Our Heritage

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

Data Intensive Applications and Challenges : A Survey on Map Reduce

R.N.Panda
Research Scholar ,Dept. of ICT
Fakir Mohan University,
Balasore, India
rnpanda2007@gmail.com

Dr Sabyasachi Pattnaik
Professor, Dept. of ICT,
Fakir Mohan University
Balasore, India
spattnaik40@yahoo.co.in

Abstract

It is already true that Big Data has drawn huge attention from researchers in information sciences, policy and decision makers in governments and enterprises. As the speed of information growth exceeds Moore's Law at the beginning of this new century, excessive data is making great troubles to human beings. However, there are so much potential and highly useful values hidden in the huge volume of data. A new scientific paradigm is born as dataintensive scientific discovery (DISD), also known as Big Data problems. A large number of fields and sectors, ranging from economic and business activities to public administration, from national security to scientific researches in many areas, involve with Big Data problems. On the one hand, Big Data is extremely valuable to produce productivity in businesses and evolutionary breakthroughs in scientific disciplines, which give us a lot of opportunities to make great progresses in many fields. There is no doubt that the future competitions in business productivity and technologies will surely converge into the Big Data explorations. On the other hand, Big Data also arises with many challenges, such as difficulties in data capture, data storage, data analysis and data visualization. This paper is aimed to demonstrate a close-up view about Big Data, including Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies we currently adopt to deal with the Big Data problems. We also discuss several underlying methodologies to handle the data deluge, for example, granular computing, cloud computing, bio-inspired computing, and quantum computing.

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

Keywords:*Interest Locality; Data Groupings; CloudBLAST,Cloudburst*

Introduction

Big Data has been one of the current and future research frontiers. Gartner defined big data as "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". He listed big data in "Top 10 Strategic Technology Trends For 2013" and "Top 10 Critical Tech Trends For The Next Five Years" .Big Data is a collection of very huge data sets with a great diver-sity of types made it difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms. In 2012, Gartner gave a more detailed definition as: "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". More generally, a data set canbe called Big Data if it is formidable to perform capture, curation, analysis and visualization on it at the current technologies. With diversified data provisions, such as sensor networks, telescopes, scientific experiments, and high throughput instruments, the datasets increase at exponential rate . The off-the-shelf techniques and technologies that we used to store and analyse data cannot work efficiently and satisfactorily. The challenges arise from data capture and data curation to data analysis and data visualization. In many instances, science is lagging behind the real world in the capabilities of discovering the valuable knowledge from massive volume of data. Based on precious knowledge, we need to develop and create new techniques and technologies to excavate Big Data and benefit our specified purposes. Big Data has changed the way that we adopt in doing businesses, managements and researches. Data-intensive science especially in data-intensive computing is coming into the world that aims to provide the tools that we need to handle the Big Data problems. Data-intensive science is emerging as the fourth scientific paradigm in terms of the previous three, namely empirical science, theoretical science and computational science. Thousand years ago, scientists describing the natural phenomenon was only based on human empirical evidences, so we call the science at that time as empirical science. It is also the beginning of science and classified as the first paradigm. Then, theoretical science emerged hundreds years ago as the second paradigm, such as Newtons Motion Laws and Keplers Laws. However, in terms of many complex phenomenon and problems, scientists have to turn to scientific simulations, since theoretical analysis is highly complicated and sometimes unavailable and infeasible. Afterwards, the third science paradigm was born as computational branch. Simulations in large of fields generate a huge volume of data from the experimental science, at the same time, more and more large data sets are generated in many pipelines. There is no doubt that the world of science has changed just because of the increasing data-intensive

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

applications. The techniques and technologies for this kind of data-intensive science are totally distinct with the previous three.

Big Data in commerce and business

According to estimates, the volume of business data worldwide, across almost companies, doubles every 1.2 year. Taking retail industry as an example, we try to give a brief demonstration for the functionalities of Big Data in commercial activities. There are around 267 million transactions per day in Wal-Mart's 6000 stores worldwide. For seeking for higher competitiveness in retail, Wal-Mart recently collaborated with Hewlett Packard to establish a data warehouse which has a capability to store 4 petabytes (see the size of data unit in Appendix A) of data, i.e., 4000 trillion bytes, tracing every pur-chase record from their point-of-sale terminals. Taking advantage of sophisticated machine learning techniques to exploit the knowledge hidden in this huge volume of data, they successfully improve efficiency of their pricing strategies and adver-tising campaigns. The management of their inventory and supply chains also significantly benefits from the large-scale warehouse.

In the era of information, almost every big company encounters Big Data problems, especially for multinational corpora-tions. On the one hand, those companies mostly have a large number of customers around the world. On the other hand, there are very large volume and velocity of their transaction data. For instance, FICO's falcon credit card fraud detection sys-tem manages over 2.1 billion valid accounts around the world. There are above 3 billion pieces of content generated on Face-book every day. The same problem happens in every Internet companies. The list could go on and on, as we witness the future businesses battle fields focusing on Big Data.

Big Data in society administration

Public administration also involves Big Data problem. On one side, the population of one country usually is very large. For another, people in each age level need different public services. For examples, kids and teenagers need more edu-cation, the elders require higher level of health care. Every person in one society generates a lot of data in each public section, so the total number of data about public administration in one nation is extremely huge. For instance, there are almost 3 terabytes of data collected by the US Library of Congress by 2011. The Obama administration announced the Big Data re-search and development initiative in 2012, which investigate addressing important problems facing the government by make use of Big Data. The initiative was constitutive of 84 different Big Data programs involving six departments. Thesimilar thing also happened in Europe. Governments around the world are facing adverse conditions to improve their productivity. Namely, they are required to be more effective in public administration. Particularly in the recent global recession, many governments have to provide a higher level of public services with significant budgetary constraints. Therefore, they should take

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

Big Data as a potential budget resource and develop tools to get alternative solutions to decrease big budget deficits and reduce national debt levels.

According to McKinsey's report, Big Data functionalities, such as reserving informative patterns and knowledge, provide the public sector a chance to improve productivity and higher levels of efficiency and effectiveness. European's pub-lic sector could potentially reduce expenditure of administrative activities by 15–20 percent, increasing 223 billion to 446 billion values, or even more.

Big Data in scientific research

Many scientific fields have already become highly data-driven with the development of computer sciences. For instance, astronomy, meteorology, social computing, bioinformatics and computational biology are greatly based on data-intensive scientific discovery as large volume of data with various types generated or produced in these sci-ence fields.. How to probe knowledge from the data produced by large-scale scientific simulation? It is a certain Big Data problem which the answer is still unsatisfiable or unknown. For instances, a sophisticated telescope is regarded as a very large digital camera which generate huge number of universal images. For example, the Large Synoptic Survey Telescope (LSST) will record 30 trillion bytes of image data in a single day. The size of the data equals to two entire Sloan Digital Sky Surveys daily. Astronomers will utilize computing facilities and advanced analysis methods to this data to investigate the origins of the universe. The Large Hadron Collider (LHC) is a particle accelerator that can generate 60 terabytes of data per day. The patterns in those data can give us an unprecedented understanding the nature of the universe. 32 petabytes of climate observations and simulations were con-served on the discovery supercomputing cluster in the NASA Center for Climate Simulation (NCCS). The volume of human genome information is also so large that decoding them originally took a decade to process. Otherwise, a lot of other e-Science projects are proposed or underway in a wide variety of other research fields, range from environmental science, oceanography and geology to biology and sociology. One common point exists in these disciplines is that they generate enormous data sets that automated analysis is highly required. Additionally, centralized repository is necessary as it is impractical to replicate copies for remote individual research groups. Therefore, centralized storage and analysis approaches drive the whole system designs.

Big Data opportunities and challenges

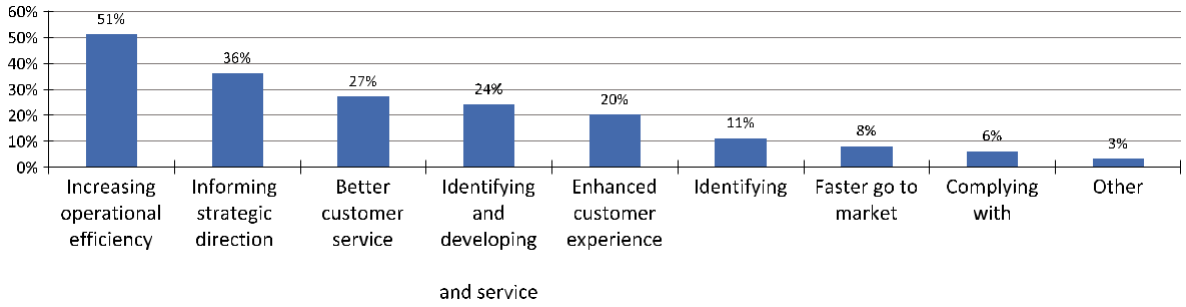
Opportunities

Recently, several US government agencies, such as the National Institutes of Health (NIH) and the National Science Foundation (NSF), ascertain that the utilities of Big Data to data-intensive

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

decision-making have profound influences in their future developments. Consequently, they are trying to developing Big Data technologies and techniques to facilitate their missions after US government passed a large-scale Big Data initiative. This initiative is very helpful for building new capabilities for exploiting informative knowledge and facilitate decision-makers. From the Networking Information Technology Research and Development (NITRD) program which is recently recognized by President's Council of Advisors on Science and Technology (PCAST), we know that the bridges between Big Data and knowledge hidden in it are highly crucial in all areas of national priority. This initiative will also lay the groundwork for complementary Big Data activities, such as Big Data infrastructure projects, platforms development, and techniques in settling complex, data-driven problems in sciences and engineering. Finally, they will be put into practice and benefit society. According to the report from McKinsey institute, the effective use of Big Data has the underlying benefits to transform economies, and delivering a new wave of productive growth. Taking advantages of valuable knowledge beyond Big Data will become the basic competition for today's enterprises and will create new competitors who are able to attract employees that have the critical skills on Big Data. Researchers, policy and decision makers have to recognize the potential of harnessing Big Data to uncover the next wave of growth in their fields. There are many advantages in business section that can be obtained through harnessing Big Data as illustrated in Fig. including increasing operational efficiency, informing strategic direction, developing better customer service, identifying and developing new products and services,



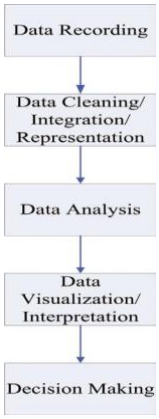
Challenges

Opportunities are always followed by challenges. On the one hand, Big Data bring many attractive opportunities. On the other hand, we are also facing a lot of challenges when handle Big Data problems, difficulties lie in data capture, storage, searching, sharing, analysis, and visualization. If we cannot surmount those challenges, Big Data will become a gold ore but we do not have the capabilities to explore it, especially when information surpass our capability to harness. One challenge is existing in computer architecture for several decades, that is, CPU-

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

heavy but I/O-poor. This system imbalance still restraint the development of the discovery from Big Data.The CPU performance is doubling each 18 months following the Moore's Law, and the performance of disk drives is also doubling at the same rate. However, the disks' rotational speed has slightly improved over the last decade. The conse-quence of this imbalance is that random I/O speeds have improved moderately while sequential I/O speeds increase with density slowly. Moreover, information is increasing at exponential rate simultaneously, but the improvement of informa-tion processing methods is also relatively slower. In a lot of important Big Data applications, the state-of-the-art tech-niques and technologies cannot ideally solve the real problems, especially for real-time analysis. So partially speaking, until now, we do not have the proper tools to exploit the gold ores completely.Different challenges arise in each sub-process when it comes to data-driven applications. In the following subsections, we will give a brief dis-cussion about challenges we are facing for each sub-process.



Knowledge discovery process.

Data capture and storage

Data sets grow in size because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies, remote sensing, software logs, cameras, microphones, radio-frequency identification readers, wireless sensor networks, and so on. There are 2:5 quintillion bytes of data created every day, and this number keeps increas-ing exponentially. The world's technological capacity to store information has roughly doubled about every 3 years since the 1980s. In many fields, like financial and medical data often be deleted just because there is no enough space to store these data. These valuable data are created and captured at high cost, but ignored finally. The bulk storage requirements for experimental data bases, array storage for large-scale scientific computations, and large output files are

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

reviewed in.Big Data has changed the way we capture and store data, including data storage device, data storage architecture, data access mechanism. As we require more storage mediums and higher I/O speed to meet the challenges, there is no doubt that we need great innovations. Firstly, the accessibility of Big Data is on the top priority of the knowledge discovery process. Big Data should be accessed easily and promptly for further analysis, fully or partially break the restraint: CPU-heavy but I/O-poor. In addition, the under-developing storage technologies, such as solid-state drive (SSD) and phase-change memory (PCM),, may help us alleviate the difficulties, but they are far from enough. One significant shift is also under-way, that is the transformative change of the traditional I/O subsystems. In the past decades, the persistent data were stored by using hard disk drives (HDDs). As we known, HDDs had much slower random I/O performance than sequential I/O performance, and data processing engines formatted their data and designed their query processing methods to work around this limitation. But, HDDs are increasingly being replaced by SSDs today, and other technologies such as PCM are also around the corner. These current storage technologies cannot possess the same high performance for both the sequential and random I/O simultaneously, which requires us to rethink how to design storage subsystems for Big Data processing systems.

Data transmission

Cloud data storage is popularly used as the development of cloud technologies. We know that the network bandwidth capacity is the bottleneck in cloud and distributed systems, especially when the volume of communication is large. On the other side, cloud storage also lead to data security problems the requirements of data integrity checking. Many schemes were proposed under different systems and security models.

Data curation

Data curation is aimed at data discovery and retrieval, data quality assurance, value addition, reuse and preservation over time. This field specifically involves a number of sub-fields including authentication, archiving, management, preservation, retrieval, and representation. The existing database management tools are unable to process Big Data that grow so large and complex. This situation will continue as the benefits of exploiting Big Data allowing researchers to analysebusiness trends, prevent diseases, and combat crime. Though the size of Big Data keeps increasing exponentially, current capability to work with is only in the relatively lower levels of petabytes, exabytes and zettabytes of data. The classical approach of managing structured data includes two parts, one is a schema to storage the data set, and another is a relational database for data re-trieval. For managing large-scale datasets in a structured way, data warehouses and data marts are two popular approaches. A data warehouse is a relational

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

database system that is used to store and analyze data, also report the results to users. The data mart is based on a data warehouse and facilitate the access and analysis of the data warehouse. A data warehouse is mainly responsible to store data that is sourced from the operational systems. The preprocessing of the data is necessary before it is stored, such as data cleaning, transformation and cataloguing. After these preprocessing, the data is available for higher level online data mining functions. The data warehouse and marts are Standard Query Language (SQL) based dat-abases systems.

Conclusion

MapReduce has efficiency and scalability in most of the studies . It is used for generating and processing big Data in various different applications.The purpose of this essay is to review, MapReduce,its architecture, big data and an appropriate use of programming model in conjunction with the applications of MapReduce in big data have been discussed thoroughly. Also, we have surveyed and analyzed the implementations of MapReduce. The applications of MapReduce framework in different contexts like the cloud, multi-core system, and parallel computation have been investigated precisely. This paper examines and categorized a number of applications which have been surveyed in MapReduce Framework based on Graph processing, Join and parallel queries, optimizing frameworks, multi-core systems, and data allocation. The goal of the MapReduce Framework is to provide an abstraction layer between the faulttolerance, data distribution and other parallel systems tasks, and the implementation details of the specific algorithm. Obviously, the requirements of MapReduce applications is growing rapidly. This survey gives the reader a general review of the MapReduceapplicationsand it will be a good introductory reference to improve the article that is easier to comprehend.

References

1. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big-data-fact-sheet-final-1.pdf.
2. http://quantumcomputers.com.
3. Karmasphere Studio and Analyst, 2012. <http://www.karmasphere.com/>.
4. Pentaho Business Analytics, 2012. <http://www.pentaho.com/explore/pentaho-business-analytics/>.
5. Sqlstream, 2012. <http://www.sqlstream.com/products/server/>.
6. Storm, 2012. <http://storm-project.net/>.
7. AbzetdinAdamov. Distributed file system as a basis of data-intensive computing, in: 2012 6th International Conference on Application of Information
8. and Communication Technologies (AICT), pp. 1–3 (October).

ISSN: 0474-9030

Vol-68-Issue-1-January-2020

9. Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, UmeshwasDayal, Michael Franklin, Johannes Gehrke, Laura Haas, Jiawei Han Alon
10. Halevy, H.V. Jagadish, AlexandrosLabrinidis, Sam Madden, YannisPapakonstantinou, Jignesh Patel, Raghu Ramakrishnan, Kenneth Ross, Shahabi
11. Cyrus, Dan Suciu, Shiv Vaithyanathan, Jennifer Widom, Challenges and Opportunities with Big Data, CYBER CENTER TECHNICAL REPORTS, Purdue
12. University, 2011.
13. ByungikAhn, Neuron machine: Parallel and pipelined digital neurocomputing architecture, in: 2012 IEEE International Conference on Computational
14. Intelligence and Cybernetics (CyberneticsCom), 2012, pp. 143–147.
15. James Ahrens, Kristi Brislawn, Ken Martin, BerkGeveci, C. Charles Law, Michael Papka, Large-scale data visualization using parallel data streaming,
16. IEEE Comput. Graph. Appl. 21 (4) (2001) 34–41.
17. Chris Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, 2008. <http://www.wired.com/science/discoveries/
18. magazine/16-07/pb-theory>.
19. Ernesto Andrianantoandro, SubhayuBasu, David K Karig, Ron Weiss, Synthetic biology: new engineering rules for an emerging discipline, Mol. Syst.
20. Biol. 2 (2006).
21. ItamarArel, Derek C. Rose, Thomas P. Karnowski, Deep machine learning – a new frontier in artificial intelligence research, IEEE Comput. Intell. Mag. 5
22. (4) (2010) 13–18.
23. Aditya Auradkar, ChavdarBotev, Shirshanka Das, Dave DeMaagd, Alex Feinberg, PhanindraGanti, Bhaskar Ghosh Lei Gao, Kishore Gopalakrishna,
24. Brendan Harris, Joel Koshy, Kevin Krawez, Jay Kreps, Shi Lu, Sunil Nagaraj, NehaNarkhede, Sasha Pachev, Igor Perisic, Lin Qiao, Tom Quiggle, Jun Rao,
25. Bob Schulman, Abraham Sebastian, Oliver Seeliger, Adam Silberstein, Boris Shkolnik, ChinmaySoman, RoshanSumbaly, KapilSurlaker, Sajid
26. Topiwala, Cuong Tran, BalajiVaradarajan, JemiahWesterman, Zach White, David Zhang, Jason Zhang, Data infrastructure at linkedin, in: 2012 IEEE
27. 28th International Conference on Data Engineering (ICDE), 2012, pp. 1370–1381.
28. ArshdeepBahga, Vijay K. Madisetti, Analyzing massive machine maintenance data in a computing cloud, IEEE Trans Parallel Distrib. Syst. 23 (10)