American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-6, Issue-4, pp-46-51
www.ajer.org
Research Paper Open Access

Big Data Analytics Challenges and Solutions in Cloud

Ms. Manju Sharma
College of Computer Science & Information Systems, Jazan University Jazan, Kingdom of Saudi Arabia

ABSTRACT: *Cloud Computing is a technology getting popularity in the field of sharing of data, hardware and software resources. It's all about sharing of computing resources rather than getting local servers or personal devices. Cloud is a service provided or available through Internet Cloud Computing is a computing in which services are delivered through the Internet. The goal of Cloud Computing is to make use of unite the millions of data and its user into single platform. Cloud computing we are using huge group of servers with specialized connections to distribute data processing among the servers. Big Data is a collection of huge volume of data structured and unstructured data that are so large and difficult to get process using traditional databases and software technologies. In big data massive volume of distributed data can be handled or stored in clouds. Cloud computing is the best solution for storage of massive amount of data.*

Keywords: *Big data, cloud computing, Hadoop, cloud computing environments, RDBMS, delimitation.*

I. INTRODUCTION

Cloud computing, the word Cloud (also phrased as "the cloud") is used as a metaphor for "*the Internet*," so the phrase *cloud computing* means "a type of Internet-based computing," where different services — such as servers, storage and applications — are delivered to an organization's computers and devices through the Internet[18]. Instead of installing a software suite for each computer, this technology requires to install single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user. There's a significant workload shift, in a cloud computing system [10].

The term "Big Data [11]" is believed to be originated from the Web search companies who had to query loosely structured very large distributed data .With large and complex data, computation becomes difficult to be handled by the traditional data processing applications which trigger the development of big data applications [12]. Google's map reduce framework and apache Hadoop are the defacto software systems [13] for big data applications, in which these applications generates a huge amount of intermediate data. Manufacturing and Bioinformatics are the two major areas of big data applications. Another important application for Hadoop is Bioinformatics which covers the next generation sequencing and other biological domains. Bioinformatics [14] which requires a large scale data analysis, uses Hadoop. Cloud computing gets the parallel distributed computing framework together with computer clusters and web interfaces.

In the Journal of Science 2008, "Big Data" is defined as "the representation of the progress of human cognitive processes, which usually includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time". So also , the definition of big data as given by the Gartner defined it as 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" [2]. Big Data can be described as a massive volume of structured and unstructured data which are so large and very difficult to process this data using traditional methods and recent software technologies. [3]. Furthermore, Big data is the elusive, all-encompassing name given to enormous datasets stored on enterprise servers for example, data stored in Google (which organizes 100 trillion Web pages), Facebook (with 1 million gigabytes of disk storage, its data keep increasing on a daily basis), and YouTube (which contains 20 petabytes of new video content per year).

Big data is also used in science, for scientific applications such as weather forecasting, earthquake prediction, seismic processing, molecular modeling, and genetic sequencing. Many of these applications require servers to run with tens of petabytes of storage, such as the Sequoia (Lawrence Livermore) and Blue Waters (NCSA) supercomputers [4]. The three main terms that generally signify Big Data are:
The three main terms that generally signify Big Data are:

i. Volume: This has to do with the amount of data generated on a daily basis which is so large and keeps increasing with time.
ii. Variety: Today data is created in different type, form and formats such as emails, video, audio, transactions etc.
iii. Velocity: This has to do with the speed it takes to produce data and how fast this data produced needs to be processed on time to meet individual demand.

The other two properties that need to be critically consider when talking about Big Data are Variability and Complexity as depicted by [31].

i. Variability: this goes along with velocity, and it has to do with how inconsistent the flow of data can be with respect to time and far it can go.
ii. Complexity: the complexity of the data must be considered especially when we have multiple source of data. The data must be rearranged in such a format that will be suitable for processing[1].

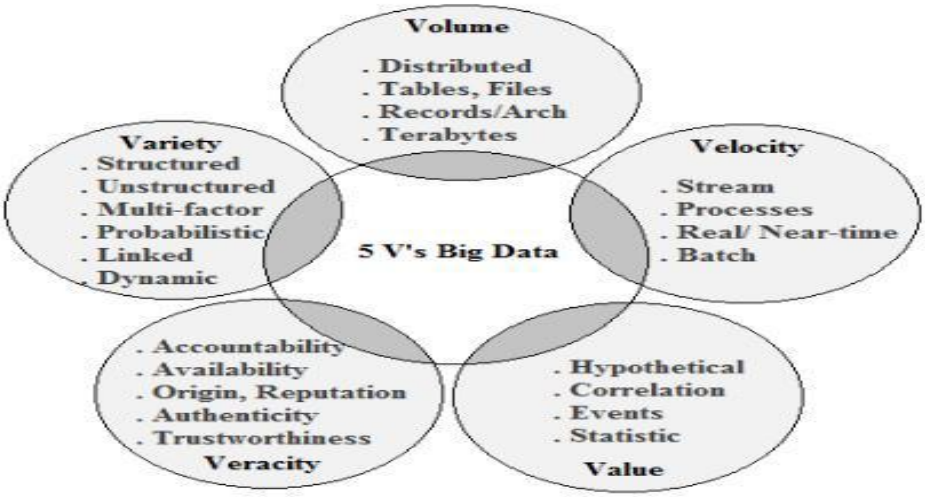

Fig. 1 Properties of Big Data[2]

II. HADOOP

Hadoop(a cloud computing framework) is a Java based distributed system, is a new framework in the market. Since Hadoop is new and still being developed to add more features, there are many security issues which need to be addressed. Researchers have identified some of the issues and started working on this. Some of the notable outcomes, which are related to our domain and helped us to explore, are presented below. The World Wide Web consortium has identified the importance of SPARQL which can be used in diverse data sources. Later on, the idea of secured query was proposed in order to increase privacy in privacy/utility tradeoff. Here, Jelena, of the USC Information Science Institute, has explained that the queries can be processed according to the policy of the provider, rather than all query processing. Bertino et al published a study on access control for XML Documents [15]. In the study, cryptography and digital signature technique are explained, and techniques of access control to XML data document is stressed for secured environment. Later on, he published another study on authentic third party XML document distribution [16] which imposed another trusted layer of security to the paradigm.

Figure 2: Big Data [29]

Hadoop is a tool used by many organizations for managing and analysing the data. Hadoop uses parallel execution of data using large clusters of tiny machines or nodes which results in faster execution. And even data is distributed among the nodes so the node failure can be easily handled. MapReduce is a programming style, for Distributed processing on Hadoop. It contains the two functions; Map function will take the input as key/value pair and splits the data on several nodes for processing. Reduce function combines the results from Map function. The architecture and algorithm are implemented using Hadoop[25]

Kevin Hamlen and et al proposed that data can be stored in a database encrypted rather than plain text. The advantage of storing data encrypted is that even though intruder can get into the database, he or she can't get the actual data. But, the disadvantage is that encryption requires a lot of overhead. Instead of processing the plain text, most of the operation will take place in cryptographic form. Hence the approach of processing in cryptographic form added extra to security layer.IBM researchers also explained that the query processing should take place in a secured environment. Then, the use of Kerberos has been highly effective. Kerberos is nothing but a system of authentication that has been developed at MIT. Kerberos uses an encryption technology along with a trusted third party, an arbitrator, to be able to perform a secure authentication on an open network. To be more specific, Kerberos uses cryptographic tickets to avoid transmitting plain text passwords over the wire. Kerberos is based upon Needham-Schroeder protocol.

Airavat [17] has shown us some significant advancement security in the Map Reduce environment. In the study, Roy and et al have used the access control mechanism along with differential privacy. They have worked upon mathematical bound potential privacy violation which prevents information leak beyond data provider's policy. The above works have influenced us, and we are analyzing various approaches to make the cloud environment more secure for data transfer and computation.

III. CLOUD COMPUTING ENVIRONMENTS

In current scenario, the challenges of security in cloud computing environments can be categorized into network level, user authentication level, data level, and generic issues. These Cloud computing technology comes with numerous security issues and this could be due to the fact that it encompasses many technologies which may include networks, databases, operating systems, virtualization, resource allocation, containerization, resource scheduling, transaction management, load balancing, concurrency control, managing contents distribution in a content delivery network (CDN) and memory management. Hence, security issues of these systems and technologies exist in cloud computing. For example, the security of the network that interconnects the systems in the cloud must be much secured. Also, containerization and virtualization paradigm in cloud computing bring about several security concerns. For example, the mapping of containers and virtual machines to the physical machines has to be done in a secured way. The security issues associated with cloud computing devices and environments can be categorized into the following: network level, user authentication level, data level, and generic issues as depicted by [28], [29].

Network level: The challenges associated with network level will include issues with network protocols and network security, such as distributed nodes, distributed data, Internode communication.

User Authentication level: The issues and challenges associated with user authentication level includes encryption/decryption techniques, authentication methods which may include issues with administrative rights for nodes, authentication of applications and nodes, logging etc.

Data level: The issues and challenges associated with data level will include data integrity and availability issues such as data protection and the distribution of data.

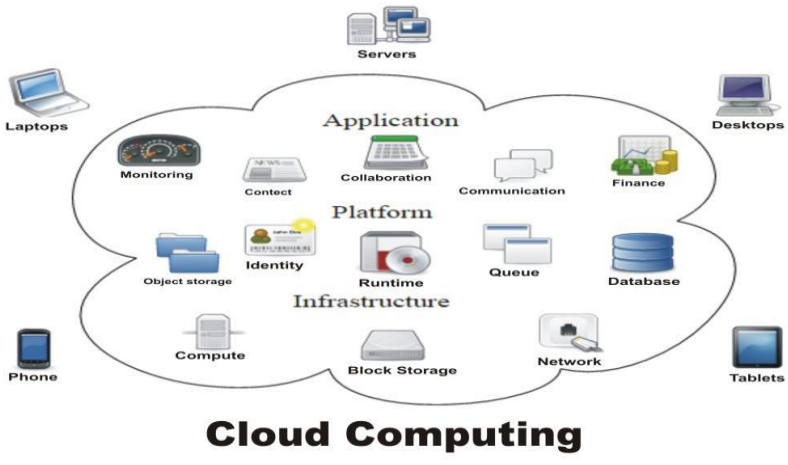

Figure 3: Cloud computing scenario [29] [30]

The issues and challenges associated with general level security issues includes issues with traditional security tools, and use of different technologies On the other hand, for big data security challenges, they are overblown by the three key characteristics of big data which are volume, variety, and velocity. Some of the unique treats that causes security vulnerabilities in big data are: Large-scale cloud infrastructures, diversity of data sources and formats, as well as the streaming nature of data acquisition and high volume inter-cloud migration.[1],[28].Cloud computing has become one of the hottest buzzwords in the IT area. Many companies and institutions are rushing to define clouds and provide cloud solutions in various ways. However, there is still no widely accepted definition for Cloud computing. A cloud is a type of distributed data center which delivers infrastructures as services. It consists of massive resources, and provides some mechanisms to provide, reimage, workload rebalance, de-provide, and monitor those resources. It represents as one or more unified resource entities, and renders users/applications with services to access those resources without knowing the detailed information[32].

IV. TRADITIONAL METHODS

Cloud computing has been revolutionizing the IT industry by adding flexibility to the way IT is consumed, enabling organizations to pay only for the resources and services they use [9]. In an effort to reduce IT capital and operational expenditures, organizations of all sizes are using Clouds to provide the resources required to run their applications. Clouds vary significantly in their specific technologies and implementation, but often provide infrastructure, platform, and software resources as services [5-6].Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage [13]. Analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly avail-able on the Web [10]. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources. This paradigm is being popularly termed as Big Data [12].In today's competitive market, being able to explore data to understand customer behavior, segment customer base, offer customized services, and gain insights from data provided by multiple sources is key to competitive advantage. Although decision makers would like to base their decisions and actions on insights gained from this data [9], making sense of data, extracting non obvious patterns, and using these patterns to predict future behavior are not new topics.

Knowledge Discovery in Data (KDD) [10] aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining [12], more specifically, aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions [7, 8, and 10].

RDBMS: The first and probably most obvious way of dealing with Big Data is by using traditional data warehousing architectures based on standard RDBMS. In this case, data are extracted from various internal and external sources, selected, aggregated, and loaded into a data warehouse. Different business intelligence tools can then be used to analyze and access the data. As volume and velocity of the data to be processed steadily increased since the 1980 [36].most contemporary companies revert to parallelized RDBMS to handle the large amounts of data [37].Consequently, data are stored on multiple machines, tables are partitioned over the nodes in a cluster, andan application layer allows for accessing the different data portions on the different nodes. The goal of such an architecture is to provide linear speed-up as well as scale-up [35][37].

V. DELIMITATIONS

The baseline to evaluate the effectiveness of big data and analytics platform which provides organizations a solution stack that is designed specifically for enterprise use. The Big Data and Analytics platform provides the ability to start small with one capability and easily add others over the big data journey because the pre-integration of its components reduces implementation time and cost.To find out the key insights we identify possible gaps in technology and frame work like cloud-supported, big data computing and analytics, challenges in big data computing and analytics and their solutions were not thoroughly discussed in previous researches.

Cloud computing comes with numerous big data challenges because it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, the security challenges of these systems

and technologies are applicable to cloud computing. For example, it is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely. Data security not only involves the encryption of the data, but also ensures that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms also have to be secure. The big data issues are most acutely felt in certain industries, such as telecoms, web marketing and advertising, retail and financial services, and certain government activities. The data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. This study will highlight these challenges and proposed techniques to get the solutions in the cloud.

Massive data sets are hard to understand, and models and patterns hidden within them cannot be identified by humans directly, but must be analyzed by computers using data mining techniques. The world of big data present rich cross-media contents, such as text, image, video, audio, graphics and so on. For cross-media applications and services over the Internet and mobile wireless networks, there are strong demands for cross-media mining because of the significant amount of computation required for serving millions of Internet or mobile users at the same time. On the other hand, with cloud computing boom-ing, new cloud-based cross-media computing paradigm emerged, in which users store and process their cross-media application data in the cloud in a distributed manner. Cross-media is the outstanding characteristics of the age of big data with large scale and complicated processing task. Cloud-based Big Data plat-forms will make it practical to access massive compute resources for short time periods without having to build their own big data farms. We propose a frame-work for cross-media semantic understanding which contains discriminative modeling, generative modeling and cognitive modeling. In cognitive modeling, a new model entitled CAM is proposed which is suitable for cross-media semantic understanding. A Cross-Media Intelligent Retrieval System (CMIRS), which is managed by ontology-based knowledge system KMS phere, will be illustrated[33].

VI. CONCLUSION

The purpose of the present study is big data and analytics challenges and solutions in the cloud. The researcher will identify its influences regarding framework and technical challenges like Data variety, Data storage, Data integration, Data Processing and Resource Management. Hence the purpose of this study is illuminating the big Data processing, analytics and resource management. This study covers title of the study, significance of the study, aims and objectives of the study, research hypothesis and research design. This study is designed based upon descriptive study as it aims to study the big data and analytics challenges and elaborate its solutions in the cloud. The research design contains the Literature review Theoretical and experimental analysis. Architectural factors impacts on big data and analytics on the cloud. The Data preparation and data alignment is a good solution for big data and analytics challenges. This study combines both primary and secondary research methods. Thus, gathering and analyzing the data will be done on the basis of existing research.

REFRENCES

[1]. Awodele .O ,Awodele .O, Kuyoro S.O , Osisanwo F.Y International Journal Of Computer Applications (0975 – 8887) Volume 133 – No.12, January 2016 14 Big Data And Cloud Computing Issues.

[2]. R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud Computing And Emerging It Platforms: Vision, Hype, And Reality For Delivering Computing As The 5th Utility, Future Gener. Comput. Syst. 25 (6) (2009) 599–616.

[3]. Douglas And Laney (2008), "The Importance Of „Big Data‟: A Definition," 2008. "Big Data: Science In The Petabyte Era," Nature 455 (7209): 1, 2008.

[4]. Katal, A., Wazid M, AndGoudar R.H. (2013) "Big Data: Issues, Challenges, Tools And Good Practices." Noida: 2013, Pp. 404 – 409, 8-10 Aug. 2013.

[5]. Ji, Changqing., Li, Yu., Qiu, Wenming., Awada, Uchechukwu., Li, Keqiu (2012) Big Data Processing Incloud Computing Environments, 2012 Internationalsymposium On Pervasive Systems, Algorithms Andnetworks, 1087-4089/12 $26.00 © 2012 Ieeedoi 10.1109/I-Span.2012.9 Pg 17-23

[6]. M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above The Clouds: A Berkeley View Of Cloud Computing, Technical Report Ucb/Eecs-2009-28, Electrical Engineering And Computer Sciences, University Of California At Berkeley, Berkeley, Usa (February 2009).

[7]. F. Schomm, F. Stahl, G. Vossen, Marketplaces For Data: An Initial Survey, Sigmod Record 42 (1) (2013) 15–26.

[8]. T.H. Davenport, J.G. Harris, Competing On Analytics: The New Science Of Winning, Harvard Business Review Press, 2007.

[9]. T.H. Davenport, J.G. Harris, R. Morison, Analytics At Work: Smarter Decisions, Better Results, Harvard Business Review Press, 2010.

[10]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The Kdd Process For Extracting Useful Knowledge From Volumes Of Data, Commun. Acm 39 (11) (1996) 27–34.

[11]. R.L. Grossman, What Is Analytic Infrastructure And Why Should You Care? AcmSigkdd Explorations Newsletter 11 (1) (2009) 5–9.

[12]. E.A. King, How To Buy Data Mining: A Framework For Avoiding Costly Project Pitfalls In Predictive Analytics, Dmreview 15(10). I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools And Techniques, Third Ed., Morgan Kaufmann, 2011.

[13]. Marcos D. Assunçao , Rajkumar Buyya, Rodrigo N. Calheiros , Silvia Bianchi , Marco A.S. Netto (2015), Big Data Computing And Clouds: Trends And Future Directions, J. Parallel Distrib. Comput. 79–80 (2015) 3–15, Www.Elsevier.Com/Locate/Jpd
[14]. VenkataNarasimhaInukollu1 ,Sailaja Arsi1 And SrinivasaRaoRavuri (2014) Security Issues Associated With Big Data In Cloud Computing, International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014.
[15]. A, Katal, Wazid M, And Goudar R.H. "Big Data: Issues, Challenges, Tools And Good Practices.". Noida: 2013, Pp. 404 – 409, 8-10 Aug. 2013.
[16]. F.C.P, Muhtaroglu, Demir S, Obali M, AndGirgin C. "Business Model Canvas Perspective On Big Data Applications." Big Data, 2013 Ieee International Conference, Silicon Valley, Ca, Oct 6-9, 2013, Pp. 32 - 37.
[17]. Zhao, Yaxiong , And Jie Wu. "Dache: A Data Aware Caching For Big-Data Applications Using TheMapreduce Framework." Infocom, 2013 Proceedings Ieee, Turin, Apr 14-19, 2013, Pp. 35 - 39.
[18]. Xu-Bin, Li , Jiang Wen-Rui, Jiang Yi, ZouQuan "Hadoop Applications In Bioinformatics." Open Cirrus Summit (Ocs), 2012 Seventh, Beijing, Jun 19-20, 2012, Pp. 48 - 52.
[19]. Bertino, Elisa, SilvanaCastano, Elena Ferrari, And Marco Mesiti. "Specifying And Enforcing Access Control Policies For Xml Document Sources." Pp 139-151.
[20]. E, Bertino, Carminati B, Ferrari E, Gupta A , And Thuraisingham B. "Selective And Authentic Third-Party Distribution Of Xml Documents."2004, Pp. 1263 - 1278.
[21]. Kilzer, Ann, Emmett Witchel, Indrajit Roy, VitalyShmatikov, AndSrinath T.V. Setty. "Airavat: Security And Privacy For Mapreduce."
[22]. Http://Www.Webopedia.Com/Term/C/Cloud_Computing.Html.
[23]. 5 Venkata N. I., Sailaja A., And Srinivasa R. R. (2014) Security Issues Associated With Big Data In Cloud Computing, International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014 Doi: 10.5121/Ijnsa.2014.6304 45
[24]. Applications Of Big Data: Current Status And Future Scope Sabia,SheetalKalraIssn (Print): 2319-2526, Volume -3, Issue -5, 2014 International Journal On Advanced Computer Theory And Engineering (Ijacte)
[25]. Analyzing Big Data Using Hadooptanuja A SwethaRamana D Ijirst –International Journal For Innovative Research In Science & Technology| Volume 3 | Issue 03 | August 2016 Issn (Online): 2349-601
[26]. Ieee Journal Of Biomedical And Health Informatics, Vol. 19, No. 4, July 2015 1209 Marco Viceconti, Peter Hunter, And Rod Hose Big Data, Big Knowledge: Big Data For Personalized Healthcare.
[27]. Venkata N. I., Sailaja A., And Srinivasa R. R. (2014) Security Issues Associated With Big Data In Cloud Computing, International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014 Doi: 10.5121/Ijnsa.2014.6304 45
[28]. Saranya A, Muthukumar, V.P. (2015) Security Issuesassociated With Big Data In Cloud Computing International Journal Of Multidisciplinary Research And Development Volume: 2, Issue: 4, 580-585 April 2015www.Allsubjectjournal.Com E-Issn: 2349-4182 P-Issn: 2349-5979
[29]. Venkata N. I., Sailaja A., And Srinivasa R. R. (2014) Security Issues Associated With Big Data In Cloud Computing, International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014 Doi: 10.5121/Ijnsa.2014.6304 45
[30]. Mell P., Grance T., 2011 Nist Special Publication 800-145: The Nist Definition Of Cloud Computing. Availableat:Http://Csrc.Nist.Gov/Publications/Nistpubs/800-145/Sp800145.Pdf.
[31]. Katal, A., Wazid M, AndGoudar R.H. (2013) "Big Data: Issues, Challenges, Tools And Good Practices." Noida: 2013, Pp. 404 – 409, 8-10 Aug. 2013.
[32]. QinghuaZheng, Jie Yang, Haifei Li, Mu Qiao, 2009 Ninth Ieee International Conference On Advanced Learning Technologies, An E-Learning Ecosystem Based On Cloud Computing Infrastructure .
[33]. Big Data Mining In The Cloud , Z. Shi, D. Leake, And S. Vadera (Eds.): Iip 2012, IfipAict 385, Pp. 13–14, 2012
[34]. Information Security In Big Data: Privacy And Data Mining, Lei Xu, Chunxiao Jiang, JianWang,Jian Yuan, And Yong Ren.
[35]. Katharina Ebner, ThiloBühnen, Nils Urbach ,2014 47th Hawaii International Conference on System Science, Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments.
[36]. Pavlo, A., Paulson, E., Rasin, A., Abadi, D.J., Dewitt, D.J., Madden, S., and Stonebraker, M., "A Comparison of Approaches to Large-Scale Data Analysis", Proceedings of the 2009 ACM SIGMOD International Conference onManagement of Data, 2009, pp. 165-178.
[37]. Dewitt, D., and Gray, J., "Parallel Database Systems: The Future of High Performance Database Systems", Commun. ACM, 35(6), 1992, pp. 85-98.