International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

Big Data Challenges: Data Analysis Perspective

Riya Lodha[Å*], Harshil Jain[Å] and Lakshmi Kurup[Å]

[Å]Computer Engineering Department, D.J.Sanghvi College of Engineering, Mumbai University, Mumbai, India

Accepted 10 Sept 2014, Available online 01 Oct 2014, **Vol.4, No.5 (Oct 2014)**

Abstract

Big Data concern large-volume, growing data sets that are complex and have multiple, autonomous sources. Big Data are now rapidly expanding in all science and engineering domains with the fast development of networking and increase in the data storage and collection capacity. This paper reviews big data characteristics and challenges. In particular, we discuss big data variety, big data integration and cleaning, big data reduction, big data indexing and querying, and finally big data analysis and mining. We further focus on the challenges in big data analysis like heterogeneity, scale, timeliness, privacy and human collaboration; and present few real-world scenarios where these challenges are faced.

Keywords: *big data, challenges, big data analysis, mining, heterogeneous, information, complex.*

1. Introduction

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. The first presidential debate between President Barack Obama and Governor Mitt Romney on 4 October 2012 triggered more than 10 million tweets within 2 hours. Among all these tweets, the most discussed ones actually revealed the public interests, such as the discussions related to medicare and vouchers. As compared to generic media, such as radio or TV broadcasting, online discussions provide a new means to sense the public interests and generate feedback in real time.

These examples illustrate the increasing importance of Big Data in applications where the amount of data to be collected has increased exponentially. Also, the commonly used software tools are incapable to process such large amount of heterogeneous data within a feasible time limit. Hence, researchers have summarized three important aspects of big data that go beyond the ability of our current data processing technology. They are Volume, Velocity and Variety, also known as 3Vs.

2. Big Data

Gartner's definition- "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 concern large-volume, growing data sets which are complex, with multiple, autonomous sources. The size of such data is in petabytes (1 petabyte=1024 terabytes), exabytes (1 exabyte=1024 petabytes) or even zettabytes (1 zettabyte=1024 exabytes).

3. Big Data Characteristics

Big Data deals with large-volume. Big data seeks to explore complex and evolving relationships among data. It also has heterogeneous and autonomous sources with distributed and decentralized control. Hence, it is an extreme challenge for discovering useful knowledge from the Big Data.

3.1 Huge Data with Heterogeneous and Diverse Dimensionality

Big Data is the huge volume of data and is represented by its heterogeneous and diverse dimensionalities. This is because different schema for data recording is followed by different collectors of information. The data representations depend on the nature of different applications and are hence diverse.

3.2 Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources are able to generate and collect information without involving (or relying on) any centralized control. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors. More specifically, the local government regulations also impact on the wholesale management process and result in restructured data representations and data warehouses for local markets (Xindong Wu *et al*, 2014).

*Corresponding author: **Riya Lodha**

3.3 Complex and Evolving Relationships

The complexity and the relationships associated with Big Data increases with an increase in the volume of the data. For example, major social network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process on the data (Xindong Wu *et al*, 2014).

4. Challenges with Big Data

The existing computational techniques can be applied with some extensions to overcome at least some aspects of the Big Data problem. Fig. 1 shows the steps involved in Big Data processing.

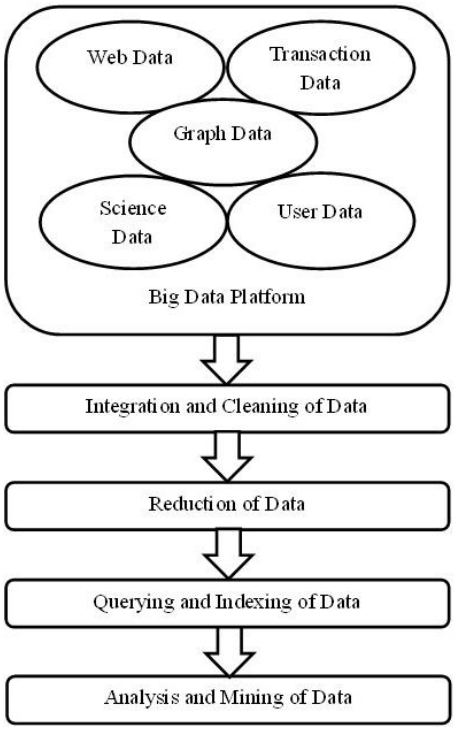

Fig. 1 Steps involved in Big Data processing

4.1 Data Acquisition and Recording

Big Data does not arise out of a vacuum: it is recorded from some data generating source. One difficulty is to define the filters so that they do not discard important information. A major part of this data is of no importance and thus can be filtered using the filters. And then it can be reduced by orders of magnitude.

The second big challenge is to generate the right metadata automatically. Thus, using it to describe what and how data is recorded and also the way it is measured. Human burden in recording metadata can be greatly minimized by metadata acquisition systems.

Data provenance is another important issue. Unless the information can be interpreted and carried, recording information about the data at its source is not useful along through the data analysis pipeline. Thus we need research both into generating suitable metadata and into data systems that carry the provenance of data and its metadata through data analysis pipelines.

4.2 Big Data Variety

Due to the various sources, heterogeneity is a natural characteristic of big data as discussed earlier.
User generated contents (UGCs): UGCs from applications with massive users. Examples are tweets, blogs, discussions, photos/videos posted and shared by users of many Web applications.
Transactional data: Transactional data that are generated by a large scale system due to massive operations/transactions processed by the system. Web logs, business transactions, feeds of moving objects, reports of sensor networks, reads of radio-frequency identifications are examples of big transactional data.
Scientific data: Scientific data that are collected from data-intensive experiments or applications. Examples are celestial data, high-energy physics data, genome data, healthcare data.
Web data: Web data that are crawled and processed to support applications such as Web search and mining. As the World Wide Web contains billions of pages, it is quite easy to generate a huge Web corpus of numerous unstructured Web pages.
Graph data: Graph data that are formed by a very huge number of information nodes, and the links between the nodes. Examples are social networks and RDF knowledge bases. Although structured, graph data are more expensive to process than relational data because ad-hoc local topology (pattern) in graphs complicates the processing of graph data (Jinchuan Chen *et al*, 2013).

4.3 Big Data Integration and Cleaning

Since the data is heterogeneous, it is not enough merely to record it and throw it into a repository. Data analysis is considerably more challenging than to simply locate, identify, understand, and cite data. All of this has to happen in a completely automated manner for effective large-scale analysis. For this, differences in data structure and semantics are required to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, a lot of additional work is required in order to achieve automated error-free difference resolution. Existing work on data cleaning assumes well-recognized constraints on valid data or well-understood error models; for many emerging Big Data domains these do not exist.

4.4 Big Data Reduction

Data reduction is the reduction of multitudinous amounts of data down to the meaningful parts. Further speaking, data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. On considering big data, making a profound transformation in computing, such as different sampling methods, aggregation (computing descriptive statistics), dimensionality reduction techniques, etc., is a feasible and

effective approach before big data analysis and management. Instead of operating on complex and large raw data directly, big data reduction tools enable the execution of various data analytic and manage tasks. Therefore big data open new chances and challenges to these techniques which have been well-studied in the past. Furthermore, big data open doors for interesting new approaches of data reduction.

4.5 Big Data Query and Indexing

In the era of big data, various forms of data from all kinds of fields have walked into every corner of our life. When it comes to query and indexing of big data, some challenges arise inevitably. First, the size of digital information in the age of big data is too huge for most softwares and people to manage and process. Also, a single machine cannot hold the sea-like big data, which should be stored in a distributed system. Therefore, the index of big data, distinguishing from the traditional index structure, should be built based on distributed system and corresponding new query theory should be encouraged. Second, big data not only refers to data sets that are very large in size, but also covers data sets that are complex in structure, high dimensional and heterogamous. It is all of these factures that make big data become a real challenge for us. Consequently, traditional methods for indexing and query with small structured data sets are not adequate any more (Jinchuan Chen *et al*, 2013).

4.6 Big Data Analysis and Mining

Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Mining requires clean and integrated data which is efficiently accessible, declarative query and mining interfaces, mining algorithms that are scalable, and big-data computing environments. Concurrently, the quality and trustworthiness of the data can be improved with the help of data mining. It will also facilitate understanding of the semantics of data and provide intelligent querying functions. Error correction and ambiguity removal is possible with the help of the knowledge developed from data.

The database systems host the data and provide SQL querying while the analytics packages perform data mining and statistical analyses which are forms of non-SQL processing. The current Big Data analysis lacks coordination between database systems and the analytics packages. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

5. Challenges in Big Data Analysis

5.1 Heterogeneity and Incompleteness

A great deal of heterogeneity is comfortably tolerated when humans consume information. In fact, the subtlety and richness of natural language can provide valuable depth. Machine analysis algorithms, however, expect homogeneous data, and cannot understand this subtlety. Hence, data must be carefully structured as a first step in data analysis.

Some incompleteness and some errors in data are likely to remain even after data cleaning and error correction. This incompleteness and these errors must be managed during data analysis. Doing this in a correct way is a challenge.

5.2 Scale

Size is the first thing anyone thinks of with Big Data. After all, the word "big" is there in the very name. A challenging issue for many decades has been managing large and rapidly increasing volumes of data. Earlier, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. A major shift underway now is that data volume is scaling faster than compute resources, and CPU speeds are static.

5.3 Timeliness

The flip side of size is speed. The time taken for analyzing the data is proportional to the size of the data set to be processed. The system design that deals with size of the data will also result in a system that can process a given size of data set faster.

There are many situations in which the result of the analysis is required immediately. We need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. Scanning the entire data set to find suitable elements is obviously impractical. Rather, to permit finding qualifying elements quickly, index structures are created in advance. But each index structure is designed to support only some classes of criteria. With new analysis on Big Data, new types of criteria are specified, and new index structures to support such criteria are needed to be devised. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

5.4 Privacy

The privacy of data is another huge concern, and it only increases in the context of Big Data. There is great public fear regarding the inappropriate use of personal data, due to linking of data from multiple sources. Managing privacy is a problem which has to be addressed from both technical and sociological perspectives in order to realize the promise of big data. A very important direction is to rethink security for information sharing in Big Data use cases.

5.5 Human Collaboration

There remain many patterns that humans can easily detect but computer algorithms have a hard time finding in spite of the tremendous advances made in computational analysis. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop.

In today's complex world, in order to really understand what is going on, it often takes multiple experts from different domains. A Big Data analysis system must share exploration of results obtained from multiple human experts. When it is too expensive to assemble an entire team together in one room, these multiple experts may be separated in space and time. The data system has to not only accept this distributed expert input but also support their collaboration.

A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. We also need a framework to use in analysis of such crowd-sourced data with conflicting statements.

As we all believe, once we own the good way to analyze and mine the big data, it can bring us the big value. However, the analysis and mining of the big data is very challenging due to its dynamic, noisy, inter-related, heterogeneous and untrustworthy properties.

Table 1 Challenges faced in real-world scenarios

Challenges in Big Data Analysis	Real-world Scenario
Heterogeneity	• Ancestry.com helps people build out their family tree and connect with their family history • To do so it maintains more than 11 billion records and 4 petabytes of **heterogeneous content**-historical records, birth records, death records, war and immigration records, even yearbooks—often in handwritten format
Scale	• The California Independent System Operator (ISO) manages electricity flow for 80 percent of California's power grid, delivering 289 million megawatt-hours annually to about 35 million consumers, more than 25,000 circuit-miles of power lines • It uses a software to correlate and analyze **massive volumes of data from multiples sources**—including weather feeds, sensors, metering equipment and more
Timeliness	• Oregon Health & Science University (OHSU) is a public university in Oregon with two associated hospitals, a level 1 trauma center and a children's hospital • It makes use of a software to track the **real-time location and status** of 4,000 infusion pumps for infusing fluids, medication or nutrients into a patient's circulatory system
Privacy	• Facebook, the most popular social networking website, encourages its users to use their real names and upload personal information on their profile page • Facebook makes use of two personal information aggregation techniques called "connections" and "instant personalization" that assure **anyone has access even to personal information you may not have intended to be public**
Human Collaboration	• DPR Construction is the general contractor for the $1.5 billion UCSF Medical Center at Mission Bay • DPR is using 3D technology to give its designers the ability to integrate data on air flow, building orientation, floor spacing, environmental resiliency, building performance, etc. into a single virtual model where the **information interacts in real-time, allowing architects, designers and construction teams to understand, visualize and interpret** hundreds of millions of data markers together in a fully operational environment

Conclusions

Managing and mining Big Data have shown to be not only a challenging but also a very compelling task. It is driven by real-world applications and key industrial stakeholders and initialized by national funding agencies. Big Data keyword literally deals with volumes of data; the key characteristics are 1) huge with heterogeneous and diverse sources of data, 2) autonomous with decentralized and distributed control, and 3) complex and evolving relationships in data. A "big mind" is required to consolidate data for maximum values; this is suggested by such combined characteristics.

1) Data integration- In order to utilize the information in big data, it is an important research topic about data integration in big data. Recently, researchers suggest utilizing users and/or crowds for improving the quality of integrated data. So crowd sourcing in data integration is a promising topic.
2) Data reduction- On considering big data, making a profound transformation in computing, such as different sampling methods, aggregation (computing descriptive statistics), dimensionality reduction techniques, etc., is a feasible and effective approach before big data analysis and management. Here the main challenge is how to run the traditional machine learning and statistics algorithms on big data.
3) Data querying and indexing- Big data querying and indexing need to modify the existing query optimization and indexing strategy in distributed system. Some traditional concerns to reduce I/O cost may not be useful in big data scenarios.
4) Data analysis and mining- The challenges in Big Data Analysis are 1) Heterogeneity and Incompleteness, 2) Scale, 3) Timeliness, 4) Privacy, 5) Human Collaboration.

Computing platforms that give high performance are required to support Big Data mining which impose systematic designs. This helps us view the full power of Big Data.

Finally, generally speaking, data is increasing with an exponential speed nowadays. However, corresponding information technology falls behind comparatively. Hence there is much remaining work for us to do about the data so that we could face the challenges brought by big data.

References

Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, Xuan Zhou (2013), Big data challenge: a data management perspective, *Frontiers of Computer Science*, 2013, Volume 7, Issue 2, pp 157-164 DOI: 10.1007/s11704-013-3903-7.

Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding (2014), Data Mining with Big Data, *Knowledge and Data Engineering, IEEE Transactions* on , vol.26, no.1, pp.97,107, DOI: 10.1109/TKDE.2013.109.