Chapter 6
On Efficient Acquisition and Recovery Methods for Certain Types of Big Data

George Avirappattu
Kean University, USA

ABSTRACT

Big data is characterized in many circles in terms of the three V's – volume, velocity and variety. Although most of us can sense palpable opportunities presented by big data there are overwhelming challenges, at many levels, turning such data into actionable information or building entities that efficiently work together based on it. This chapter discusses ways to potentially reduce the volume and velocity aspects of certain kinds of data (with sparsity and structure), while acquiring itself. Such reduction can alleviate the challenges to some extent at all levels, especially during the storage, retrieval, communication, and analysis phases. In this chapter we will conduct a non-technical survey, bringing together ideas from some recent and current developments. We focus primarily on Compressive Sensing and sparse Fast Fourier Transform or Sparse Fourier Transform. Almost all natural signals or data streams are known to have some level of sparsity and structure that are key for these efficiencies to take place.

1. INTRODUCTION

The scientific community as well as the intelligence agencies have traditionally led the field in collection and compilation of vast amounts of electronic data. Search engines (such as Google, Yahoo!, and Microsoft) and e-commerce started amassing exponentially increasing amounts of data starting in the early 2000's. After social networks, like Facebook or Twitter arrived, with hundreds of millions of users, electronic data collection increased to a level beyond imagination.

Deriving actionable information from the data collected has challenged the best minds in many disciplines. Efficient storage and retrieval of data on demand needed new thinking. From this need, many new technologies including the "Hadoop – MapReduce" ecosystem, with an ever increasing number of components was born. There are several scientific communities and commercial or public entities hard at

DOI: 10.4018/978-1-4666-9840-6.ch006

work to exploit this newest opportunity in spite of the unforeseen challenges in doing so. The traditional analysis of digital data was limited to one's own computing domain, often represented by an academic or corporate structure. However, with the advancement of computing and networking technologies that lead to big data, there seems to be a paradigm shift in what we even consider to fit the definition of "data".

The word "data" is readily conceptualized my most of us. However these concepts vary widely. Even most current dictionaries have generic and varying definitions of the term. According to Oxford dictionary, data means, "Facts and statistics collected together for reference or analysis". Oxford goes on to specify it meaning in Computing as, "the quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media" and its meaning in Philosophy as, "things known or assumed as facts, making the basis of reasoning or calculation." Merriam-Webster defines data as "facts or information used usually to calculate, analyze, or plan something or as information that is produced or stored by a computer". However, with the introduction of the World Wide Web in the early nineties and its success in providing connectivity to digital information everywhere (and the subsequent development of unforeseen levels of acquisition, storage, and analysis capabilities of digital information) one may wonder whether these definitions suffice what we consider as data.

Useful data can generally be considered as information of any kind that may evoke any of our senses about past or present. Such information often is embedded with high levels of sparsity and redundancy, especially in one or other of its alternate representations. Any event that has occurred or is occurring and could lead to some form of sensation or thought in one or more of us can be regarded as source of useful data. Data sources that interest us can perhaps be divided into two broad categories: data that can be attributed to humans, and data that can be attributed to non-humans.

Some examples of the first kind are e-mails, internet searches, tweets, articles (scientific or otherwise), creative works including audio and video, commercial transactions and the census. In this case since humans act as both the source and recipient, we have complete control of how the related data is perceived or interpreted. The second kind can be sourced mostly to observations of natural phenomena around us, as in oceanography, seismology, geology and meteorology, astronomy, high energy physics, biology, and chemistry. This type of data allows us perhaps our own impression or interpretation of what actually is taking place.

Analytics on both types of data holds promise. But strategies for analysis, however, may differ. The former will always be discrete and finite in size and dimension, no matter the volume, velocity, variety, or any other characteristics. At least theoretically, it may not need as much processing in acquisition, storage, and retrieval. The latter, on the other hand tends to be continuous and infinite in size and perhaps in dimension but full of sparsity and redundancy.

Regardless, analytics to divulge meaningful information from any data has better potential when they are used collectively through aggregation, composition, or integration. For example, individual transactions by themselves are unlikely targets for analytics (although perhaps with information gathered from analytics one can and may go back to subsets or individual data.)

Analytics, even after identifying patterns visually or otherwise, will largely have to be based more and more on principles of computational techniques, whether mathematical or statistical. Much of the information content in any type of data depends on the amount and type variation contained in it (Rudder, 2014). Identifying this would require on data in non-quantitative forms, (here forth we will assess only digital data regardless of source,) whether structured, unstructured, or qualitative, be transformed into quantitative. After all, analytics on big data is only promising if the analytical techniques used