Big Data in Computer Science Research

(Trends, Challenges and Privacy Concerns*)*

Thomas Swed
Lewis University

***Abstract*— Among major areas of research today is that of data management and specifically, the challenges associated with "Big Data". Big Data is a term that refers to massive, complex data sets as well as the use of analytic methods to discern meaning and gather insights from data. This paper will discuss the major trends and characteristics that have led to the emergence of Big Data, the challenges associated with it, and concerns over privacy in the data mining process.**

Keywords—Big Data, Data Mining, Privacy, Data Trends, Cloud Computing

I. Introduction

Data collection has increased exponentially in recent years due to significant advances in smart devices, inexpensive storage, social software and the ability to connect many devices to a network, including cars and even our homes. Once data has been collected it must be processed, cleaned and analyzed. Improvements in multicore processors, solid state storage and open source software have led to the ability to process vast amounts of data. Some challenges of mining the data include scalable data infrastructures, dealing with diverse data sets and cloud services. Data mining refers to the process of analyzing data to discover insights and patterns. Throughout this process, a major concern arises over protecting an individual's sensitive information to prevent unwanted disclosure.

II. Big Data Characteristics And Trends

A. The Ubiquity of Data

Over the past several years there has been an explosion of data generation coming from a variety of sources. Information is being gathered by smart devices, cameras, microphones, and sensors. Many devices now have the ability to be connected due to the emerging Internet of Things (IoT), including home appliances, vehicles and other devices. Structured data is readily available on the Web, while efforts are being directed toward extracting and structuring information from software logs and sensors. Solid state storage as well as multicore processors continue to mature and have become increasingly inexpensive, further encouraging data storage and retrieval. Cost-effective approaches have developed to cultivate the volume, velocity and variability of large data sets, making them much less daunting. Information is both available and feasible to be utilized by virtually anyone. In short, data is everywhere, with much excitement over exploring the insights that can be gleaned.

B. Key Terms[1]

- *Volume* defines the amount of data generated and stored. As data continues to increase, greater value is placed on the aggregated result rather than individual records.
- *Velocity* refers to the speed at which information is generated. Challenges have arisen over the magnitude and complexity of information needing to be ingested.
- *Variability* measures the type and nature of the data, including text, images, video, audio and more. Incomplete and inconsistent data as well as incompatible formats often present some of the biggest difficulties to accurately use large amounts of information.
- *Value* quantifies the usefulness of data in decision making and analysis.
- *Complexity* refers to the relationship the data has to itself. Data sets are often dependent, connected and overlapping such that even a slight change in one more components can affect the behavior of the system as a whole.

C. Data Mining

Once data has been created and stored it must be interpreted if insights are to be discovered. Data mining, also referred to as knowledge discovery from data (KDD) is the process of obtaining useful discoveries or patterns from information through analysis [2]. There are four basic steps in this process [3].

1) Data preprocessing: this step includes selecting a specific subset of data relevant to the task at hand, cleaning the data to account for inconsistencies, and combining data from different sources.
2) Data transformation: data contains redundant or irrelevant information that is removed through tasks such as feature selection and feature transformation.
3) Data mining: tasks such as cluster analysis, anomaly detection, classification and regression are performed. This is where statistical and complex analysis happens.
4) Pattern evaluation and presentation: discernment is needed to determine applicable patterns and knowledge resulting from the mining process. The results are then

presented in a simple way, often with the aid of visualizations.

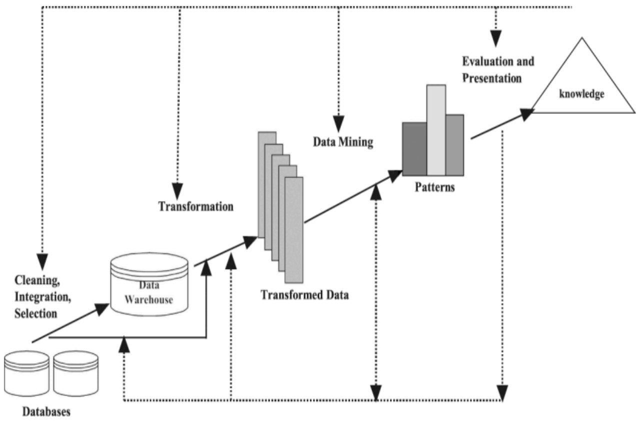

Fig. 1. An overview of the KDD process [4]

III. CHALLENGES

There are penetrating challenges faced when considering Big Data. Data management systems have significantly higher demands than traditional database systems. Because of these demands, the data mining process is labor-intensive and requires concentrated effort. Existing systems are being reexamined to develop solutions that will scale to the necessary demands of the data while at the same time remaining user-friendly. A few of the most common challenges are examined below, including developing scalable data infrastructures, coping with diversity, processing and understanding of data, and cloud services [5].

A. Scalable Data Infrastructures

With the onset of parallel processing, diverse and unstructured data has become much more scalable. Programming models such as MapReduce – an implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster -- have become increasingly popular [6]. Declarative and higher level languages have applied these models over existing ones to enable the use of accessible Big Data platforms. Due to the adoption of declarative languages for processing information, traditional query technics are being pursued. In order to effectively query a large data set, a great amount of parallelism over large clusters of processors is needed. In order to compensate for these strains and reduce resource demands, processors are being developed to integrate data sampling, data mining and machine learning calculations into their work flows.

B. Coping With Diversity

Traditionally, data has been stored in structured relational databases while using a declarative language to query it. Increasingly more information is being generated and stored in an unstructured, mixed form that relational databases do not handle well. Incompatible data formats, incomplete data, non-aligned data structures, and inconsistent data semantics represent significant challenges when interacting and querying these data sets [7]. To handle this diversity, platforms are needed that can integrate and manage structured as well as unstructured data. These systems must be adapted to mix multiple types of data processing, such as querying data with SQL and then analyzing it with R [8]. The ability to integrate different systems to maximize processing power still needs developing.

C. Processing and Understanding Data

Currently, there are remarkably few tools that process data from start to finish. The process of taking data in its raw form to gather insights requires significant effort through many steps along the way. Many steps also require work to be performed by someone with significant computer abilities, including experience with programming and analytic software. The tools that do exist are typically expensive and focus on specific steps in the method. Significant improvements in technologies to deal with each step of the process are needed, along with a way to combine these technologies into comprehensive systems.

D. The Cloud

Reduced hardware costs, scalability and the relative ease of administration have led to widespread adoption of Cloud Computing. Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [9]. It is offered in a variety of models. The three most standard are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [10].

- *Software as a Service (SaaS)* offers access to software applications through a monthly or yearly subscription. It eliminates all of the overhead involved with deploying an application (licensing, installation, support and updating) and provides accessibility and scalability.

- *Platform as a Service (PaaS)* allows users to develop, run and manage applications by providing access to hardware. The user only has control over the deployed application and configuration settings however, rather than the network, servers, operating system or storage. Examples include Amazon Web Services (AWS) and Windows Azure.

- *Infrastructure as a Service (IaaS)* provides hardware and other necessary components to a user, allowing a greater level of control over the system.

Because data does not lend itself well to adaptability and change, building a Data Platform to account for continually changing storage and networking demands is a tremendous feat. Data replication presents challenges to account for scaling applications. When using Platforms as a Service as a data platform, tasks that would have been performed by a database administrator (DBA) in a traditional database would need to be automated to account for variance. Solutions are still being developed to solve these issues.

IV. Privacy Concerns

As data continues to be generated and stored, a security risk is presented for unwanted disclosure of personal or sensitive information. This is particularly a concern with cloud based platforms as users have little to no control over the security of the data sets. Service providers are able to access the data at any time, without a user knowing. In order to gain insights through analysis from information, some amount of personal data must be provided. The challenge in this instance is to protect an individual's privacy while performing data mining algorithms. In an attempt to address these concerns, a research process known as privacy-preserving data mining, or PPDM, has been developed. To better address privacy concerns, users may be divided into four separate categories, each with different solutions [11].

1) *Data Provider* refers to anyone who provides information to be used in the analysis process. Unfortunately, once personal information regarding an individual has been provided to a company or stored on the cloud, there is always the possibility of it being accessed undesirably, either through hacking or dishonest practices. Security measures have been developed to encrypt sensitive data, prevent websites from tracking a user's online activities, and block unwanted scripts or advertisements from running.

2) *Data Collector* refers to anyone gathering information from a data provider and passing it on to a data miner. The main concern of the data collector is to ensure that a use's sensitive information has been anonymized or modified before passing it on to be analyzed. This process is referred to as *privacy preserving data publishing (PPDP)* [12]. While there are several methods for performing PPDP, it is the job of the data collector to utilize an appropriate method to protect sensitive information.

3) A *Data Miner* is anyone performing an algorithm on data to perform analysis and discover insights. The concerns for a data miner include discovering sensitive information through the mining process and working with other analysts. In either case measures must be taken to modify or remove sensitive information before it is published or analyzed by another data miner. Approaches to achieve this are known as *privacy-preserving data mining* (PPDM) [13].

4) *Decision Maker* refers to anyone who uses analyzed data to make decisions. While much responsibility rests on the data collector and data miner to safeguard privacy, a decision maker is also responsible for ensuring that the results presented to him are protected and only distributed to authorized parties.

V. Conclusion

We live in an evolving digital age. As Big Data continues to generate and develop, challenges are presented to manage, analyze and protect it. Through disciplined approaches to handling this data and overcoming its challenges, many are recognizing the potential of exciting new discoveries.

References

[1] Walunj Swapnil K., Yadav Anil H., Yadav Riteshkumar and Yadav Satish L.. Article: Big Data: Characteristics, Challenges and Data Mining. *IJCA Proceedings on International Conference on Advances in Information Technology and Management* ICAIM 2016, July 2016, 2-3.

[2] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren. Article: Information Security in Big Data: Privacy and Data Mining. *IEEE Access (Vol 2.)*, October 2014, 1149.

[3] Xu, 1149.

[4] Xu, 1150.

[5] Daniel Abadi et al. Article: The Beckman Report on Database Research. *Communications of the ACM, Vol. 59 No. 2,* February 2016, 94.

[6] Jeffrey Dean, Sanjay Ghemawat. Article: MapReduce: Simplified Data Processing on Large Clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation,* December 2004, 1.

[7] Swapnil K., 2-3.

[8] Abadi et al, 96.

[9] Peter Mell, Timothy Grance. The NIST Definition of Cloud Computing, *National Institute of Standards and Technology: U.S. Department of Commerce.* September 2011, pp. 7.

[10] Qusay F. Hassan, Article: Demystifying Cloud Computing. *CrossTalk Magazine,* January/February 2011, pp. 2.

[11] Xu, 1151.

[12] Xu, 1154.

[13] Xu, 1160.