Future Prospects and Challenges of Big Data with Big Data Problems

Reshu Grover, Manisha Agarwal

Abstract: ***As we aware, every second, petabytes of data is being generated from various sources as mentioned in previous para at a rapid speed, other sources may be stock market transactions data, sales and marketing data, sensors data, web documents, internet images, movies, multimedia data and lot many. These Big Data are as volume, variety, velocity, veracity and value in 5 V's, which explains the complexity of Big Data. Due to adoption of Big Data analytics, there is demand of efficient technologies to manage heterogeneous data]. 21st century has marked the advent of Rich-Data concept. Different societal application such as disaster management, urban planning & monitoring, health hazards etc. has reinvigorated the significance of Geographic Information Systems (GIS). Newer and advance technologies are generating means to produce location-based dataset called as geospatial datasets. Increasing influx of user-generated data has amplified geospatial data and has outdid conventional computation requirement limits. The similitude of Big data and data in consideration has evolved the concept and appellation as "Spatial Big Data" (SBD). This paper is an effort towards enlistment and analysis of the different SBD concepts and technologies in the contemporary time. The main endeavor of this paper is to critically analyze the different technologies in the present day and identify the existing technical inadequacies in the existing systems***

Index Terms**: Big data,** ***Spatial Big Data, Social networking***

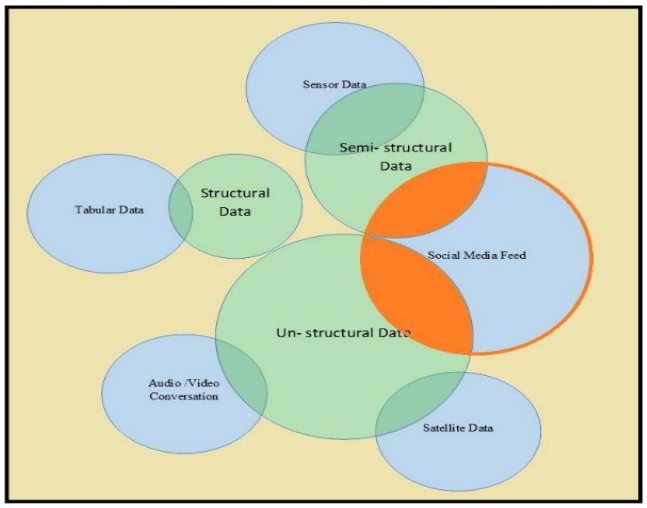
I. INTRODUCTION

Non-Traditional data (un-structured) and its characteristics is the biggest concern of analysists in the highly evolving data analysis scenario. In the contemporary world, advancement in technologies are rapidly and dynamically generating humongous amount of data for further analysis. These generated data could belong to the same category/class of information but the occurrence might happen at geographically diverse locations, leading it to be considered as different category. The basic natures of such data is predominantly unstructured &/or semi-structured and can be of heterogeneous-format, multi-scale, multi-resolution and multi-temporal as well leading to increase in complexity of handling of such data sets (fig 1). . During the last decade, it has been perceived that approximately 80% of the data being generated is linked with geographical location markers, consequently making data storage in spatial form a compulsion [22]. Various organizations have identified spatial information crucial for decision making in their business objectives [20]. To cope with position complexity these data are linked with location, hence there is exponential increase of data size. Geographic Information System (GIS) is one such platform which has provided solutions regarding operating location intrinsic data. It was in the mid-19th century when new perspectives in the field of cartography came to light. This perspective combines cartography and computational principles to produce automated digitized maps. The difference between traditional data and big data is shown in Table 1.

Table 1 Differences between big data and traditional data

Feature	Traditional Data	Big Data
Volume	GB	Constantly updated (PB or EB recently)
Generated rate	Per hour, day...	More rapid
Structure	Structured	Structured, semi-structured and un-structured
Data source	Centralized	Fully distributed
Data integration	Easy	Difficult
Data store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real-time

Fig. 1.: Different form of Data



Canadian government played a very important role in emergence of GIS with government initiatives . One such initiative was taken by Roger Tomlinson, an

Revised Manuscript Received on July 22, 2019.

Reshu Grover, Research Scholar, Bansthali Vidyapeeth, Jaipur, Rajasthan, India, reshugrover3@gmail.com
Manisha Agarwal, Associate Professor, Bansthali Vidyapeeth, Jaipur, Rajasthan, India

English geographer in mid-1960. Jack Dangermond founder E.S.R.I (1969), is a leading commercial firm which mostly focused on developing GIS application for government and commercial sectors.
During Cold War U.S defense agencies used associated technologies of GIS viz. global positioning systems (GPS) and Radio Frequency Identification Tags (RFID).
From late 1980 multiple advancement in the technical field of GIS and computational systems has enabled multiple products to thrive in the field of GIS related software, enabling spatial analysis to be financially feasible commodity in today's market. In the current era of smart phones, ipads and laptops, GIS and the spatial data technology has become standard highly accepted system universally.

Three is huge amounts of data are generating from difference sources, like sensor's data, IoT data, weather forecasting data, social networking data, military data, technology development data, Smart grid data etc. all data are may be in different format and different volume, all machine learning and data mining approaches cannot handle all types of data due heterogeneity or data size. So, need a proper technique or an approach that can handle, analyses accurately and efficiently. Using big data analytics, we can integrate data and handle heterogeneity of data. Which is the main feature of the big data and big Data analytics. In this paper we have introduces a data processing method for handling these using Data mining techniques, Big data tools and machine learning (ML), Deep learning with their challenges to handle these data with respect to Big Data analytics.

This paper is planned into five sections. The first section dicusses the general overview of the problem area, Characteristic, processing methods that can handle metagenetic. In section 2 we have given related work. In section 3 we have given big data and data mining tools. Section 4 shows the limitations of existing system and Final section concludes our paper.

A. Characteristic of Spatial Big Data

Spatial data are those data which is integrated with geolocation parameters. Any data that has location associated with information is defined as spatial data. Spatial data also known as geographical data has two parts the attribute data i.e. the information about the event and the spatial data i.e. the location where the event has occurred (Table 2). The property "Location" does not only mean latitude and longitude of the point or region but also projection, datum and topology, hence handling of spatial data is a complex phenomenon.

Table 2: Characteristics of Spatial Data

Spatial Data			
Attribute Data	Geographic Data		
	Location Data (Neighborhood Characteristics	Projection Data (Heterogeneous Characteristics)	Topology Data (Fuzzy Nature)

The elementary characteristic of spatial data is derived from the 1st law of geography, which states that "Everything is related to everything else but near things are more related than distant things" [24]. Therefore; spatial data is portrayed by the spatial dependency i.e. values of variable in a particular location are related to the value of same variable in its neighborhood. The second essential characteristic of spatial data is heterogeneity. It is the distinctive characteristics of each place, demonstrating that spatial data exhibit stationary characteristics not often[1]. This characteristic is also known as non- stationarity. The third and the last characteristic of spatial data is its exceptionally fuzzy nature.

In the era of smart-devices, geotagging has become a primitive feature. Multiple social media platforms associated with location- based services are generating data at alarming rate in the last few years [20]. Large number of people express their feeling, information and experiences through social media resulting in enormous data for processing. Engagement of youth on social media provide the analysist with an accurate platform. Recent studies have proven that 10% of the random tweets over the internet conation location [19]. Social media is a global platform and the real time data generated over it, is what has created the concept of Spatial Big data. Various example of spatial big data include check-ins[19] by various user globally at the same time and understanding the epicenter of each check-ins, GPS-tracking of user using smart devices, Unmanned aerial vehicle (UAV)/Wide area motion imagery (WAMI) [19] video and generating roadmaps from various user generated content, Waze, Open Street Map etc.
Spatial Big-Data is a sub part of Big Data that have location as its core property. Extend of information that produced, processed and analysis is far-fetched. Categorizing any data as spatial big data depends on the context for example in the case of disaster, copious response of the user in fraction of second grant velocity to spatial big data. Spatial Big-Data can provide answer to global question at micro level and present analyst with much refined and accurate result. Management of such intense data requires a complex tightly coupled system, which deals with the 3+1 V,'s of the spatial big data.
Volume: Amount of data generated by various digital devices over the internet is of great concern for many analysist. This voluminous data demand to ameliorate the conventional data management systems for fast processing.
Velocity: Another important feature of big data is the continuously increasing data coming from various different sources at a very fast rate. This unremitting data generate by various advance technology worldwide at rapid rate is providing velocity to Big-Data.
Variety: Nowadays, anyone can generate data about any entity in consideration. Modern technology has provided one with various means to investigate any ongoing event. Advance technologies such as sensors, RFIDs, satellite imageries, online platform etc. can generate precise data about the very same entity. These means provide same data in various different

forms have it been structured, unstructured or semi-structured. Humongous data coming from different platform for analysis result in the immense variety of Big-Data.
Veracity: The messiness of the data available for analysis is a major reason for the failure of conventional system.

Heterogeneous typed data generated over the internet result in complex preprocessing of the data. Data captured from multiple sources are unstructured in nature as well as exit in different file formats.

B. Data Processing Methods for Heterogeneous Data and Big Data Analytics

Data Cleaning : This process used to identify, and find the unstructured and unreasonable data and after finding, these data can be modify or delete or improve for enhancing the data quality [12]. Like noisy problem in healthcare data, missing some values and incomplete data, impure data problems in big data. Due to this type of problem, decision and retrieval of data and develop a big data, machine learning or deep learning approach is not easy , so to improve the data quality data cleaning is necessary [14].

Data Integration: In integration process variety of data has to be integrate, aggregate and merge on the basis of their matched or shared variable. In latest and advanced data processing and analysis approaches, various format data like structured, unstructured or semi structured can be combine that will give new data analyses way and results to make new decisions manually or using latest machine learning of bog data tools [15].

Dimension Reduction and Data Normalization: By reducing the dimensionality of the data feature matrix, we can reduce the structure of data for better analysis and getting meaning full data from heterogeneous datasets [16].

Data collection in big data and machine learning iis shown in Figure 2 [30].

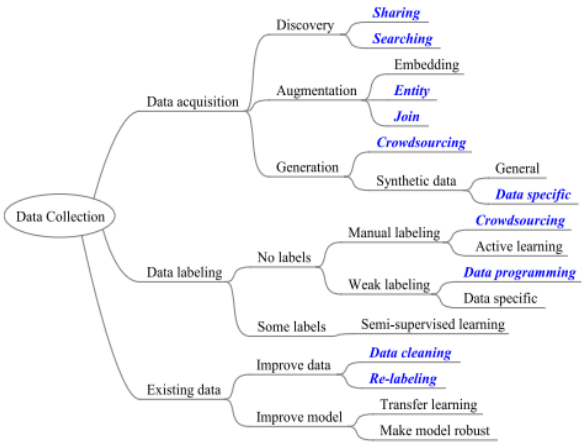


Fig. 2 Data Collection in Big Data and Machine Learning

II. RELATED WORK

Exponential growth in data volume has antiquated traditional methods for data processing and handling [23]. Eskandari L.et. al. [7] paper stated that the current technologies are creating data at accelerating rate approximately 30,000 Gigabytes per second. Secondly, report published by international data corporation state that the current decade (2010-2020) is monitoring doubling of data per year [9]. Big data analytics is the best solution for working with such kind of data. Study conducted by Veda C. Storey et.al; has thoroughly examine the characteristic of the data and stated which characteristic can be handle by which domain (technological domain or software domain)

Table 3: Characteristic of Big Data and its solution

Characteristic of Big Data	Technology Solution	Software Solution
Volume	Yes	
Velocity	Yes	Yes
Variety		Yes
Veracity		Yes

Eric Evans distributed database studies has pointed that data driven paradigm require radical change in data storage techniques, hence use of NoSQL database are encouraged [5]. NoSQL is BASE theorem-based data storage mechanism, which follow no fixed schema and provide horizontal scaling. Study conducted by Veda C. Storey et.al; has mentioned various different database and their capability to work with BASE theorem Table 4.

Table 4: NoSQL verses SQL database

Database	Relational	SQL	Column store	Scaling	Eventual Consistency	BASE	Large Data Volume	Schema
SQL	Yes	Yes	No	Limited	-	No	NO	Fixed
NoSql	No	No	Yes	Yes (Horizontal)	Yes	Yes	Yes	Yes

Another study conducted by Ali Davoudian et. al & Abdul Haseb et.al; presented classification of NoSQL Database and comparison of different database available in market Table 5.

Table 5: Different NoSQL Table

Parameter	Neo 4J	Dynamo	Couch db	Mongo db	Bigtable	Hbase	Casandra	Maria db	Create Db	C-store
Data Model	Graph	Key Value	Document	Document	Column	column	Column	column	Column	Column
Scalability	Low	High	Medium	Medium	High	High	High	High	High	Medium
Data Size	Low	High	Medium	Medium	High	High	High	High	High	High
Data Complexity	High	Low	Medium	Medium	Low	Low	Low	High	Medium	Low
Communication Protocol	SSL	HTTP	SSL	SSL	-	SSH	SSL	SSL	SSL	SSL

Microblogging is a crucial part of everyone's life. Various studies have shown

that such practices are rich source of timely data for valuable information [23]. Smith et al; has stated that 88% of US adults use internet, 77 % owns smart phone and 69 use social media. 30 % of this tagged location in their post [18].
Geo-tag information over microblogging sites is Volunteered Geographic Information (VGI), which is pervasive in nature, and make each citizen a sensor. Cugler et. al. paper has also mentioned the role of VGI in increasing velocity of geospatial data.

These improved service models provide broad geospatial information to upkeep social popularization (Ye 2008; Luo et al., 2009).
Geographical data is consolidated in hierarchical data object by spatial database. ArcGIS is one of the most important solution provider to process geospatial data.

White papers of ERSI has stated working principals of ArcGIS's geodatabase. "Working with the geodatabase: Powerful multiuser editing and sophisticated data integrity" in 2012 demonstrate the structure, working and various feature of geodatabase [27].
Another paper "Understanding Coordinate management in geodatabase" in 2007 illuminated the multi dimensionality of geodatabase that is in 2, 3 or 4 dimension. This paper also exemplifies the importance of m dimension [26].
Both SQL and NoSQL based database management systems has provided solution for geospatial data. MySQL and Postgre- SQL has incorporated OGC's SFS and SFSQL to provide the functionality of spatial analysis, but face the problem with scalability of the current data. In S. Schade, (ISPRSarchives, 2015) paper optimal treatment of the big data to handle geospatial data was introduced. In order to investigate the big data of geospatial nature presented over distribute geography a different approach needs to be reformed. To deal with the large scale of unstructured data column based NoSQL database such as Neo4j, RIAK, Hbase, MapReduce and Cassandra can be used[19]. V .Kantere et al paper stated that in case of decentralized spatial information, a peer to peer paradigm can be beneficial [24]. To achieve fast [6]. MongoDB and HBase has provider analyst with advancement of Column- oriented database systems which support OLAP or join processing. Zhange et. al. paper stated that Apache Hadoop is a framework that facilates parallelization, remote execution, data distribution, load balancing, or fault tolerance while working with SBD [29]. Other then Apache Hadoop; Pregel [18], GraphLab [17], Power-Graph [27], HaLoop [3], PrIter [26], and CIEL [17] also provide solution for processing SBD. Lee et. al. paper stated how Complex event processing (Oracle CEP and Esper) and Spatial-OLAP (JMap and GeoMondrian)can be used to manipulate SBD. Major CEP engine doesn't provide parallel processing but Interstage Big Data CEP Server by FUJITSU [13].

III. NEXT GENERATION REQUIREMENTS FOR BIG DATA ANALYSIS AND MACHINE LEARNING

As big data is the combination of three Vs (huge Volume, high Velocity, and huge Variety) that may be in inconsistency or consistent, it is widely used software analysis approach to find the meaning full data from various format data. By using these tool industries and organization are easily able to analyses their dataset for making good decision and benefit of the organization that also improved the operational efficiency of the organizations. In this section we have discussed few requirements for next generation technologies for big data, machine learning and deep learning.

Next generation future requirements of Big data, Machine learning and Deep learning technologies

1. How new approaches like Machine Learning, Deep Learning and Big Data are different from traditional data mining approaches
2. Query format for the Heterogeneous Data
3. Search Quality Improvement for heterogeneous data
4. How to Optimizing for Evolving Data due to Diverse Data Models
5. Data Inconsistency due to Inadequate Resources
6. Handle the growth of the Internet —
7. Real-time processing
8. Process complex data types with heterogeneous data
9. Efficient indexing and ranking algorithms for heterogeneous data with Information retrieval system for heterogeneous and big
10. How new issues related Machine Learning, Deep Learning and Big Data can be identified easily?
11. How Machine Learning, Deep Learning and Big Data approaches can be combined for more data samples available in market for better learning and analysis to make good decisions and reduce the risk or the organizations.
12. How all type of data can be retrieved within seconds to make decisions that will also help the clients or the customer to fats processing and they can get anything they need digitally.
13. Natural language data analysis, data modeling and Knowledge Extraction problem for big data and machine learning
14. Uncertainty problem in Machine Learning, Deep Learning and Big Data.
15. How new algorithms can be design to process heterogeneous and big data and how these approaches can improve the quality of service (QoS) generation technologies.
16. How handle the growth of data on the internet and how to process these data?

IV. LIMITATION OF EXISTING SYSTEMS

Veda C. Storey et.al. Paper has highlighted some of the shortcoming of tradition database systems while working with such data [23]. These drawback are 1) single point of failure, 2) expensive with respect to amount of data, 3)

impedance mismatching (aggregated verses atomic value [23] , and 4) distributed processing (high complexity new node to data balance [5] and performance decrease as join and transcation is difficult in distributed environment [5]).
In Grace Park et.al; paper, it states that considering the volume of the data, big data analysis may also fail due to the following error / reasons [27]: 1) lack of data content, Inaccurate Metadata and 3) batch-oriented system and their issues with real time data processing.
With increasing number of analysis proceeding at same time a new problem came into existence i.e. the problem of high concurrent access (Yang and Huang, 2013). Cugler et. al. paper stated that MapReducd framework faces problem while working with multiple iterations.

V. OUTCOME AND INFERENCE

The conventional storage technologies, designed for Spatial Big Data are providing new and integrated technologically sound architecture for the efficient ETL operations. Market pioneers like those that Hadoop's Map Reduce, Oracle 11C, PostGre-SQL and Geo-Database have successfully integrate the Spatial characteristic in the core framework and fabricated advance layer of spatial analysis. Participation of NoSQL databases in the era of SBD have further streamline the processing of unstructured spatial data. The Column oriented database such as Createdb, Bigtable, Hbase and Casandra etc. have tested processing of spatial big data with a better efficiency then SQL database.
In a distributed environment with data flowing from diverse direction at alarming rate, central processing and storage has more drawbacks them advantages. Range partition approach works with fragments to scan dataset but the efficiency decrease with increasing fragments by range Partition [30]. Spatial Proximity Partion storage SPPS approach designed by Zheng et. al. have altered the efficiency of spatial query's and scan resulting from invalid fragment in distributed column database. SPPS model has used variant of hashing algorithm i.e. Geohash and variant prefix tree to reduce scan period.
The Indexing algorithm have played a very crucial role in multiple data processing engine. Two of the most important indexing algorithms, Inverted index / hashing is applicable almost in every application for indexing unstructured data. The pros and cons of the two are in the Table 6. The combination of these two algorithms provide a better indexing algorithm for unstructured data.

Table 6: Inverted Index verses Hashing

Algorithm / Property	Inverted Index	Hashing
Full Text Scan	Yes	No
Scalability	Limited	Very High
Maintenance Cost	High	Low
Storage Overhead	High	Medium

This approach is texted on a text file with approximate 47000 indexes and seek time turns out to be best.

There are different types of data are available with different organizations, Structured Data (spreadsheets and relational databases), Semi-Structured Data(like; XML documents, NoSQL database, and some object-oriented based databases) , Unstructured Data (Examples are e-mail, MS word documents, videos, photos, audio files, presentations, Web Pages and many other kinds of business documents) and Heterogeneous Data(any type of data we have discussed previous types) [1,2].
Big data was referenced the earliest by open source project Nutch in Apache Software Foundation, which was used to describe the analysis of a large number of data sets in web search applications [9]. Different industries have some consensus, but have not a unified definition for the big data. In 2012, Gartner updated its definition in [10] as follows: "Big data is high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." The representative features of big data are 3Vs: volume, variety, velocity. International Data Corporation thinks that the 4th V of big data is value with sparse density but IBM regards as veracity is the 4th V [11].

Data is an integral part of most business-critical applications. As business data increases in volume and in variety due to technological, business, and other factors, managing this diverse volume of data becomes more difficult. A new paradigm, data virtualization, is used for data management. Although a lot of research has been conducted on developing techniques to accurately store huge amounts of data and to process this data with optimal resource utilization, research remains on how to handle divergent data from multiple data sources.

Heterogeneity of data may be in differ types or at different levels for Big data and Machine Learning

- At different language data problem means Syntactic heterogeneity.
- Semantic heterogeneity at the logical level and data modeling level for different domains.
- At data storage, schema mapping and data integration level
- At relationship level heterogeneous data, query level, and data retrieval level
- At decision making level on the basis of exiting dataset
- At technology selection level for heterogeneous data in big data and machine learning

A. General Problems of Big Data

Big data can provide big success opportunities. However, as with most emerging technologies, several characteristics are associated with big data problems that make them technically challenging. These general problems or challenges of big data can be grouped in three categories: data, process, and management.

- Data Challenges (Volume)- How to deal with the huge volumes of Heterogeneous Big data in terms of processing and storage?

- Velocity- How to respond to flood of information in a real-time manner on time for the users or query?
- Variety- How to deal with multiple format data and heterogeneous data and structured from various sources?
- Veracity- How to deal with following types of problems associated with Big Data like- Invalidity, Untruths, Missing Values or Uncertainty, Coverage of Data, Reading Values, Quality of Data?
- Value- What type of data and how much data is required for analysis and for users?

B. Need of Big Data Analytics

- Storage of Traditional and structured data, semi-structured data, unstructured data and heterogeneous at one place.
- Insight into hidden data and processing of all types of data
- Resource utilization with minimum availability of hardware and software with cheaper cost
- Quick analysis of Big data for market requirements
- Performance analysis and improvement using Big Data analytics tools and frameworks, Query optimization and efficient result for the users
- Privacy of data and user's information

VI. CONCLUSION

There is no best solution available in the market and a lot of option is present to improve upon the existing platforms, to keep up with the ever-increasing demand.

Machine learning seems to be eating the world with a new breed of high-value data-driven applications in image analysis, search, voice recognition, mobile, and office productivity products. To paraphrase Mike Stonebreaker, machine learning is no longer a zero-billion-dollar business. As the home of high-value, data-driven applications for over four decades, a natural question for database researchers to ask is: what role should the database community play in these new data driven machine-learning-based applications?

Big data and Machine learning can be used to discover hidden patterns, removing the uncertainty, Finding the useful information to make good decision in the organizations or at user level

In closing the discussion, we emphasize that the opportunities and challenges brought by big data are very broad and diverse, and it is clear that no single technique can meet all demands. In this sense, big data also brings a chance of "big combination" of techniques and of research. The above challenges about data mining, big data machine learning can be further research topics.

As the Big data has big potential that may provide the better solution for the organizations. As now a day's data are available in various format and generating in exponential way, there should be some approach that can handle this heterogeneity at different level.

In our research identification and selection, we have selected the latest topic, that may be merge with heterogeneous data in Big Data.

REFERENCES

1. Amol Barewar Mansi A. Radke U. Deshpande(2014) : Geo Skip List Data Structure - storing spatial data and efficient search of geographical locations, 978-1-4799-3080-7/14/$31.00_c 2014 IEEE
2. Borthakur, D.: The hadoop distributed file system: Architecture and design. Hadoop Project Website 11, 21 (2007)
3. Bu, Y., Howe, B., Balazinska, M., Ernst, M.D.: Haloop: Efficient iterative data processing on large clusters. Proceedings of the VLDB Endowment 3(1-2), 285–296 (2010)
4. Ls Carlo Strozzi, NoSQL Relational Database Management System: Home Page, URL http://www.strozzi.it/cgibin/CSA/tw7/I/en_US/nosql/Home-Page, 1998 (accessed 05.07.13).
5. Corbellini, A., Mateos, C., Zunino, A., Godoy, D., & Schiaffino, S. (2017). Persisting big-data: The NoSQL landscape. Information Systems, 63, 1–23.
6. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Communications of the ACM 51(1), 107–113 (2008)
7. Eskandari L, Huang Z, Eyers D (2016) P-Scheduler: adaptive hierarchical scheduling in apache storm.In: Proceedings of the Australasian Computer Science Week Multi -conference , ACSW 2016, No. 26. ACM Press, New York
8. Ghemawat, S., Gobioff, H., Leung, S.: The google file system. In: ACM SIGOPS Operating Systems Review, vol. 37, pp. 29–43. ACM (2003)
9. Gregory Giuliani, Nicolas Ray, Anthony Lehmann(2011): Grid- enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities, Future Generation Computer Systems 27 (2011) 292–303
10. Guan, X., & Chen, C. (2014). Using social media data to understand and assess disasters. Natural Hazards, 74(2), 837–850.
11. Karnitis, G., & Arnicans, G. (2015). Migration of Relational Database to Document-Oriented Database: Structure Denormalization and Data Transformation.7th International Conference on Computational Intelligence, Communication Systems and Networks.
12. Lee, J.-G., & Kang, M. Geospatial Big Data: Challenges and Opportunities. Big Data Research, 2(2), 74–81, 2015.
13. LIU, Z., GUO, H., & WANG, C. (2016). Considerations on Geospatial Big Data. IOP Conference Series: Earth and Environmental Science, 46, 012058. doi:10.1088/1755-1315/46/1/012058
14. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D.,Guestrin, C., Hellerstein, J.M.: Graphlab: A new framework for parallel machine learning. arXiv preprint arXiv:1006.4990 (2010)
15. Malewicz, G., Austern, M., Bik, A., Dehnert, J., Horn, I., Leiser, N., Czajkowski, G.: Pregel: a system for largescale graph processing. In: Proceedings of the 2010 international conference on Management of data, pp. 135– 146.ACM (2010)
16. Murray, D.G., Schwarzkopf, M., Smowton, C., Smith, S.,Madhavapeddy, A., Hand, S.: Ciel: a universal execution engine for distributed data-flow computing. In: Proceedings of the 8th USENIX conference on Networked systems design and implementation, p. 9 (2011)
17. Nguyen, T., Larsen, M. E., O'Dea, B., Nguyen, D. T., Yearwood, J., Phung, D., … Christensen, H. (2017). Kernel-based features for predicting population health indices from geocoded social media data. Decision Support Systems, 102, 22– 31. doi:10.1016/j.dss.2017.06.010
18. Ningyu Zhang, Guozhou Zheng, Huajun Chen (2014): HBaseSpatial: a Scalable Spatial Data Storage Based on HBase, 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and

Communications, 978-1-4799- 6513-7/14 $31.00 © 2014 IEEE, DOI 10.1109/TrustCom.2014.83.

19. Ruiz-Medina, M. D. (2012). New challenges in spatial and spatiotemporal functional statistics for high-dimensional data. Spatial Statistics, 1, 82–91.
20. Song, Z., Chen, J., & Ye, J. Y. (2014). A Mobile Storage System for Massive Spatial Data. Advanced Materials Research, 962-965,2730–2734.
21. S. Schade (2015) : BIG DATA BREAKING BARRIERS – FIRST STEPS ON A LONG TRAIL, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-7/W3, 36th International Symposium on Remote Sensing of Environment, 11–15 May 2015, Berlin, Germany
22. Storey, V. C., & Song, I.-Y. (2017). Big data technologies and Management: What conceptual modeling can do. Data & Knowledge Engineering, 108, 50–67.doi:10.1016/j.datak.2017.01.001
23. Sun, D., Yan, H., Gao, S., Liu, X., & Buyya, R. (2017). Rethinking elastic online scheduling of big data streaming applications over high- velocity continuous data streams. The Journal of Supercomputing, 74(2), 615–636.
24. Tobler, W., A Computer model simulating urban growth in the Detroit region Economic geography, 1970.46 : P 234-240
25. Understanding coordinate management in geodatabase (2007): As ESRI white paper June 2007. Working with geodatabase (2012): Powerful multiuser editing and sophisticated data integrity. An ESRI white paper February 2012
26. Zhang, N., Zheng, G., Chen, H., Chen, J., & Chen, X. (2014). HBaseSpatial: A Scalable Spatial Data Storage Based on HBase. 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. doi:10.1109/trustcom.2014.83
27. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications 1(1), 7–18 (2010)
28. Zhang, Y., Gao, Q., Gao, L., Wang, C.: Priter: a distributed framework for prioritized iterative computations. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, p. 13. ACM (2011).
29. Zheng, K., Gu, D., Fang, F., Zhang, M., Zheng, K., & Li, Q. (2017). Data storage optimization strategy in distributed column-oriented database by considering spatial adjacency. Cluster Computing, 20(4), 2833–2844.
30. Yuji Roh, Geon Heo, Steven Euijong Whang, Member, IEEE " A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective" Wireless Communications and Mobile Computing Volume 2018, Article ID 8738613, 19 pages.