Review

Big Data: Infrastructure, technology progress and challenges

Lidong Wang¹ and Cheryl Ann Alexander²

¹Department of Engineering Technology, Mississippi Valley State University, USA
²Technology and Healthcare Solutions, Inc., USA

Received 13 April, 2015; Accepted 31 June, 2015; Published 17 July, 2015

Big Data is multidisciplinary. It has great impacts on scientific discoveries and value creation. It has potential applications in government administration, Homeland Security, life sciences and health care, manufacturing, and supply chain management and global business, etc. Big Data process and value chain covers data generation, collection, cleaning, migration, integration, storage, analysis and modeling, value mining and discoveries, visualization, and data governance and security. This paper deals with Big Data characteristics, infrastructure, impacts, applications, and technology progress. Big Data challenges and future research are also presented in this paper.

Key words: Big data, Big Data analytics, data collection, data management, data mining, machine learning, information security, hacking, Hadoop, MapReduce algorithm

INTRODUCTION

Big data is a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques (Demchenko et al., 2013). Characteristics of big data can be categorized into volume, velocity, variety, veracity, and value.

Volume: It means data size.

Velocity: It suggests that information is generated at a rate that exceeds those of traditional systems (O’Leary, 2013).

Variety: This refers to heterogeneity of data types, representation, and semantic interpretation (Jagadish et al. 2014).

Veracity: This includes two aspects: data consistency (or certainty) and data trustworthiness (Demchenko et al., 2013). It refers to the accuracy, truthfulness, and reliability of the data (O’Leary, 2013).

Value: It is defined by the added-value that the collected data can bring to the intended process, activity, or predictive analysis/hypothesis. Data value is closely related to the data volume and variety (Demchenko et al., 2013). Big data is often noisy, dynamic, heterogeneous, inter-related, and untrustworthy. Nevertheless, even noisy big data could be more valuable than tiny samples (Jagadish et al., 2014). Outliers can be even more interesting because critical innovations or revolutions may happen outside the average tendencies (George et al., 2014).

High volume data have following five key sources (George et al., 2014):

Public data: Held by governments, governmental organizations, and local communities.

Private data: Held by private firms, non-profit organizations, and individuals.

Data exhaust: Can be ambient data that are passively collected, non-core data with limited or zero value. It also can be from information-seeking behavior such as Internet searches and telephone hotlines. This can be used to infer people’s needs, desires, or intentions.

*Corresponding author: lwang22@students.tntech.edu
Table 1. Types of Big Data

<table>
<thead>
<tr>
<th>Type</th>
<th>Data descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Data stored in a fixed field Relational database, spreadsheet, etc.</td>
</tr>
<tr>
<td>Unstructured</td>
<td>Data not stored in a fixed field; text-recognizable documents, image/video/voice data, etc.</td>
</tr>
<tr>
<td>Semi-structured</td>
<td>Not stored in a fixed field but including metadata, schema, etc.; e.g., XML, HTML text, and so forth.</td>
</tr>
</tbody>
</table>

Source: Chun and Lee, 2014 modify

Community data: Being often unstructured data—especially text such as consumer reviews on products.

Self-quantification data: Revealed by the individual through quantifying personal actions and behaviors such as monitoring exercise and movement via wristbands.

Variety indicates that there are various types of data, and they could be classified into structured, unstructured and semi-structured data sets. Table 1 shows the three types of big data.

BIG DATA INFRASTRUCTURE, IMPACTS AND APPLICATIONS

Big Data includes the following infrastructural and architectural aspects (Bellini et al., 2013):

Scalability: This may have impact on several aspects of the big data solution (e.g. data storage, data processing, rendering, computation, and connection, etc.). In most cases, scalability is obtained by using distributed and/or parallel architectures.

High availability: Availability refers to the ability of the community of users to access a system exploiting its services. A high availability leads to increased difficulties in guarantee data updates, preservations and consistency in real time.

Computational process management: Big Data has to cope with the needs of controlling computational processes by means of allocating them on a distributed system, putting them in execution on demand or periodically, killing them, recovering processing from failure, returning eventual errors, and scheduling them over time, etc.

Workflow automation: Big data processes are typically formalized in the form of process workflows from data acquisition to results production.

Cloud computing: The cloud capability helps Big Data to obtain much storage space and computation power.

Self-healing: This refers to the capability of a system to autonomously solve failure problems.

Big data sources include Internet data, location data (e.g. mobile device data and geospatial data), images (e.g. surveillance and satellites), transaction, social media, and devices data (e.g. sensors and RFID devices), etc. (Chan, 2013). Streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge from what is happening now, allowing organizations to react quickly. Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring, log records or click-streams in web exploring, manufacturing processes, email, blogging, Twitter posts, and others (Bifet, 2013). In order to integrate data from multiple heterogeneous data sources, two processes are needed: data migration and data integration. Data migration is to retrieve and collect data from the data sources and store them into the third data source in a format specified. Data integration processes are: first, to check if each data exists in the Database (DB) and then update data; second, to remove or combine duplicated data from the heterogeneous data (Woo, 2013).

Big Data has great impacts on value creation, improved productivity, and scientific discoveries, etc. Its domains and application include: government, Homeland Security, life sciences, physical sciences, education, health care, manufacturing, transportation, smart cities, retail, location-based services, supply chain management, and communication and media, etc. (Zhang, 2013).

The integration of linguistic methods with traditional e-discovery techniques was proposed to identify deceptive texts within a given large body of written work, such as email. A set of linguistic features associated with deception were identified and a prototype classifier was constructed to analyze texts and describe the features’
METHODS AND TECHNOLOGY PROGRESS OF BIG DATA

Big Data technology can be in two situations. One is big volumes of data, but "small analytics." The focus is on running SQL analytics (count, sum, maximum, minimum, and average) on large amounts of data. The other is big analytics on big volumes of data. Big analytics means data clustering, regressions, machine learning, and other complex analytics on very large amounts of data. At the present time, people tend to run big analytics using statistical packages, such as R, SPSS and SAS (Stonebraker, 2013).

The technology for big data include stream processing, data mining, machine learning, cloud computing, crowd sourcing, parallel computing, time series analysis, natural language processing, and visualization (tag cloud, clustergram history flow, spatial information flow), etc.

Big data analysis mainly includes extraction, cleaning, integration, analysis and modeling, and interpretation, etc. (George et al., 2014; Zhang, 2013). Data cleansing is the process of eliminating noisy, erroneous and inconsistent data. Error detection is done by using statistical, clustering, and pattern-based and association rules (Rajesh, 2013). The research of big data in the form of stream data includes classification mining, clustering mining, frequent item mining and other traditional stream data mining issues (Wu, 2014).

There are various ways to analyze data (Rajesh, 2013):

**OLAP:** It is an acronym for On-Line Analytical Processing. It comprises software tools which enable users to do ad hoc querying and analysis on multidimensional data.

**Data Mining:** It is the process of discovering meaningful new correlations, patterns and hidden trends from large volumes of data.

**Dashboards:** They are real-time visualization tools that help in communicating or monitoring critical business indicators.

**Analytics:** Analytics comprises a range of techniques such as gathering, analyzing and interpreting the wealth of data.

Distributed file systems, cluster file systems, and parallel file systems are main tools for Big Data. Computational methods for analyzing big data include cloud computing, cluster computing, and graphics processing unit (GPU) computing, etc. (Merelli, 2014).

**Cloud computing:** It improves Big Data storage and analysis.

**Cluster computing:** The data parallel approach and the parallelization paradigm that subdivide the data to analyze among almost independent processes are suitable solutions for many kinds of Big Data analysis.

**GPU computing:** HPC technologies are the forefront of accelerated data analysis revolutions. The use of accelerator devices such as GPUs represents a cost effective solution.

A client-server architecture for Big Data was presented. The server level architecture for Big Data consists of parallel computing platforms that can handle the associated volume and speed. There are three prominent parallel computing options: clusters or grids, massively parallel processing (MPP), and high performance computing (HPC). The client level architecture consists of NoSQL databases, distributed file systems and a distributed processing framework. Apache Hadoop is a framework for distributed processing of large data sets across clusters of computers, and is designed to scale up from a few servers to thousands of machines, each offering local computation and storage. The two critical components for Hadoop are the Hadoop distributed file system (HDFS) and MapReduce. MapReduce is the distributed processing framework for parallel processing of large data sets. It distributes computing jobs to each server in the cluster and collects the results (Chan, 2013).

MapReduce is a programming model for processing big data. Hadoop is the open-source implementation of MapReduce. It can be used in text analysis, web indexing, and graph processing, etc. (Rajesh, 2013). Modular and decentralized approaches seem to be a cost effective alternative. Hadoop, for example, is prominent open-source top-level Apache data-mining warehouse. It is built on the top of a distributed clustered file system that can take the data from thousands of distributed PC and server hard disks (Hilbert, 2013).

In order to decrease the MapReduce processing time, pre-processing the input data is often needed. In the
context of stream data, such as RFID and sensor data, the data itself is raw data obtained directly from RFID reader or sensor hub. The common types of errors in RFID or sensor data are duplicate reads. Redundancy can happen at the reader level or the data level. Reader level redundancy occurred when there is more than one reader or sensor hub used to cover a specific location. Data level redundancy occurred as data streams. It happens when an RFID tag or a sensor tag stays at the same place for a period of time. Duplicate stream data removal technique considers location filtering. Three algorithms were proposed. They are: Intersection Algorithm, Relative Complement Algorithm, and Randomization Algorithm (Jeon, 2013).

Support vector machine (SVM) is an effective pattern classification method, but it cannot solve pattern classification problems efficiently in big data. Core vector machine (CVM) is a new pattern classification method. It transforms kernel method into an equivalent minimum enclosing ball (MEB) problem and is combined with computational geometry, thus CVM solves the problem of time and space requiring in processing big data. CVM can not only ensure the classification accuracy but also economize running time and storage space (Dong, 2014).

Big data visualization has two major challenges: perceptual and interactive scalability. Visualizing every data point may overwhelm users' perceptual capacities; reducing the data through sampling or filtering can elide interesting structures or outliers. Querying large data can incur disrupting fluent interaction. Even with data reduction methods like binned aggregation, high dimensionality or fine-grained bins cannot ensure real-time processing. For perceptual scalability, binned aggregation was selected as a primary data reduction strategy. Methods for real-time interactive querying (e.g., brushing and linking) among binned plots were proposed. These were implemented in imMens, a browser-based visual analysis system that uses WebGL for data processing and rendering on the graphics processing unit (GPU) (Liu et al., 2013).

Big Data visualization analyst tools must meet the following requirements (Gorodov and Gubarev, 2013):

i. Analyst being able to use more than one data representation view at once.
ii. Active interaction between user and analyzable view.
iii. Dynamical change of factors number during working process with view.

Some Big Data visualization methods are as follows (Gorodov and Gubarev, 2013):

TreeMap: This method is based on space-filling visualization of hierarchical data.

Circle Packing: This method is a direct alternative to Treemap besides the fact that as primitive shape it uses circles, which also can be included into circles from a higher hierarchy level.

Sunburst: It is also an alternative to Treemap, but it uses Treemap visualization, converted to polar coordinate system. The variable parameters are not width and height, but a radius and arc length. This method can be adapted to show data dynamics, using animation.

Circular network diagram: Data object are placed around a circle and linked by curves based on the rate of their relatedness.

Parallel coordinates: This method allows visual analysis to be extended with multiple data factors for different objects.

Streamgraph: Streamgraph is a type of a stacked areagraph, which is displaced around a central axis, resulting inflowing and organic shape.

The classification and properties of the above visualization methods are shown in Table 2.

CHALLENGES OF BIG DATA

Large volume, variety, and velocity of big data create challenges in storage, curation, search, retrieval, and visualization; veracity generates data uncertainty handling complications (Zhang, 2013). Generally, Big Data has following challenges (Jagadish et al., 2014; Hofman and Rajagopal, 2014; Hilbert, 2013):

Data acquisition: The raw data is often too voluminous to store it all. Effective in-situ processing has to be designed.

Information extraction and cleaning: Most data sources are notoriously unreliable. Understanding and modeling the sources of error is a first step toward developing data cleaning techniques.

Data integration and representation: Data must be carefully structured as a first step prior to data analysis. A set of data integration tools helps resolve heterogeneities in data structure and semantics. However, the cost of full integration is often formidable.

Inconsistency and incompleteness: Even after error correction has been applied, some incompleteness and some errors in data are likely to remain.

Timeliness: As data grow in volume, real-time techniques are needed to summarize and filter what is to be stored.

Privacy and data ownership: As the value of data is increasingly recognized, the value of the data owned by
Table 2. Visualization methods classification and properties

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Big Data Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treemap</td>
<td>Large Volume (only hierarchical data)</td>
</tr>
<tr>
<td>Circle packing</td>
<td>Large Volume (only hierarchical data)</td>
</tr>
<tr>
<td>Sunburst</td>
<td>Large Volume + Velocity (data dynamics)</td>
</tr>
<tr>
<td>Circular network diagram</td>
<td>Large Volume + Variety</td>
</tr>
<tr>
<td>Parallel coordinates</td>
<td>Large Volume + Variety + Velocity (data dynamics)</td>
</tr>
<tr>
<td>Streamgraph</td>
<td>Large Volume + Velocity (data dynamics)</td>
</tr>
</tbody>
</table>

Source: Gorodov and Gubarev (2013).

an organization becomes a central consideration.

Data governance and security: The ability to distinguish between open data, data with access restrictions, and data that are not available outside a data source.

Visualization: Perceptual and interactive scalability.

Interoperability of isolated data silos: Large parts of valuable data lurk in “data silos” of different departments, regional offices, and specialized agencies. Fragmentation impedes the massive and timely exploitation of data. Data interoperability standards are becoming a pressing issue for the Big Data paradigm.

Specifically, storage system challenges (Yiu, 2013) in the Big Data era is as follows:

i. Data overload puts pressure on storage system performance.
ii. Increased server virtualization heavily impacts storage.
iii. Disaster recovery concepts are now a must-have. Storage managers need to develop comprehensive back-up plans.
iv. Systems need to be able to multi-task.
v. Cloud infrastructures impact storage systems.

Specifically, emerging directions and future challenges for semantics-based computing on big data are as follows (Jeong and Ghani, 2014):

Big dirty data (BDD): This data is ambiguous, inconsistent, inaccurate, incomplete, and redundant.

Integrated searching in big data: A tool such as ontology seems like a good starting point for this kind of 360-degree semantic search.

Transferring IT solutions: It is from reactive to proactive. The processing of data stored through past events is referred to as the reactive semantic approach. Analyzing the existing large volume of data predict future events is called the proactive semantic approach.

High-speed (streaming) data capturing and consumption.

Security issues in Big Data.

Challenges in information security are our major concerns. Big data can be misused through abuse of privilege by those with access to the data and analysis tools; curiosity may lead to unauthorized access and information may be deliberately leaked. There are three major risk areas: information life cycle, data provenance, and technology unknowns. For information life cycle, there may be no obvious owner for the data to ensure its security. What will be discovered by analysis may not be known at the beginning. The ownership of the data may be subject to dispute. Big data can be secured in transit preferably using encryption (Small, 2013). Information security helps protect privacy. Examples of privacy challenges (Nunan and Domenico, 2013) are listed as follows:

i. Different sets of data that were not previously considered as privacy implications are combined in ways that threaten privacy.
ii. Security problems such as hacking or other unauthorized access.
iii. Data are collected autonomously and independent of human activity. This raises ethical concerns.
iv. Combinations of data for which there are no capabilities to analyze could suffer privacy breaches in the future.

CONCLUSIONS AND FUTURE RESEARCH

Big data is large in volume, variety, velocity, value, and veracity. Data value is closely related to the data volume and variety (multiple data forms, e.g. structured data, unstructured data, and semi-structured data). There is much noise even false information in big data. However, there are hidden high-valued data mixed with raw and noise data in big data.

There are a lot of technologies in Big Data such as data capture, stream processing, parallel computing,
archiving, database management, data mining, and dashboards, etc. Big Data technologies can unlock significant value by making information transparent and usable based on Big Data analytics.

Big Data has challenges such as consolidating segmented or siloed data, processing unstructured or semi-structured data, processing continuously streaming data, data privacy and information security, and lack of unified standards, etc. Mining high-speed data streams and sensor data, privacy-preserving data mining, real-time processing complicated data (with heterogeneity, inconsistency and incompleteness), and big data visualization, etc. can be future research topics.

REFERENCES


