Big Data Mining: A Comprehensive Analysis

Hemlata
Department of Computer Science & Applications
Maharshi Dayanand University
Rohtak, Haryana, India

Abstract : Big Data is a term used to identify the datasets that cannot be managed due to their large size and complexity, with the current methodologies or data mining software tools. Big Data mining is the ability of extracting useful information from the large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most important opportunities for the next few years. In this review paper, a broad overview of the topic, its current status, its characteristics through HACE Theorem, Big Data generation and acquisition and Big Data challenges are summarised. A brief overview of the technologies of Big Data Analytics such as Hadoop and MapReduce is also presented.

IndexTerms **- Big Data, Data mining, HACE Theorem, Hadoop, HDFS, MapReduce.**

I. DEFINITION OF BIG DATA

In general, big data means the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time. Because of different concerns, scientific and technological enterprises, research scholars, data analysts, and technical practitioners have different definitions of big data.

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." On the basis of this definition, in May 2011, McKinsey & Company, a global consulting agency announced Big Data as the next frontier for innovation, competition, and productivity. Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software. This definition includes two connotations: First, datasets' volumes that conform to the standard of big data are changing, and may grow over time or with technological advances; Second, datasets' volumes that conform to the standard of big data in different applications differ from each other.[1]

As a matter of fact, big data has been defined as early as 2001. Doug Laney, an analyst of META (presently Gartner) defined challenges and opportunities brought about by increased data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety, in a research report [2]. Although such a model was not originally used to define big data, Gartner and many other enterprises, including IBM [3] and some research departments of Microsoft [4] still used the "3Vs" model to describe big data within the following ten years [5]. In the "3Vs" model, Volume means, with the generation and collection of masses of data, data scale becomes increasingly big; Velocity means the timeliness of big data, specifically, data collection and analysis, etc. must be rapidly and timely conducted, so as to maximumly utilize the commercial value of big data; Variety indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data [1]. However, others have different opinions, including IDC, one of the most influential leaders in big data and its research fields. In 2011, an IDC report defined big data as "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis." [6] With this definition, characteristics of big data may be summarized as four Vs, i.e., Volume (great volume), Variety (various modalities), Velocity (rapid generation), and Value (huge value but very low density). Such 4Vs definition was widely recognized since it highlights the meaning and necessity of big data, i.e., exploring the huge hidden values. This definition indicates the most critical problem in big data, which is how to discover values from datasets with an enormous scale, various types, and rapid generation.[1][7].

II. CHARACTERISTICS OF BIG DATA: HACE THEOREM [8]

HACE Theorem: Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. In view of these characteristics, it becomes an extreme challenge to discover knowledge or mine useful data from Big Data.

1. **H**uge Data: One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations.
2. **A**utonomous Sources: Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. The organizations which originate the data have their own method of representation of the data and no control can be applied to it. That is why this characteristic of autonomous sources is very challenging.
3. **C**omplex and Evolving Relations: With the increase in volume of Big data the complexity of the data increases. The relationship between individual data, complicate the whole data representation and the extraction process from the data.

This complication is becoming part of the reality for Big Data Applications, where the key is to take the complex (non-linear, many to many) data relationships, along with the evolving changes to discover useful patterns from Big Data collections.

III. BIG DATA GENERATION AND ACQUISITION

Value chain of big data can be generally divided into four phases: data generation, data acquisition, data storage, and data analysis. If we take data as a raw material, data generation and data acquisition are an exploitation process, data storage is a storage process and data analysis is a production process that utilizes the raw material to create new value.[9]

1. **Data generation**: Data generation is the first step of big data. Given Internet data as an example, huge amount of data in terms of searching entries, Internet forum posts, chatting records, and microblog messages, are generated. Data could be generated using Enterprise Data, IoT Data, Bio Medical Data, Data Generation from Other fields. [1].
2. **Big data acquisition**: Big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once we collect the raw data, we shall utilize an efficient transmission mechanism to send it to a proper storage management system to support different analytical applications[10].
3. **Data collection**: Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. Common data collection methods are Log Files, Sensing & Methods for acquiring network data. The current network data acquisition technologies mainly include traditional Libpcap-based packet capture technology, zero-copy packet capture technology, as well as some specialized network monitoring software such as Wireshark, SmartSniff, and WinNetCap.Data transportation Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis. Big data is mainly stored in a data center. Data transmission consists of two phases: Inter-DCN transmissions and Intra-DCN transmissions. As a strengthening technology, Zhou et al. in [11] adopt wireless links in the 60GHz frequency band to strengthen wired links.
4. **Data pre-processing** : Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is undoubtedly a waste to store meaningless data[12]. Some relational data pre-processing techniques are Integration, Cleaning & Redundancy elimination. Generally, data integration methods are accompanied with flow processing engines and search engines [13]. Authors in [14] discussed data cleaning in e-commerce by crawlers and regularly re-copying customer and account information. Apart from the data pre-processing methods, specific data objects shall go through some other operations such as feature extraction. Such operation plays an important role in multimedia search and DNA analysis [15].

IV. DIFFERENCE BETWEEN BIG DATA AND DATA MINING [4]

Big data	**Data mining**
Big data is a term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information
Big data is the asset	Data mining is the handler which provide beneficial result
Big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data.	Data mining refers to the operation that involve relatively sophisticated search operation

V. PILLARS OF BIG DATA [16]

1. Big Table – Relational, Tabular format – rows & columns.
2. Big Text – All kinds of unstructured data, natural language, grammatical data, semantic data.
3. Big Metadata – Data about data, taxonomies, glossaries, facets, concepts, entity.
4. Big Graphs – object connections, semantic discovery, degree of separation, linguistic analytic, subject predicate object.

VI. Big Data Challenges

A. *Security and Privacy*

The private information of a person is very secreat which the person might not want the Data owner to know or any other person to know about them [17]. But in Big Data the data is so diverse and big in volume that the security and privacy is the most important issue and challenge.

B. *Trust*

When any organization or any individual store the data online, it is the expectation by default that its data is secure. Each data storing site (node) should give the confidence in the clients that their data is secure. But this trust is very challenging as the data is so huge and vast in volume [18].

C. *Storage Issues*

The quantity and variety of data is so large that it is a very big challenge now a day. The quantity increases with a fraction of seconds, so it is very difficult to manage the system which stores this huge data[17]. The storage available is not enough for storing such type of data. Social Media and Sensor devices are a great contributor of data.

D. *Transport Issues*

After storage, the transportation of the data from the storage device to the user node is again a challenge. For communication over network we require high bandwidth infrastructure to transfer of data[19].

E. *Processing Issues*

Processing of large amount of data need large processors and large time. For example, a petabyte of data needs to be processed. In other words, assume that the data is chunked into bloks of 8 words, so 1 petabyte = 1K terabyte. Also assume that the processor expand 100 instructions on one block at 10 gigahertz, the time required for end-to-end processing would be 10 nanoseconds [20]. To process 1K terabyte would require a total processing time of roughly 350 years. Thus effective processing of petabyte of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information [21].

F. *Data Management and sharing*

As we know that for data to have any value it needs to be discoverable, accessible and usable. The requirement of these values increases when the data is big [22]. These requirements must be achieved while following all privacy laws. Data also needs to be accurate, complete and timely if it is to be used for support complex analysis and decision making [23]. For these reasons, management and governance focus needs to be on making data open and available via standardized API's, formats and metadata [24].

G. *Requirement of diverse skills*

Since Big Data is very wide concept, so it requires the individuals with diverse and new skill sets which should not be limited to technical but also extend to research, analytical, interpretive and creative ones. So, there is a requirement of training programs to be organized by different organizations [25].

H. *Technical Challenges*

a) **Fault tolerance computing**: With the emergence new technologies like Big Data and Cloud Computing, it is desired that if any failure occurs, the damage should be within acceptable threshold. To device a fault tolerant system is extremely hard. So, we should reduce the probability of failure to an "acceptable" level [26].

b) **Scalability**: It is a very big challenge as the processor technology has changed in recent years. It leads towards cloud computing via high level sharing of resources[27].

c) **Quality of Data**: In Big Data we collect a huge amount of data but it is very expensive in terms of storage. Sometimes, for decision making we need large amount of data e.g. for prediction etc. But other times we required the correct data as compared to huge data [27]. So, it is very difficult to judge which data is relevant. It is also a challenge to ensure how much data would be enough for decision making.

d) **Heterogeneous Data**: Big Data is a collection of huge amount of unstructured data [28]. To manage this unstructured heterogenous data is very cumbersome and costly.

VII. Technologies Available for Big Data

A. *Hadoop*

Hadoop [29] is written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system. Presently, it is used on large amounts of data. With Hadoop, enterprises can harness data that was previously difficult to manage and analyze. Hadoop is used by approximately 63% of organizations to manage huge number of unstructured logs and events (Sys.con Media, 2011). In particular,

Hadoop can process extremely large volumes of data with varying structures (or no structure at all). Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data.[30]

1. *HDFS*.: This paradigm is applied when the amount of data is too much for a single machine. HDFS is more complex than other file systems given the complexities and uncertainties of networks. Cluster contains two types of nodes. The first node is a name-node that acts as a master node. The second node type is a data node that acts as slave node. This type of node comes in multiples. Aside from these two types of nodes, HDFS can also have secondary name-node[30]. HDFS stores files in blocks, the default block size of which is 64 MB. All HDFS files are replicated in multiples to facilitate the parallel processing of large amounts of data.
2. *HBase*.: HBase is a management system that is open-source, versioned, and distributed based on the BigTable of Google. This system is column- rather than row-based, which accelerates the performance of operations over similar values across large data sets. For example, read and write operations involve all rows but only a small subset of all columns. HBase is accessible through application programming interfaces (APIs) such as Thrift, Java, and representational state transfer (REST) [31]. These APIs do not have their own query or scripting languages. By default, HBase depends completely on a ZooKeeper instance.
3. *ZooKeeper*: ZooKeeper maintains, configures, and names large amounts of data. It also provides distributed synchronization and group services. This instance enables distributed6 The Scientific World Journal processes to manage and contribute to one another through a name space of data registers (z-nodes) that is shared and hierarchical, such as a file system [32]. Alone, ZooKeeper is a distributed service that contains master and slave nodes and stores configuration information.
4. *HCatalog:* HCatalog manages HDFS. It stores metadata and generates tables for large amounts of data. HCatalog depends on Hive metastore and integrates it with other services, including MapReduce and Pig, using a common data mode [33]l. With this data model, HCatalog can also expand to HBase. HCatalog simplifies user communication using HDFS data and is a source of data sharing between tools and execution platforms.
5. *Hive:* Hive structures warehouses in HDFS and other input sources, such as Amazon S3. Hive is a subplatform in the Hadoop ecosystem and produces its own query language (HiveQL). This language is compiled by MapReduce and enables user-defined functions (UDFs). The Hive platform is primarily based on three related data structures: tables, partitions, and buckets. Tables correspond to HDFS directories can be distributed in various partitions and, eventually, buckets.
6. *Pig:* The Pig framework generates a high-level scripting language (Pig Latin) and operates a run-time platform that enables users to execute MapReduce on Hadoop. Pig is more elastic than Hive with respect to potential data format given its data model. Pig has its own data type, map, which represents semistructured data, including JSON and XML.
7. *Mahout:* Mahout is a library for machine-learning and data mining. It is divided into four main groups: collective filtering, categorization, clustering, and mining of parallel frequent patterns [34]. The Mahout library belongs to the subset that can be executed in a distributed mode and can be executed by MapReduce.
8. *Oozie:* In the Hadoop system, Oozie coordinates, executes, and manages job flow [35]. It is incorporated into other Apache Hadoop frameworks, such as Hive, Pig, Java MapReduce, Streaming MapReduce, and Distcp Sqoop. Oozie combines actions and arranges Hadoop tasks using a directed acyclic graph (DAG). This model is commonly used for various tasks.
9. *Avro:* Avro serializes data, conducts remote procedure calls, and passes data from one program or language to another. In this framework, data are self-describing and are always stored based on their own schema because these qualities are particularly suited to scripting languages such as Pig.
10. *Chukwa:* Currently, Chukwa is a framework for data collection and analysis that is related to MapReduce and HDFS [36]. This framework is currently progressing from its development stage. Chukwa collects and processes data from distributed systems and stores them in Hadoop. As an independent module, Chukwa is included in the distribution of Apache Hadoop.
11. *Flume:* Flume is specially used to aggregate and transfer large amounts of data (i.e., log data) in and out of Hadoop. It utilizes two channels, namely, *sources* and *sinks*. Sources include Avro, files, and system logs, whereas sinks refer to HDFS and HBase. Through its personal engine for query processing, Flume transforms each new batch of Big Data before it is shuttled into the sink.

Table 1 Hadoop Components and their Functionalities [30]

Hadoop Component	Functions
HDFS	Storage and replication
MapReduce	Distributed processing and fault tolerance
HBASE	Fast read/write access
HCatalog	Metadata
Pig	Scripting

Hive	SQL
Oozie	Workflow and scheduling
Zookeeper	Coordination
Kafka	Messaging and data integration
Mahout	Mavhine learning

Table 2 Hadoop Usage [30]

	Specified Use	Used by
1	Searching	Yahoo, Amazon, Zvents
2	Log processing	Facebook, Yahoo, ContexWeb.Joost, Last.fm
3	Analysis of video and images	New York Times, Eyelike
4	Data Warehouse	Facebook, AOL
5	Recommendation systems	Facebook

B. MapReduce[11]

Big data has generated a whole new industry of supporting architectures such as MapReduce. MapReduce is a programming framework for distributed computing which was created by Google using the divide and conquer method to break down complex big data problems into small units of work and process them in parallel [12]. MapReduce can be divided into two stages [11]:

1. **Map Step**: The master node data is chopped up into many smaller subproblems. A worker node processes some subset of the smaller problems under the control of the JobTracker node and stores the result in the local file system where a reducer is able to access it.
2. **Reduce Step**: This step analyzes and merges input data from the map steps. There can be multiple reduce tasks to parallelize the aggregation, and these tasks are executed on the worker nodes under the control of the JobTracker.

Figure 1 MapReduce Processing Pipeline

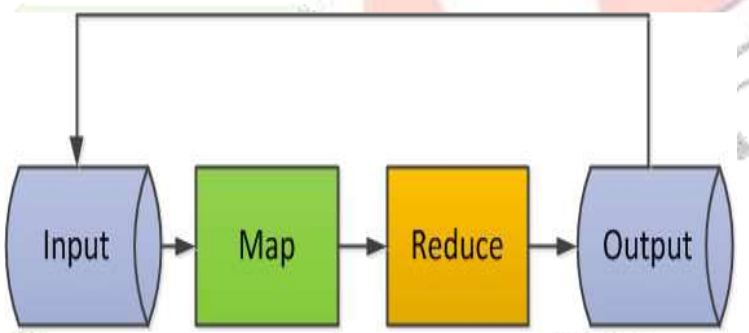
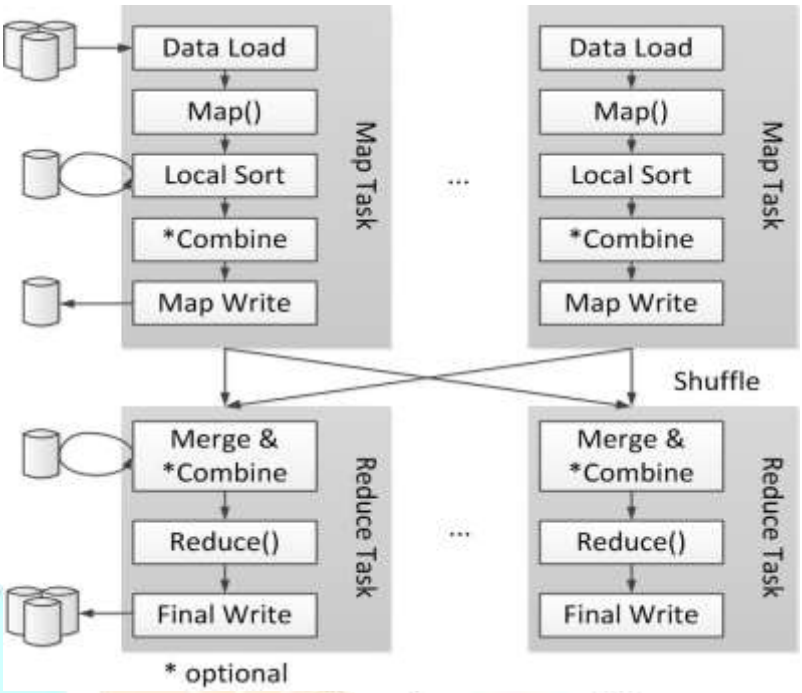


Figure 2 Architecture of MapReduce Model



REFERENCES

[1] Min Chen · Shiwen Mao · Yunhao Liu, "Big Data: A Survey", © Springer Science+Business Media New York 2014, published online: 22 january 2014.

[2] Laney D (2001), "3-d data management: controlling data volume, velocity and variety", META Group Research Note, 6 February.

[3] Zikopoulos P, Eaton C et al (2011) "Understanding big data: analytics for enterprise class hadoop and streaming data" McGraw-Hill Osborne Media.

[4] Meijer E (2011), "The world according to linq. Communications of the ACM 54(10):45–51.

[5] Beyer M (2011) "Gartner says Solving big data challenge involves more than just managing volumes of data" Gartner. http://www.gartner.com/it/page.jsp.

[6] Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView, pp 1–12.

[7] Mayer-Sch¨onberger V, Cukier K (2013), "Big data: a revolution that will transform how we live, work, and think" Eamon Dolan/Houghton Mifflin Harcourt.

[8] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, VOL. 26, NO. 1, January 2014.

[9] Subita Kumari, "Big Data: A Detailed Review", IJIFR, Volume-1, Issue-7, March 14.

[10] Hemlata, Dr. Preeti Gulia, "Comprehensive Study of Open- Source Big Data Mining Tools", International Journal of Artificial Intelligence and Knowledge Discovery, e-ISSN: 2231- 0312, Vol. 6, Issue 1, January, 2016.

[11] Firat Tekiner and John A. Keane ,"Big Data Framework", 2013 IEEE International Conference on Systems, Man, and Cybernetics.

[12] Xin Luna Dong, Divesh Srivastava, "Big Data Integration", ICDE Conference, IEEE 2013

[13] Kumari, Subita, and Pankaj Gupta. "Proposed Architecture of MongoDB-Hive Integration." International Journal of Applied Engineering Research 12.15 (2017): 5000-5004.

[14] Kohavi R, Mason L, Parekh R, Zheng Z, "Lessons and challenges from mining retail e-commerce data", Mach Learn 57(1-2):83–113(2004).

[15] Kamath U, Compton J, Dogan RI, Jong KD, Shehu A, "An evolutionary algorithm approach for feature generation from sequence data and its application to dna splice site prediction", IEEE/ACM Transac Comput Biol Bioinforma (TCBB) 9(5):1387–1398(2012)

[16] D.Usha and Aslin A.P.S., "A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce", Internetional Journal of Current Engineering and Technology, ©2014 INPRESSCO®

[17] Hemlata, Gulia, P. (2018). DCI3 Model for Privacy Preserving in Big Data. In Big Data Analytics (pp. 351-362). Springer, Singapore.

[18] Kumari, Subita, and Pankaj Gupta. "Implementation of CouchDBViews." Big Data Analytics. Springer, Singapore, 2018. 241-251.

[19] Hemlata, Gulia, Preeti. "Novel Algorithm for PPDM of Vertically Partitioned Data." International Journal of Applied Engineering Research 12.12 (2017): 3090-3096.
[20] Dhruba Borthakur et al, "Apache Hadoop Goes Realtime at Facebook", SIGMOD .'11, June 12.–16, 2011, Athens, Greece,Copyright 2011 ACM .
[21] Wei Fan, Albert Bifet," Mining Big Data: Current Status, and Forecast to the Future," SIGKDD Explorations,Volume 14, Issue 2.
[22] Hemlata Chahal, "ID3 Modification and Implementation in Data Mining" International Journal of Computer Applications (0975-8887) Volume 80- No7, October 2013
[23] Seref SAGIROGLU and Duygu SINANC, "Big Data: A review", IEEE January 2013.
[24] Duren Che, Mejdl Safran, Zhiyong Peng, " From Big Data to Big Data Mining: Challenges, Issues and Opportunities", © Springer-Verlag Berlin Heidelberg, 2013.
[25] Zhou X, Zhang Z, Zhu Y, Li Y, Kumar S, Vahdat A, Zhao BY, Zheng H (2012) Mirror on the ceiling: flexible wireless links for data centers. ACM SIGCOMM Comput Commun Rev 42(4):443–454
[26] Leung K-S, Lee KH, Wang J-F, Ng EYT, Chan HLY, Tsui SKW, Mok TSK, Tse PC-H, Sung JJ-Y, "Data mining on dna sequences of hepatitis b virus", IEEE/ACM Transac Comput Biol Bioinforma 8(2):428–440(2011).
[27] Agrawal, D. & Aggarwal, C. (2001), On the design and quantification of privacy preserving data mining algorithms, in 'Proccedings of the 20th ACM SIGACT-SIGMODSIGART Symposium on Principles of Database Systems', Santa Barbara, California, USA.
[28] Zaiying Liu, Ping Yang and Lixiao Zhang (2013) "A Sketch of Big Data Technologies" IEEE 2013 Seventh International conference on Internet Computing for Engineering and Science.
[29] A. Hadoop, "Hadoop," 2009, http://hadoop.apache.org/.
[30] Nawsher Khan et-al, "Big Data: Survey, Technologies, Opportunities, and Challenges", Hindawi Publishing Corporation The Scientific World Journal Volume 2014, Article ID 712826, 18 pages http://dx.doi.org/10.1155/2014/712826
[31] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler , "The Hadoop Distributed File System", IEEE 2010.
[32] Du Zhang "Inconsistencies in Big Data", 12th IEEE Int. Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC'13) , 2013
[33] Raja Appuswamy, Christos Gkantsidis, Dushyanth Narayanan, Orion Hodson, and Antony Rowstron," Scale-up vs Scale-out for Hadoop: Time to rethink?", SoCC'13,1–3 Oct. 2013, Santa Clara, California, USA, ACM 978-1-4503-2428-1.
[34] Mirko K¨ampf and Jan W. Kantelhardt, "Hadoop.TS: Large-Scale Time-Series Processing", International Journal of Computer Applications (0975 - 8887) Volume 74 - No. 17, July 2013.
[35] Dennis A. Ludena R. , Alireza Ahrary, "A Big data Approach for a new ICT Agriculture Application Development", 2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, IEEE Computer society,2013.
[36] Huang Z, Shen H, Liu J, Zhou X, "Effective data coreduction for multimedia similarity search", In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, pp 1021–1032(2011).