A Survey on Challenges and Advantages in Big Data

LENKA VENKATA SATYANARAYANA

Dept. of CSE, Aditya Institute of Technology and Management, Tekkali, AP, India

Abstract

Big data, which refers to the data sets that are too big to be handled using the existing database management tools, are emerging in many important applications, such as Internet search, business informatics, social networks, social media, genomics, and meteorology. Big data presents a grand challenge for database and data analytics research. The exciting activities addressing the big data challenge. The central theme is to connect big data with people in various ways. Particularly, This paper will showcase our recent progress in user preference understanding, context-aware, on-demand data mining using crowd intelligence, summarization and explorative analysis of large data sets, and privacy preserving data sharing and analysis.The primary purpose of this paper is to provide an in-depth analysis of different platforms available for performing big data analytics. This paper surveys different hardware platforms available for big data analytics and assesses the advantages and drawbacks of Big Data

Keywords

Data Volume, Data Velocity, Data Variety, Data Value, Data Management Issues, Challenges in Big Data, Big Data Technologies

I. Introduction

Big data is defined as large amount of data which requires new technologies and architectures to make possible to extract value from it by capturing and analysis process.New sources of big data include location specific data arising from traffic management, and from the tracking of personal devices such as Smartphones. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are need to be understood. Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools. The difficulties can be related to data capture, storage, search, sharing, analytics and visualization etc.

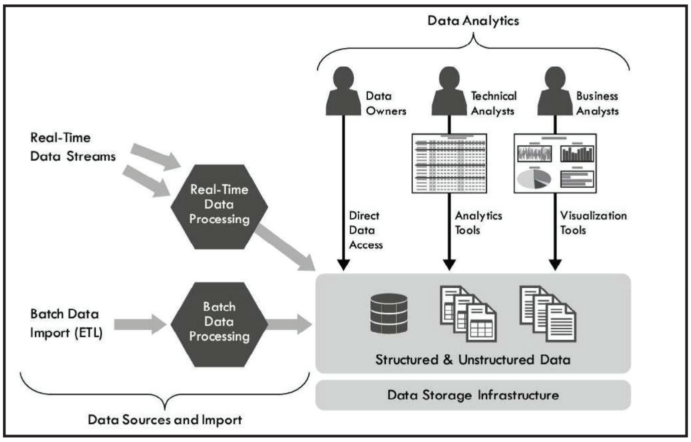

Fig. 1: Big Data Architecture

Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. The various challenges faced in large data management include – scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. In addition to variations in the amount of data stored in different sectors, the types of data generated and stored—i.e., encoded video, images, audio, or text/numeric information; also differ markedly from industry to industry

II. Big Data Characteristics

A. Data Volume

The Big word in Big data itself defines the volume. At present the data existing is in petabytes (1015) and is supposed to increase to zettabytes (1021) in nearby future. Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it.

B. Data Velocity

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows and aggregated.

C. Data Variety

Data variety is a measure of the richness of the data representation – text,images video, audio, etc. Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents.

D. Data Value

Data value measures the usefulness of data in making decisions. Data science is exploratory and useful in getting to know the data, but “analytic science” encompasses the predictive power of big data. User can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and can also rank it according to the dimensions they require. These reports help these people to find the business trends according to which they can change their strategies.

E. Complexity

Complexity measures the degree of interconnectedness (possibly very large) and interdependence in big data structures such that a small change (or combination of small changes) in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all

III. Issues in Big Data

The issues in Big Data are some of the conceptual points that should be understood by the organization to implement the technology effectively. Big data Issues are need not be confused with problems but they are important to know and crucial to handle.

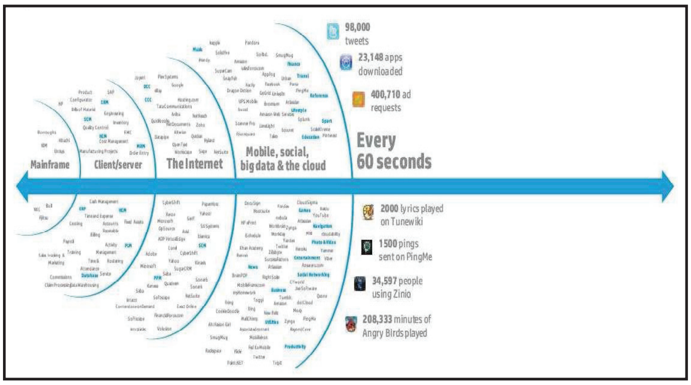
98,000 tweets
23,148 apps downloaded
400,710 ad requests
Every 60 seconds
Mainframe
Client/server
The Internet
Mobile, social, big data & the cloud
2000 lyrics played on Tunewiki
1500 pings sent on PingMe
34,597 people using Zinio
208,333 minutes of Angry Birds played

Fig. 2: Explosion in size of Data

A. Issues related to the Characteristics

Data Volume As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.

Data Velocity Our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion. E-Commerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data velocity management is much more than a bandwidth issue.

Data Variety All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.

Data Value As the data stored by different organizations is being used by them for data analytics. It will produce a kind of gap in between the Business leaders and the IT professionals the main concern of business leaders would be to just adding value to their business and getting more and more profit unlike the IT leaders who would have to concern with the technicalities of the storage and processing.

Data Complexity One current difficulty of big data is working with it using relational databases and desktop statistics/visualization packages, requiring massively parallel software running on tens, hundreds, or even thousands of servers. It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

B. Storage and Transport Issues

The quantity of data has exploded each time we have invented a new storage medium. The difference about the most recent data explosion, mainly due to social media, is that there has been no new storage medium. Moreover, data is being created by everyone and everything, (from Mobile Devices to Super Computers) not just, as here to fore, by professionals such as scientist, journalists, writers etc.

Current disk technology limits are about 4 terabytes (1012) per disk. So, 1 Exabyte (1018) would require 25,000 disks. Even if an Exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an Exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer time to transmit the data from a collection or storage point to a processing point than the time required to actually process it. To handle this issue, the data should be processed "in place" and transmit only the resulting information. In other words, "bring the code to the data", unlike the traditional method of "bring the data to the code." .

C. Data Management Issues

Data Management will, perhaps, be the most difficult problem to address with big data. Resolving issues of access, utilization, updating, governance, and reference (in publications) have proven to be major stumbling blocks. The sources of the data are varied - by size, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them like- documents, drawings, pictures, sound and video recordings, models, software behaviors, user interface designs, etc., with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance.Unlike the collection of data by manual methods, where rigorous protocols are often followed in order to ensure accuracy and validity, Digital data collection is much more relaxed. Given the volume, it is impractical to validate every data item. New approaches to data qualification and validation are needed. The richness of digital data representation prohibits a personalized methodology for data collection. To summarize, there is no perfect big data management solution yet. This represents an important gap in the research literature on big data that needs to be filled.

D. Processing Issues

Assume that an Exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 Exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for endto- end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of Exabyte of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable .

IV. Challenges in Big Data

The challenges in Big Data are usually the real implementation hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant results.

A. Privacy and Security

It is the most important challenges with Big data which is sensitive and includes conceptual,technical as well as legal significance.

1. The personal information (e.g. in database of a merchant or social networking website) of a person when combined with external large data sets, leads to the inference of new facts about that person and it's possible that these kinds of facts about the person are secretive and the person might not want the data owner to know or any person to know about them.
2. Information regarding the people is collected and used in

order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.

3. Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.
4. Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

B. Data Access and Sharing of Information

If the data in the companies information systems is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner. This makes the data management and governance process bit complex adding the necessity to make data open and make it available to government agencies in standardized manner with standardized APIs, metadata and formats thus leading to better decision making, business intelligence and productivity improvements.Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness.

C. Analytical Challenges

The main challenging questions are as:

1. What if data volume gets so large and varied and it is not known how to deal with it?
2. Does all data need to be stored?
3. Does all data need to be analyzed?
4. How to find out which data points are really important?
5. How can the data be used to best advantage?

Big data brings along with it some huge analytical challenges. The type of analysis to be done on this huge amount of data which can be unstructured, semi structured or structured requires a large number of advance skills. Moreover the type of analysis which is needed to be done on the data depends highly on the results to be obtained i.e. decision making. This can be done by using one of two techniques: either incorporate massive data volumes in analysis or determine upfront which Big data is relevant.

D. Human Resources and Manpower

Since Big data is at its youth and an emerging technology so it needs to attract organizations and youth with diverse new skill sets. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive and creative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations. Moreover the Universities need to introduce curriculum on Big data to produce skilled employees in this expertise.

E. Technical Challenges

1. Fault Tolerance

With the incoming of new technologies like Cloud computing and Big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault-tolerant computing is extremely hard, involving intricate algorithms. It is simply not possible to devise absolutely foolproof, 100% reliable fault tolerant machines or software. Thus the main task is to reduce the probability of failure to an "acceptable" level. Unfortunately, the more we strive to reduce this probability, the higher the cost.

Two methods which seem to increase the fault tolerance in Big data are as:

- First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation.
- Second is, one node is assigned the work of observing that these nodes are working properly.

If something happens that particular task is restarted.

But sometimes it's quite possible that that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the output of the previous computation of task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

2. Scalability

The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs,even complex machine learning tasks. There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having the same performance between sequential and random data transfer. Thus, what kinds of storage devices are to be used; is again a big question for data storage.

3. Quality of Data

Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Business Leaders will always want more and more data storage whereas the IT Leaders will take all technical aspects in mind before storing all the data. Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

4. Heterogeneous Data

Unstructured data represents almost every kind of data being produced like social media interactions, to recorded meetings, to handling of PDF documents, fax transfers, to emails and more. Working with unstructured data is cumbersome and of course costly too. Converting all this unstructured data into structured one is also not feasible. Structured data is always organized into highly mechanized and manageable way. It shows well integration with database but unstructured data is completely raw and unorganized.

V. Big Data Technologies

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. Technologies being applied to big data include Massively Parallel Processing (MPP) databases, data mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems. Real or near-real time information delivery is one of the defining characteristics of Big Data Analytics. A wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyze, and visualize big data. These techniques and technologies draw from several fields including statistics, computer science, applied mathematics, and economics. This means that an organization that intends to derive value from big data has to adopt a flexible, multidisciplinary approach

A. Big Data Technologies

The International Data Corporation (IDC) study predicts that overall data will grow by 50 times by 2020, driven in large part by more embedded systems such as sensors in clothing, medical devices and structures like buildings and bridges. The study also determined that unstructured information - such as files, email and video - will account for 90% of all data created over the next decade. But the number of IT professionals available to manage all that data will only grow by 1.5 times today's levels (World's data will grow by 50X in next decade, IDC study predicts , 2011). The digital universe is 1.8 trillion gigabytes in size and stored in 500 quadrillion files. And its size gets more than double in every two years time frame. If we compare the digital universe with our physical universe then it's nearly as many bits of information in the digital universe as stars in our physical universe

Given a very large heterogeneous data set, a major challenge is to figure out what data one has and how to analyze it. Unlike the previous two sections, hybrid data sets combined from many other data sets hold more surprises than immediate answers. To analyze these data will require adapting and integrating multiple analytic techniques to "see around corners", e.g., to realize that new knowledge is likely to emerge in a non-linear way. It is not clear that statistical analysis methods, as Ayres argues, are or can be the whole answer. A Big Data platform should give a solution which is designed specifically with the needs of the enterprise in the mind. The following are the basic features of a Big Data Platform
offering

Comprehensive - It should offer a broad platform and address all three dimensions of the Big Data challenge -Volume, Variety and Velocity.

- Enterprise-ready - It should include the performance, security, usability and reliability features.
- Integrated - It should simplify and accelerates the introduction of Big Data technology to enterprise. It should enable integration with information supply chain including databases, data warehouses and business intelligence applications.
- Open source based - It should be open source technology with the enterprise-class functionality and integration.
- Low latency reads and updates
- Robust and fault-tolerant
- Scalability
- Extensible
- Allows adhoc queries
- Minimal maintenance

Table 1: Tools Available for Big Data

Tool	Tool Description
IBM InfoSphereBigInsights	It is open source Apache Hadoop with IBM Big Sheet which is able to analyze data in its native format without imposing any schema/structure and enables fast adhoc analysis.
WX2 Kognitio Analytical Platform	It is a fast and scalable in-memory analytic database platform
ParAccel Analytic Platform	It is a columnar, massively parallel processing (MPP) analytic database platform. It has strong features for the query optimization and compilation, compression and network interconnect.
SAND Analytic Platform	It is a columnar analytic database platform that achieves linear data scalability through massively parallel processing (MPP).

VI. Big Data Advantages

The Big Data has numerous advantages on society, science and technology. It is unto the way that how it is used for the human beings. Some of the advantages (Marr, 2013)are described below:

A. Understanding and Targeting Customers

This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models.

B. Understanding and Optimizing Business Process

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. HR business processes are also being improved using big data analytics. This includes the optimization of talent acquisition.

C. Improving Science and Research

Science and research is currently being transformed by the new possibilities big data brings. Take, for example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of our universe – how it started and works - generate huge amounts of data. The CERN datacentre has 65,000 processors to analyse its 30 petabytes of data.

D. Improving Healthcare and Public Health

The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. Just think of what happens when all the individual data from smart watches and wearable devices can be used to apply it to millions of people and their various diseases. The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone.

E. Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings. Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses

F. Improving Security and Law Enforcement

Big data is applied heavily in improving security and enabling law enforcement. The revelations are that the National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us). Others use big data techniques to detect and prevent cyber-attacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data use it to detect fraudulent transactions.

VII. Conclusion

Here some of the important issues are covered that are needed to be analyzed by the organization while estimating the significance of implementing the Big Data technology and some direct challenges to the infrastructure of the technology. The commercial impacts of the Big data have the potential to generate significant productivity growth for a number of vertical sectors. They should also create healthy demand for talented individuals who are capable to help organizations in making sense of this growing volume of raw data. In short, Big Data presents opportunity to create unprecedented business advantages and better service delivery. A regulatory framework for big data is essential. That framework must be constructed with a clear understanding of the ravages that have been wrought on personal interests by the reduction of information to data, its centralization, and its expropriation but the biggest gap is the lack of the skilled managers to make decisions based on analysis by a factor of 10x. Growing talent and building teams to make analytic-based decisions is the key to realize the value of Big Data.

References

[1] Hansen, C.,“Big Data: A Scientific Visualization Perspective”, SCI Institute Professor of Computer Science, University of Utah, 2013.

[2] Emmanuel Letouzé (May 2012), “Big Data for Development: Challenges & Opportunities”, [Online] Available: http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-GlobalPulseMay2012.pdf.

[3] Aveksa Inc.,"Ensuring “Big Data”, Security with Identity and Access Management", Waltham,MA: Aveksa, 2013.

[4] Hewlett-Packard Development Company,"Big Security for Big Data. L.P.: Hewlett-Packard Development Company", 2012.

[5] Kaisler, S., Armour, F., Espinosa, J. A., Money, W.,"Big Data: Issues and Challenges Moving Forward", International Confrence on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Soceity, 2013.

[6] Katal, A., Wazid, M., Goudar, R. H.,"Big Data: Issues, Challenges, Tools and Good Practices", IEEE, pp. 404-409, 2013.

[7] Marr, B. (2013, November 13),"The Awesome Ways Big Data is used Today to Change Our World", Retrieved November 14, 2013, from LinkedIn: https://www.linkedin.com/today/post/article/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-tochange-our-world

[8] Patel, A. B., Birla, M., Nair, U.,"Addressing Big Data Problem Using Hadoop and Nirma University", Gujrat: Nirma University, 2013.

[9] Singh, S., Singh, N.,"Big Data Analytics", International Conference on Communication, Information & Computing Technology (ICCICT) (pp. 1-4). Mumbai: IEEE, 2012.

[10] The 2011 Digital Universe Study: Extracting Value from Chaos. (2011, November 30). Retrieved from EMC: http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

[11] World's data will grow by 50X in next decade, IDC study predicts. (2011, June 28). Retrieved from Computer World: http://www.computerworld.com/s/article/9217988/World_s_data_will_grow_by_50X_in_next_decade_IDC_study_predicts

[12] Dumbill E: What Is Big Data? An Introduction to the Big Data Landscape. In Strata 2012: Making Data Work. O'Reilly, Santa Clara, CA O'Reilly; 2012.