Chapter 7
Big Data Models and the Public Sector

N. Nawin Sona
Government of India, India

ABSTRACT

This chapter aims to give an overview of the wide range of Big Data approaches and technologies today. The data features of Volume, Velocity, and Variety are examined against new database technologies. It explores the complexity of data types, methodologies of storage, access and computation, current and emerging trends of data analysis, and methods of extracting value from data. It aims to address the need for clarity regarding the future of RDBMS and the newer systems. And it highlights the methods in which Actionable Insights can be built into public sector domains, such as Machine Learning, Data Mining, Predictive Analytics and others.

INTRODUCTION

Big Data with its Volume, Variety and Velocity is an emerging area in the public sector. The public sector is a depository of enormous citizen related, infrastructure, economic, operational, statistical and archival data. Rapid growth in data with the 3Vs is matched by proliferation of new technological options to store, process and analyze them. However, despite the widespread use in the web based private enterprises, public sector adoption of Big Data technologies is limited. On a global scale, The United Nations E-Government Survey (2014) does not indicate extensive usage of Big Data technologies.

The use of Big Data has been pioneered in the USA, but is uneven across the public sector (Desouza, 2014). In the Executive of Office of the President of United States' factsheet, Big Data Across the Federal Government (2009), various uses are quoted from Departments of Defense, Homeland Security, Energy, Veterans Administration, Health and Human service among others. In USA, the TechAmerica Foundation-SAP Public Sector Survey (2012) is very indicative: although 63% of Federal and 76% of State IT officials say it is very important and extremely beneficial, the largest implementation barriers mentioned include a) privacy and policy concerns b) demonstrating the level of Return on Investment and c) lack of clear ownership of Big Data within the organization.

DOI: 10.4018/978-1-5225-7501-6.ch007

This gap needs to be filled by better understanding within the public sector about the technology and its methods and approaches. Palmer (2006) estimates that only half percent of all data is being analyzed for insight. Enormous data stand un-analyzed due to the limitations of traditional Business Intelligence tools, lack of flexibility of data integration and paucity of suitable algorithms to deliver analytics.

BACKGROUND

The newer Big Data technologies now enable the public sector to understand the patterns, and the ecosystem generating it, model the data, enable analytics, and in effect, extract actionable insights and eventually achieve prescriptive models. Various surveys, such as the World Economic Forum's "Global Information Technology Report 2014", have shown that there are effective and potential use cases of Big Data in public sector in improving Citizen Services, Medicine, Public Health, Education, Crime Mapping and prediction, emergencies such as epidemics, natural calamities and terrorist attacks, Urban Planning, Utilities planning, Power grid load balancing, Smart Cities, Smart Buildings, Predictive Modeling in fraud detection. Big Data also, has the potential for data driven policy making.

BIG DATA MODELS

Understanding the Data

Types of Data

Data is acknowledged as the "new oil" (Palmer, 2006). Today there is emergence of newer data formats and evolution of older ones. The RDBMS, SQL based systems relied on structured schema into which the highly cleaned and precisely formatted data was stored and processed. Structured data is a sine qua non in the transactional RDBMSs and for financial data where consistency and security reign supreme. Whereas, Semi-structured data (XML, JSON, RDF, YAML) are usually tagged and are Machine Readable and the aim is to make it Human Readable. And Unstructured data such as books, video, images, audio files, web pages, social media content, blog posts, comments on forums, streaming multimedia, are Human Readable data, and the aim is to make it Machine Readable. Other important data-types are Geospatial, Machine-to-machine (M2M) transmitted from mobile devices and sensors.

In formulating a Big Data deployment plan, it is necessary to understand the domain and its peculiar problems. The existing core transactional, operational data available for Business Intelligence (hereinafter BI), MIS, other ERP or CRM should be estimated. The frequency of data collection, the sources, whether streaming or static, freshness or staleness of data, frequency of updating and analysis are all factors that must be considered.

Big Data is useful where RDBMSs are limited by schema, query boundaries, cost of cleaning data and their lack of feasibility in processing unstructured and semi structured data. The question is not whether the choice is between Big Data and RDBMSs; but whether to use both Big Data *and* RDBMSs, or not. Hence, understanding available data relevant to the domain, which can be combined and mined for insight, would be crucial in designing a Big Data deployment model.