Chapter 90
Opportunities and Challenges of Big Data in Public Sector

Anil Aggarwal
University of Baltimore, USA

ABSTRACT

Data has always been the backbone of modern society. It is generated by individuals, businesses and governments. It is used in many citizen-centric applications, including weather forecasts, controlling diseases, monitoring undesirables etc. What is changing is the source of data. Advances in technology are allowing data to be generated from any devise at any place in any form. The challenge is to "understand", "manage" and make use of this data. It is well known that government generates unprecedented amount of data (ex: US census), the question remains: can this data be combined with technology generated data to make it useful for societal benefit. Governments and non-profits, however, work across borders making data access and integration challenging. Rules, customs and politics must be followed while sharing data across borders. Despite these challenges, big data application in public sector are beginning to emerge. This chapter discusses areas of government applications and also discusses challenges of developing such systems.

BIG DATA

Data is everywhere. Data collection starts as soon as an individual steps out into the public domain. In many cases, data collection may even start within the private domain (when an individual is under surveillance). The data is a gold mine with hidden value. The challenge is to find the gold. Traditionally, data was collected and used to generate reports (ex: census reports) and maybe some intelligence (ex: homeland database). Data was well-behaved and could be represented in traditional relational or object-oriented form. This is changing as new data is generated from very different sources like social networking, weblogs, sensors and "smart" devices, to name a few. This data contains a "double gold" mine, which includes the gold value of the structured data and the additional gold value of unstructured data. The challenge is to find this "double gold". This data is being generated at an unprecedented rate. According to a recent IDC (2015) forecast "...the Big Data technology and services market will grow at a 26.4% compound annual growth rate to $41.5 billion through 2018, or about six times the growth rate of the overall information technology market. Additionally, by 2020, IDC believes that line of business

DOI: 10.4018/978-1-5225-7501-6.ch090

buyers will help drive analytics beyond its historical sweet spot of relational (performance management) to the double-digit growth rates of real-time intelligence and exploration/discovery of the unstructured worlds". Given the abundance of data it is not surprising there is a race to find "the double gold" in the data. Before we discuss application and challenges, a brief description of big data is presented.

According to Gartner.com, "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". Big data is also defined as data which cannot be processed by conventional means. The 4 Vs (velocity, variety, veracity and volume) have become the standard characteristics of big data.

Big data is typically unstructured and cannot be processed by conventional relational, structured query language (SQL), implying new techniques and new systems need to be developed, mastered and used. The next section briefly describes the big data system development process followed by opportunities and challenges.

BIG DATA DEVELOPMENT PROCESS

Big data system development process is similar to other systems, with the difference being the scale and nature of computing. The following steps are necessary for a project of this magnitude to succeed:

- Collect
- Manage/architecture
- Process
- Act

Each of these steps has its challenges. By some estimates, 90% of the data generated by devices is not useful. Therefore, the first step is to identify relevant data which requires functional expertise. Managing data would require data integration of structured and unstructured data. Several non-relational databases (key-value, key-document, graph) are emerging that can be used. However, the processing of this structure can be quite challenging. Typical processing is based on massive parallel processing that typically uses a HADOOP like structure. This requires creating clusters and replicating them. The task is to create the appropriate cluster to balance load. Once data is stored and defined it needs to be processed. Traditional languages like SQL are not appropriate. The challenge is to develop new languages that can help in processing. Once data is processed, the next step is to understand the output and apply it.

The next section describes opportunities in the public sector.

Opportunities of Big Data

Not all public sector applications are suitable for big data applications. Big data applications are suitable for the types of analysis that includes "discovery", "interrogation", "insights', etc. For example, they are suitable to study sentiment analysis (ex: citizen's reaction to a new highway), study terrorists cells (network analysis), and tracking diseases (ex: Ebola using network analysis). They are not suitable for traditional processing like tax record processing, creating budget analysis, census directory, etc. Figure 1 summarizes various applications of big data. The list is by no means complete since any public sector can find suitable big data applications.