Chapter 6
Big Data Analysis:
Big Data Analysis Pipeline and Its Technical Challenges

Rajanala Vijaya Prakash
S. R. Engineering College, India

ABSTRACT

The data management industry has matured over the last three decades, primarily based on Relational Data Base Management Systems (RDBMS) technology. The amount of data collected and analyzed in enterprises has increased several folds in volume, variety and velocity of generation and consumption, organizations have started struggling with architectural limitations of traditional RDBMS architecture. As a result a new class of systems had to be designed and implemented, giving rise to the new phenomenon of "Big Data". The data-driven world has the potential to improve the efficiencies of enterprises and improve the quality of our lives. There are a number of challenges that must be addressed to allow us to exploit the full potential of Big Data. This article highlights the key technical challenges of Big Data.

1. INTRODUCTION

Big Data has the potential to revolutionize much more than just research. Google's works on Google File System, MapReduce, and Hadoop, have led to arguably the most extensive development and adoption of Big Data technologies. They have become the indispensable foundation for applications ranging from Web search to content recommendation and computational advertising. There have been persuasive cases made for the value of Big Data for Healthcare - through home based continuous monitoring and through integration across providers (CCCc, 2011) Urban planning - through fusion of high fidelity geographical data, Intelligent transportation - through analysis and visualization of live and detailed road network data, Environmental modelling - through sensor networks ubiquitously collecting data (CCCd, 2011), Energy saving - through unveiling patterns of use, Smart materials - through the new materials genome initiative (National Science and Technology 2011), Machine translation between natural languages - through analysis of large corpora, Education - particularly with online courses (CCCb, 2011), Computational

DOI: 10.4018/978-1-5225-0182-4.ch006

social sciences - a new methodology growing fast in popularity because of the dramatically lowered cost of obtaining data (Lazar, D. et al. 2009), Systemic risk analysis in finance - through integrated analysis of a web of contracts to find dependencies between financial entities (Flood Jagadish, H.V., Kyle, A., Olken, F. And Raschid, 2011), Homeland security - through analysis of social networks and financial transactions of possible terrorists, Computer security - through analysis of logged events, known as Security Information and Event Management, or SIEM).

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved. There remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data is a major challenge, and is the one most easily recognized. Industry analysis companies like to point out there are challenges not just in *Volume*, but also in *Variety* and *Velocity* (Gartner Group 2011), and those companies should not focus on just the first of these. Variety refers to heterogeneity of data types, representation, and semantic interpretation. Velocity denotes both the rate at which data arrive and the time frame in which they must be acted upon. While these three are important, this short list fails to include additional important requirements. Several additions have been proposed by various parties, such as *Veracity*. Other concerns, such as privacy and usability, still remain. The analysis of Big Data is an iterative process, each with its own challenges, that involves many distinct phases.

2. BACKGROUND

Practically everything on the Internet is recorded. When you search on Google or Bing, your queries and subsequent clicks are recorded. When you shop on Amazon or eBay, not only every purchase, but every click is captured and logged. When you read a newspaper online, watch videos, or track your personal finances, your behaviour is recorded. The recording of individual behaviour does not stop with the Internet: text messaging, cell phones and geo locations, scanner data, employment records, and electronic health records are all part of the data.

Consider the data collected by retail stores. A few decades ago, stores might have collected data on daily sales, and it would have been considered high quality if the data was split by products or product categories. Nowadays, scanner data makes it possible to track individual purchases and item sales, capture the exact time at which they occur and the purchase histories of the individuals, and use electronic inventory data to link purchases to specific shelf locations or current inventory levels. Internet retailers observe not just this information, but can trace the consumer's behaviour around the sale, including his or her initial search query, items that were viewed and discarded, recommendations or promotions that were shown, and subsequent product or seller reviews. And in principle these data could be linked to demographics, advertising exposure, social media activity, offline spending, or credit history.

There has been a parallel evolution in business activity. Firms have moved their day to day operations to computers and then online, it has become possible to compile rich data sets of sales contacts, hiring practices, and physical shipments of goods. Increasingly, there are also electronic records of collaborative work efforts, personnel evaluations, and productivity measures. The same story also can be told about the public sector, in terms of the ability to access and analyze tax fi lings, social insurance programs, government expenditures, and regulatory activities.