percent of IP traffic[2]. In 2013, wired devices accounted for the majority of IP traffic at 56 percent. Overwhelmingly, the Global Internet traffic in 2018 will be equivalent to 64 times the volume of the entire global Internet in 2005 which suggests that bureaucracy and public officials may have to revise and reframe their capacity which would not be limited by their non-availability in office or by their multifarious tours as excuses for not attending and responding to important queries. To understand that much of the global Internet traffic which would reach 14 gigabytes (GB) per capita by 2018, rising by 5 GB per capita in 2013[3] would require additional capacities in the offices of public officials including the ability for BD analytics. As analytics is increasingly being embedded in business processes by using data-in-motion with reduced latency yet the *real time data*[4] which has to be catered to immediately and with urgency in every government, e.g., www.turn.com capacity of 10m/sec.

The variety of data is rising very fast in equivalence to its volume and velocity. The old time Data Warehouse Technology[5] used in the 1990s cannot be relevant anymore for the fact that public policy cannot depend upon an individual's understanding anymore. Besides a technically efficient administrator, what is also be needed is an equivalent expansion of key government offices towards an adoption of latest reporting tools, data mining tools (SPSS, etc.) and GIS to name a few. The data would come from various sources and would be transformed using

[2] VNI Report available at http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

[3] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

[4] Real-time data denotes information that is delivered immediately after collection. There is no delay in the timeliness of the information provided. It is of immense use to public officials as the 'Real-time data' is often used for navigation or tracking.

[5] A data warehouse is the data repository of an enterprise. It is generally used for research and decision support. For further details see Joseph M. Wilson's 'An Introduction to Data Warehousing'(a PPT from Storet Co.) and Samii, Massood (2004) International Business and Information Technology: Interaction and Transformation in the Global Economy, New Hampshire USA: Psychology Press.

Extract Transform Load[6] (ET) data inside the Warehouse. In earlier times this could be possible by untrained or less trained 'babudom' as it was more or less a structured content but to allow the earlier capacity to continue would be to play havoc with public policy. The public policy spaces would then be littered with consultants, each one asking for their fee and pulling information to their vested commercial interests. Currently, data content is unstructured for lack of a directed objective. Once policy formulation begins differentiation within larger objectives i.e., climate change as a main theme may add ever growing specificities such as coastal regulations, disaster risk reduction, ecosystem studies, disease control , food security and environmental changes then the need for Big Data to improve public policy formulation and implementation becomes important. To organize unstructured texts, sounds, social media blogs etc. government needs more enabling technology like the ones at IBMs Info-sphere stream platform.

Lastly but the most important requirement is the veracity (authenticity) of data for BD. Unlike governed internet data, BD comes from outside-our- control sources. Thus BD requires significant correctness and accuracy problems besides establishing and ensuring the credibility of data for target audience. Thus each Ministry of Government will have to first start with a basic data which routinely arrives at its posts and through analytics store it as Big Data. Right now much of the available data disappears or gets contaminated. Kevin Normandeau (2013) explains that BD veracity refers to the biases, noise and abnormality, the knowledge about which helps to clean the system. Many experts have added validity and volatility as important 'Vs' for BD. This may become important for the coming times when stored data could become outdated or irrelevant thereby suggesting a time period about its validity and also volatility. This is not so important for countries of South Asia which have yet to take their initial test drive on the BD highway.

3. Drivers for BD

In a compelling book of David Feinleib (2013) the author has tried to demystify Big Data as he emphasizes that to understand BD is to capture one of the most important trends of the present day world which surpasses every institutional boundary. The Changing governance paradigmatic requirements, e-governance expansion and rising number of internet and mobile users is a yeoman's task for routine administration to attend to. The new age citizen- customers are more sophisticated consumers who prefer to go on-line before taking a decision. Automation and convergence technology is speeding up faster with IVR, Kiosks

[6] ETL suggests three functions; ***e**xtract, **t**ransform, **l**oad*, combined together into one tool to pull data out of one database and transfer it to another database. Extract is the process of *reading data* from a database. Transformisthe process of *converting the extracted data* from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data. Load is the process of *writing the data* into the target database. This helps to either to shift data to data warehouse or to convert it into data marts which would store data for future usage as well as for marketing.

and mobile telephony usages penetrating the regions untouched so far with any market or governance activity. Information is being collected through a hub and spoke model in a number of South Asian countries but BD is still a distant requirement.

Methodologically, logical atomism can be seen as endorsement of *analysis*, understood as a two-step process in which one attempts to identify, for a given domain of inquiry, set of beliefs or scientific theory, the minimum and most basic concepts and vocabulary in which the other concepts and vocabulary of that domain can be defined or recast, and the most general and basic principles from which the remainder of the truths of the domain can be derived or reconstructed.

4. Origin of the Term

The origin of Big Data can be traced to the earlier analytical philosophers who discovered the mathematical logic in the way language is used. Ludwig Wittgenstein *TractatusLogico-Philosophicus*(1921) and Bertrand Russell's 'logical atomism' in his *Principia Mathematica* (1925-27, with A. N. Whitehead) inspired a debate on the fundamental building blocks of thought processes or an endorsement of analysis through which a given domain of enquiry can be defined and recast in a manner that remainder of the truths could be derived or accessed. Their logic of analysis suggested that the way human beings express themselves in their language propositions, paves the way for understanding the world more logically. Even the fundamental truths of arithmetic, are nothing more than relatively stable ways of playing a particular language-game. Big Data is a form of a revolution within ICT which paves the way for many more ideas to flow in as society advances.

It is said that the lunch table conversations during the mid 1990s at the Silicon Graphics featured the Chief Scientist John Mashey quite prominently. Douglas Laney, a veteran data analyst at Gartner declared John Mashey as the 'father of Big Data'[7]. However, the origin of the term is from scattered sources but as Victor Mayer-Schonberger and Kenneth Cukier (2013) simplify the debate by suggesting that the term has originated from the many debates on astronomy and genomics, sciences where data storage, correlation and retrieval leads to major breakthroughs in our understanding of the universe and well being of people.

It becomes fairly clear that Big Data originates out of the fundamental building blocks of language and culture which can be referred to as its genetics. The new digital forms of communication — Web sites, blog posts, tweets — are often very different from the traditional sources for the study of words, like books, news articles and academic journals.

[7] Lohr, Steve (2013) The Origins of 'Big Data': An Etymological Detective Story, New York Times,Feb.1.accessed http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/, 15.7.2014

5. How would Governance Benefit from Big Data?

5.1 Sophistication in Decision Making Tools

In earlier times the decision making involved no process except the whims and fancies of the rulers. Later it evolved into some scientific principles which formed the inflexible parameter of good decision making. Contesting this approach Herbert Simon indicated a behavioral approach to decision making but warned that a halo of preconceived thoughts around decision makers led to bounded rationality. Big Data minimizes the fuzziness of all approaches and brings logic of science in data corroboration, correlation, forecasting and predictability in decision making. It also helped in making policies more inclusive and decision making increasingly holistic, interdisciplinary and sustainable. Besides these issues, BD is also needed for improved risk management in business and in governance. A case is mentioned below.

In 2009 the FLU virus was discovered in USA. All strains were collected from the Bird Flu, Swine Flu and H1N1 and their correlation was established with the 1918 Spanish Flu which infected half a billion and killed tens of millions .The information had to be relayed back to central organizations and tabulated. This was a big challenge as officials visited this information only once a week which was a fatal time span for communicable disease spread. At such a time Google through its in house BD Analytics made 50m. Common searches that Americans share online and compared with the Communicable Disease Report Data on the spread of seasonal flu between 2003-2008. This correlation established a staggering 450m. Different mathematical models in order to test the search terms and finally helped in finding a solution. Without BD Analytics this was almost impossible or would have taken so long that the whole exercise would have become irrelevant.

5.2 Diagnostic Capability

Monitoring patient’s history, well being documents, nature of circulatory systems and frequency of infection can strengthen microscopic-long distance robotics which has enormous scope in telemedicine especially in the Third world and in Army locations. It has the ability to detect nascent heart attacks, early stages of cancer and also management of insulin levels.

Big Data has contributed to the Food and Drug Administration of USA in many ways.ie; Proteus Digital Health, a California based biomedical firm could kick-start the use of an electronic pill. It creates information which helps tissue engineering, genetic testing, DNA sequencing and source based solutions as well as early warning alerts on the basis of information corroboration and analytics.

5.3 Climate Change Related Early Warning Mechanism Systems

Climate change has brought substantial justification to have BD availability. The increasing inter-sectoral and inter-agency information such as the land, air and water bodies related changes, cloud formation, cyclones and hurricanes centred specialized data for over many hundred years and relationships to aquifers, flora and fauna, disasters and droughts, weather and crops etc. This expanse of information and the widening scope of its applicability in public policy has never existed prior to BD. Currently there is data and also the country and region based information which is scattered and much less accessed even during the period when the problem actually strikes. The meteorological data, density of population inhabitations, ecosystem services, local responses in the past to similar issues and urban planning records would combine in BD analytics to justify and enable retrofitting in decision making during troubled times of climate change.

6. Conclusion

South Asia has the world's largest number of poor. This region also has the largest number of governance challenges in terms of providing health, livelihood, education, skills and disaster mitigation and risk reduction infrastructure. There are many policy changes which have to be brought in through innovation, training and technology. Big Data is a mine of information to overcome and also escape many decisional catastrophes which are likely to come on the overloaded highway of government policies. This also requires balancing of a robust and secure public sector architecture that can accommodate the need for sharing data openly with all stakeholders. This further entails a commitment from national governments to reform and achieve well being for all citizens.

References

David Feinleib (2013) *Big Data Demystified: How Big Data Is Changing the Way We Live, Love and Learn*, USA: Big Data Group LLC.

Eaton, Chris, Deroos, Dirk, Deutsch, Tom, Lapis, George, and Zikopoulos, Paul (2012) *Understanding Big Data, Analytics for Enterprise Class Hadoop and Streamlining Data,* New York: McGraw Hill.

Normandeau, Kevin (2013) *Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity*, Sept. 12, Inside Big Data Available at http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/ (accessed on 20.6.2014).

Schmidt, Eric and Cohen, Jared (2013) *The New Digital Age: Reshaping the Future of People, Nations and Business*, New York: Alfred A. Knopf, Random House Publication.

Victor Mayer-Schonberger and Kenneth Cukier (2013) *Big Data: A Revolution that will transform how we live, work and think,* New York: Houghton Mifflin Harcourt Publishing Co.