Gurtej Singh Ubhi *et al*, International Journal of Computer Science and Mobile Applications,
Vol.5 Issue. 10, October- 2017, pg. 196-203 **ISSN: 2321-8363**
**Impact Factor: 5.515**

# Big Data Landscape: State of the Art, Trend Analysis and Potential Challenges

**Gurtej Singh Ubhi[1], Jaspreet Kaur Sahiwal[2]**

[1]M.Tech Scholar, Lovely Professional University
[2]Assistant Professor, Lovely Professional University
Email – [1] Gurtejubhi@Yahoo.com, [2] Jaspreet.14752@lpu.co.in

***Abstract— With fruition of the computer world and the propagation of the storage devices has led to an explosion of the on hand information , to add further it has been found that a large part of the same data is neither stored or used. There is a growing need for extraction of usable facts from the operational data in the real world which is decisive in fast changing situation. Since, indulgent and treatment Big data is a big dispute This paper surveys different approaches in real-time analytics of Big Data or near real-time in the detailed fields of application, as well as tools and techniques being used.***

***The survey results indicate what technologies have been used in each of the fields of application and what the reason for the choice was along with the extension of future challenges of the same. This is an attempt to contribute to epistemology with novel opinion on typical scenarios for Big Data.***

## 1. Introduction

Today's the human race produces gigantic capacity of statistics and the foundation to amass, figure and relay this information is subjugated and continue to mature. Enormous data offers a great prospect to maneuver and use it in valuable application. However, we face new nominal summons when it comes to manage, organize and process, and analyze this huge amount of data [1].

Businesses and companies can learn more about their situation and recital using Big Data analytics and they can work with better acquaintance to advance the progression of verdict creation and achieving elevated feat [2]. Analyzing vast quantities of data if it is done efficiently can greatly aid in the addressing of problems immediately as well as the introduction of smart ideas.

Advances in far-flung sensing, computation, interactions, and space have shaped huge data anthology. Information getting hold of priceless for science, regime, big business and humanity is been exigent.

To excerpt instances, exploration industries like Yahoo!, Microsoft & Amazon have shaped a lock, stock and barrel new dealing by reasonably gaining knowledge on internet plus constructing it obtainable for additional use. These assemble trillions of bytes of information as a schedule with incessantly adjoin innovative-fangled options like pouring commands, figure repossession, and satellite descriptions [25].

## 2. Research Background

Big Data usage are indubitably solitary of the focal drivers for these developments, yet these options tend to taper our view on the scalability of data processing. In this paper, we investigate realm of Big Data beyond scalability, and suggest a draft model that characterizes Big Data applications by how they generate insights and influence decision making.

The typical dialogue in Big Data inspect is located on the 4V model: velocity, volume, variety and veracity. Velocity and volume are classes of scale and do not let for a qualitative peculiarity of Big Data applications. The same holds more or less true for variety: The only difference in quality that could be argued is between applications that use a variety of data formats and sources and those that do not. Veracity again is a universal criterion that does not allow distinction, unless we consider the contrast between applications where veracity can be verified and where it cannot. This consideration already points out the necessity to shift the discussion from a technological focus to epistemology.

 

**ISSN: 2321-8363**
**Impact Factor: 5.515**

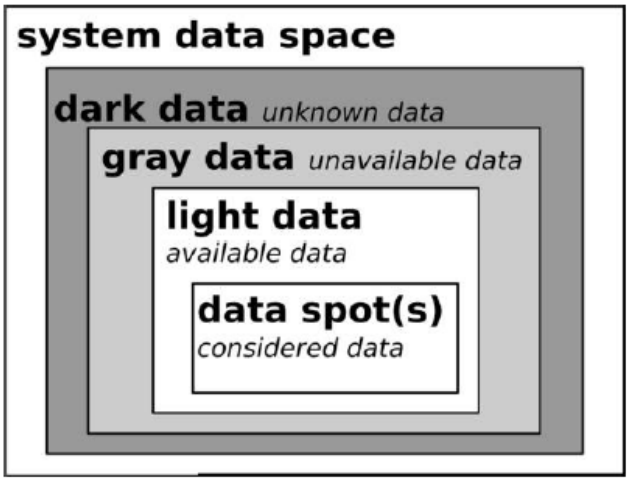


**Figure 1.** Data Spaces in Big Data

The model depicted in Figure 1 assumes, that the next of kin between key component and samples may be unclear in Big Data scenario. It differentiates amid 4 types of data:

• **Light data** is the data that is available; it is set to be allocated and used at any instance.

• **Data spots** are smaller part of the light data so as the measure in the scrutiny.

• **Gray data** is data that is not available to us, but that we can make qualified assumptions about and that we know is part of the system we are analyzing.

• **Dark data** is unknown data that we cannot qualify or quantify in any way.

We further postulate the following distinctiveness:

• Presence of dark data increases uncertainty of a system model

• Presence of gray data decreases the accuracy of a system model

• Selection of data spots out of the light data increases the discovery of small patterns and weak signals.

Big Data differ from customary data in little distinctiveness recognized like the 3 V's: Big Data volume, velocity, and variety.

Moreover, several supplementary eccentricities of Big Data are just came as original V's like value and veracity.

**Volume**: The aspect of digital data in 2011 was probable at 1.8 Zettabytes, plus it seems we have to look forward to covenant with 50 times additional value via year 2020 [3]. The Internet along with smart phones contributes amazingly in producing this data.

**Velocity**: Vast capacity of information are formed and Big Data sets are produced speedily each second. Organizing, using along with dispensation the data as it is unruffled to be incorporated in the verdict creation in factual-instant applications is typically nearly all vital procedural confront [4].

**Variety**: There are a lot of data types like messages, images, and videos, sensors data, production dealings, and fiscal and political news. Composed Big Data beyond the prearranged data includes unstructured data of all varieties [5].

 

**ISSN: 2321-8363**
**Impact Factor: 5.515**

**Value**: There is typically unidentified precious data in the enormous quantity of data placed plus vacant. The trait of Big Data is by using technologies, worth can be collected as of immature data [5].

**Veracity**: The significant facet of Big Data is the eminence of collected data, that can diverge deeply based on precise breakdown.

## 3. Review of Literature

It has been put into a variety of issues related to Big Data. This segment highlights the efforts plus foremost assistance by the researchers under an assortment of areas.

**Jasmin Azemovic and Denis Music [8]** have offered storing methods in support of amorphous information that included:
i) Data inhabit in relational milieu. ii) Data resides faint relational situation was association and iii) Data stored inside amalgam mode.

**Boris et al. [31], [10]** concentrated on the associate executive, s tored into solitary library for querying, innovation and salvage of information and for working out along with compensation. They allege that the metadata stratum plays as faultless viaduct amid formless and prearranged data situation.

**The paper [11]** focuses on issue during streaming with analyzing the streams of data. The ways for e.g. load flaking, variety, sketching and collection are coupled to observe subsets of entire data sets. In chore-like answer, the methods are allocated to update obtainable and devise novel means for competent deployment of gap with point. Charge-based painstaking clustering, classifications, alliance, and deterioration to extort familiarity as of streaming inputs.

**Shao Xiao-Dong and LI Qiang [12]** highlighted 2 key causes in blueprint and growth methods for colossal data broadcast/getting course usage in bona fide-era.

They are i) synchronizing dilemma in transmit/receiving and ii) Receiving/handling progression. Data shield hoard manner is employed for synchronizing in data distribution and message triggering. Supervision device also employs data bulwark cache in getting impediment difficulty.

**Ostrowski et al. [13]** premeditated an agenda to sustain machine learning as of formless data as sources and to amalgamate 2 or further learning device with a customary indexing process to solve issues in index based algorithm. The scaffold has 3 rounds:

**Stage i)** Preprocessing stage to spot all connected subjects.

**Stage ii)** Implementing supervised erudition system using Bayesian erudition algorithm.

**Stage iii)** Put together the overturn indexing strategy on figure of samples.

**The paper [14]** extract as well as classifies amorphous data of webpages and stored into a partially-prearranged database using XML documents. The paper gives comprehensive revision on frequent tools for mining and categorization of data. Deed ObjectModule tree may be employed in cataloging progression and created trial product of the structural design to prove how data removal from web pages.

**J. Chandrika and K.R AnandaKumar [15]** argued the significance of data stream mining class consciousness plus resources implementation. A fresh structure to widespread for any stream mining methods have been wished-for. They projected reserve wakefulness with eminence wakefulness structure.

To mitigate the time in result top-k analogous neighbors **Lien et al. [16**] locate out a novel connection measure on graph model, that premeditated the connection measures as of associations on a graph with vertices being stuff with user.

 

**ISSN: 2321-8363**
**Impact Factor: 5.515**

## 4. Areas of Application

Amid the abundant papers that were analyzed through this survey, a wide assortment of methodologies was used. A vast majority of the papers published in the turf of real-time data analytics make some contribution to exact group in everyday life. This further categorizes each into specific areas of application of the different proposed real-time data analytics methodologies.

### 4.1 Scrutiny

Existing inspection systems often undergo from the extremely slow recognition process. Hua, Jiang, and Feng propose a new method. In a jam-packed area where a child can easily get lost, this new methodology is able to examine millions of images in real-time to locate the missing child [7].Baig and Jabeen introduce a Big Data analytics scheme that monitors student conduct.

### 4.2 Internet of Things

The statistics from dissimilar sources and sensors are becoming more and more obtainable. Public and confidential data can be used to impel innovations and novel solutions to a variety of tribulations. The Internet of Things (IoT) is particularly talented in real time predictive data analytics for effective pronouncement support. Big Data Processing is caught up in our everyday life such as movable plans and wireless sensors, that are aimed at trade by billions of users' interactive data.

### 4.3 Social Media

One more chief and mounting relevance of real-time data analytics is the area of social media. Analyzing post on sites like Facebook plus Twitter can establish quite functional for illustration of the summary and building guesses about activities that transpire in detailed area of the world at sure era [19]. Authors suggest an analytics methodology to apply to Twitter, which is able to mine for patterns and detect outliers based on status updates. Facebook used a real-time data analytics method in the assembly of its messaging application [20]. Societal medium platform can be rather enlightening through a crowd sourcing stance. Nguyen and Jung offer a technique of happening recognition from side to side and the behavioral analysis of Twitter user.

### 4.4. Health Care

Real-time data analytics applications are quickly growing in the area of health care. Analyzing data in real-time can provide constant updates on the well-being of patients' health conditions along with numerous other situations [23, 24].
Sujatha, Devi, Kiran, and Manivannan propose a statistical method that uses real-time data analytics to predict the survival rate and length of stay of patients diagnosed with diabetes [27]. By collecting and analyzing data in real-time, the quality of service of health care's best practices can be better enforced [28]. Currently communication and real-time status updates on patients' health is vital to accurate and immediate treatment of clients. Wang, Qui, and Guo propose a new tele health system that provides real-time information updates [29].

### 4.5 Marketing

Deng, Gao and Vupplapati speak to the dispute by developing Big Data Mobile Marketing analytics and advertising suggestion framework [33]. Their structure ropes both offline and online promotion operation in which the chosen analytics techniques are used to employ in marketing recommendations based on composed Big Data on mobile user's profiles, admission behaviors, and mobility pattern. The project of Deng and Gao provides an authentic-point and static on-demand service for advertisers and publishers. Their scheme demands solutions to analyze the composed big advertising data, discover customers' performance styles along with ascertain, a novel model for publicity counsel [34].

### 4.6 Cyber Sanctuary

With the very fast growth of the Internet, web-based systems are facing malevolent and apprehensive files threatening their security. Mahmood and Afzal review security analytics which can help monitoring streams in real time and detect and thwart the attacks [35].

## 5. Revelation of sound known expertise firm

The appearance of expertise and applications have led to Big Data, which is composite due to the degree of information being shaped each day in terms of terabytes, enterprises talking in stipulations of pet bytes. There is various application presented with industries are also deceitful and budding their own paraphernalia to grip big data predicament. The minority of them are formulated at this juncture.

 

**ISSN: 2321-8363**
**Impact Factor: 5.515**

**A. IBM**
Big Data is innovative epoch in data investigation plus abuse, and IBM has set onward to assist their patrons to cooperate with big data. The IBM has premeditated two Hadoop-based tools for big data: IBM InfoSphere Big Insights which is multitalented plus competent key for conduct and dispensation Internet-level volumes of truth of data. It enhance expertise to nude the ever rising stress of the commerce firms, workflow, and adding organizational, provisioning, as well as safety skin, beside with difficult analytical computational ability from IBM Research. IBM InfoSphere Streams gives incessant calculations taking place with mammoth volumes of data sinuous, with submillisecond reply era [32].

**B. ORACLE**
About 2012 oracle came up with identity-residential catalog with the merchandise name NOSQL for big data appliances. It plants in scattered analogous dispensation background and also ropes new indoctrination add-ons. The majority patrons are paying attention towards Oracle Big data surroundings as of its skill to supervise large scale data. Most of the industries in addition to researchers persuade to towards this means [30].

**C. EMC**
EMC Big Data is developing swift plus ropes real time doling out. Looking the offer data is essential since of it ever mounting temperament. Such as others EMC come awake with big data tool: Greenplum, amid analytical capability. EMC Greenplum Appliance puts Hadoop and SQL (Hive) in similar skin. EMC chains Hadoop based data investigation by a fiscal information storage plus lofty-ability [26].

**D. HP**
HP provides actual-instant big statistics analytics stage, a gratis adaptation tool Vertica. It provide highly developed analytics, consumer responsive boundary version.

## 6. Paraphernalia and Techniques

In this section, some of the central paraphernalia and technology that are life form used in Big Data and valid-instant analytics are being reviewed. Hadoop, as a base for data storage and disseminated dispensation, is the most imperative tool being used in this area, however, for stream processing, Spark is a authoritative tool for in-memory computing which analyze information in genuine-instance. Also, for real-time data analytics there will be a requirement for data intake tools like Kafka, Storm, and Flume which are using patterns to import data in a way that is set to be applied on clusters.

### 6.1 Hadoop

The Apache Hadoop software library allows for the allocation dispensation of large data sets crossways cluster of computers as using simple indoctrination models. It can scale up from single server to countless machines, each of them having their own storage and dispensation.

Here there is no need to depend on hardware to achieve high ease of use. The records itself is calculated to sense and lever failure at the application stratum; it mounts a decidedly-obtainable overhaul on top of a bunch of computers, each of which might be level to failures [36]. Hadoop is a distributed processing framework based on Java.

There are researchers, though, that make apply of merely fraction of Hadoop's functionality. Driscoll, Daugelaite, and Sleator apply MapReduce for the doling out of data except the cloud for data cargo space [38].

There are some that argue that while Hadoop is well-suited for experimentation in data analytics, it is not ideal as aiming for real-time processing [39].

### 6.2 Spark

Spark is a structure for similar dispensation of Big Data. Spark is considered to use the basis of Hadoop Map Reduce with several changes that enables it to execute more competently than Hadoop Map Reduce. Spark has its possess streaming API and independent process for nonstop batch dispensation across unreliable short time intervals. Spark runs up to 100 times faster than Hadoop in certain conditions; but it still uses Hadoop distributed file system.

 

**ISSN: 2321-8363**
**Impact Factor: 5.515**

This is the reason why most of the Big Data projects install Spark on Hadoop so that the superior of Big Data applications can be run on Spark by using the data stored in Hadoop distributed file system. So we can consider Spark as a porch of Hadoop, which has some skin for real-time analytics like being fast, simple, and supportive of applications such as machine learning, stream processing, and graph computation. Xu, Wu, Xu, Zhu, and Bass implement Spark into their idea for real-time data analytics as a tune-up.

It is capable to hold up both stream and group processing while Hadoop is prepared typically for batch dispensation. Spark provides many authentic-era dispensation and evaluation ways that Hadoop unaccompanied cannot. Therefore, to direct the data for their design, they use Spark specially [37].

**6.3 GridGain**

GridGain is based on Java an open resource device for dispensation big data in authentic instant. Grid Gain is attuned by (HDFS) Hadoop Distributed File System and an decision to Hadoop's MapReduce. GridGain gives lofty tempo actual point in time data scrutiny with in-reminiscence data dispensation. Operating Systems: Linux, Windows and OS X.

**6.4 Storm**

Storm is one more real-time working out system. It is a task parallel scattered computing scheme which can dependably process unbounded streams of importing numbers. Storm uses an independent workflow, Directed Acyclic Graphs, in its platform. Storm utilizes Zookeeper, a minion worker to manage its processes, instead of running on Hadoop clusters.

**6.5. Flume**

Data intake gear have a extremely imperative position in genuine-instant analytics. Flume is one of the gear that prepare a dispersed, reliable plus obtainable service for resourcefully importing data.
Collecting, aggregating, and bringing in huge amount of data with its flexible architecture based on streaming data flows, makes it likely for Big Data frameworks to ingest data in a way that makes it trouble-free for dispensation tools to arrive at data.
Flume is solitary of the data dispensation structure that has the skill to be functional to factual-era data analytics act alike to that of Storm [17].

**Makeshwar et al.** projected a structure to collect plus amass vast data as of a feeler network and to scrutinize the established data [18].

They searched the appropriateness of Flume in addition to Mahout to bring high presentation of calculation enhancement of Hadoop Dispersed File Organism.

**Conclusion**

The connotation and necessity of real-time Big Data analytics is exclusively clear toward the management of data generated, and in improvement of technology and in turn the facilitation of everyday life. There is a upward need for admittance to information and to draw conclusions at a rate ideal for society to achieve advantages, and be able to make on time and knowledgeable decisions.
To highlight current impact of realtime data analytics, a literature survey has been conducted. By classifying wished-for real-time data analytics methods by their different areas of application and utilized tools, future researchers and data scientists will have the ability to improve their businesses and technologies with the advantage of data analytics in real-time while using appropriate tools for their circumstances. In addition, researchers will be able to appreciate situations and applicable areas where real-time data analytics have not been taken advantage of, and from there be able to develop methods to benefit other aspects of society.

 

# References

[1] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 1–5.
[2] A. McAfee and E. Brynjolfsson. "Big Data: the management revolution." Harvard business review, Vol. 90, No. 10, pp. 60-68, 2012.
[3] Bakshi, Kapil, "Considerations for Big Data: Architecture and approach", Aerospace Conference, IEEE, 2012, pp. 1-7.
[4] N. Mohamed, J. Al-jaroodi, Real-Time Big Data Analytics: Applications and Challenges. International Conference on High Performance Computing & Simulation (HPCS), 2014
[5] P. Gupta. N, Tyagi. An Approach Towards Big Data-A Review. International Conference on Computing, Communication and Automation (ICCCA2015)
[7] Y. Hua, H. Jiang, and D. Feng. Fast: Near real-time searchable data analytics for the cloud. SC14: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 754–765, Nov 2014.
[8] Jasmin Azemovic and Denis Music, "Efficient model for detection data and data scheme tempering with purpose of valid 85 forensic analysis," Computer Engineering and Applications, Int. Conf. on Computer Engineering and Applications Manila, Philippine, June, 2009.
[10] "News: Live Mint". Are Indian companies making enough sense of Big Data?. Live Mint (2014, Nov) [Online]. Available: http://www.livemint.com/. 2014-06-23.
[11] Ibrahim et al., (2015). "Big data" on cloud computing: Review and open research issues". Information Systems, 2014, pp.98–115. doi:10.1016/j.is.
[12] Shao Xiao-Dong and Li Qiang, "A strategy for continuously big capacity data transmit/receive process handling in real-time," 2nd Int.Conference on Advanced Computer Control (ICACC), 2010 (Volume:5 ).
[13] David et al., "A discrete simulation framework for part replenishment optimization", SIMULTECH, 2013, pp.467-473.
[14] Couldry et al., "Advertising, Big Data, and the Clearance of the Public Realm: Marketers New Approaches to the Content Subsidy". Int. Journal of Communication: 2014, pp.1710–1726,
[15] J. Chandrika and K.R Ananda Kumar, "data stream querying: challenges and issues", Int. Conf. on Computer Applications, ISBN: 978-981-08- 7300-4, 2011.
[16] Lien and Phuong, "Collabotative filtering with a graph-based similarity measure," Int. Conf. on Computing, Management and Telecommunications, 2014, pp.251-256.
[17] C. Wang, I. A. Rayan, and K. Schwan. Faster, larger, easier: reining real-time Big Data processing in cloud. In Proceedings of the Posters and Demo Track, page 4. ACM, 2012.
[18] P.B. Makeshwar, A. Kalra, N.S. Rajput, K.P. Singh, Computational Scalability with Apache Flume and Mahout for Large Scale Round the Clock Analysis of Sensor Network Data, National Conference on Recent Advances in Electronics & Computer Engineering, 2015.
[19] A. Bifet. Mining Big Data in real time. Informatica, 37(1), 2013.
[20] D. Borthakur, J. Gray, J. S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash. Apache Hadoop goes realtime at facebook. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pages 1071–1080. ACM, 2011.
[23] T. B. Murdoch and A. S. Detsky. The inevitable application of Big Data to health care. Jama, 309(13):1351– 1352, 2013.
[24] W. Raghupathi and V. Raghupathi. Big Data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1):1, 2014.
[25] Randal E et al., "Big-Data computing: creating revolutionarybreakthroughs in commerce, science, and society", Computing Community Consortium. (Dec, 2008). Version 8, pp.1-7.
[26] Wang W. K, "A knowledge-based decision support system for measuring the performance of government real estate investment," Expert Systems with Applications, 2005, 29(4), pp.901–912.
[27] V. Sujatha, S. P. Devi, S. V. Kiran, and S. Manivannan. Bigdata analytics on diabetic retinopathy study (drs) on real-time data set identifying survival time and length of stay. Procedia Computer Science, 87:227–232, 2016.
[28] F. A. Batarseh and E. A. Latif. Assessing the quality of service using Big Data analytics: With application to healthcare. Big Data Research, 2015.
[29] J. Wang, M. Qiu, and B. Guo. Enabling real-time information service on telehealth system over cloud-based Big Data platform. Journal of Systems Architecture, 2016.

 

**ISSN: 2321-8363**
**Impact Factor: 5.515**

[30] Tseng & Lin, "Text mining techniques for patent analysis," Information Processing and Management, 2007. 43, pp.1216–1247.
[31] Boris Plejic et al., "Transforming Unstructured Data from Scattered Sources into Knowledge", 2008 IEEE, Int. Symp. on Knowledge Acquisition and Modeling Workshop Proceedings, DOI: 10.1109/KAMW.2008.4810643.
[32] Wang et al., "Self-associated concept mapping for representation, elicitation and inference of knowledge," Knowledge-based Systems, 2008. 21(1), pp.52–61.
[33] L. Deng, J. Gao and C. Vupplapati, Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing, IEEE first international conference on Big Data computing services and applications, 2015.
[34] L. Deng, J. Gao, An Advertising Analytics Framework Using Social Network Big Data, 5th International Conference on Information Science and Technology, 2015.
[35] T. Mahmood, U. Afzal. Security Analytics: Big Data analytics for cyber security. In 2nd National conference on Information Assurance(NCIA), 2013.
[36] http://Hadoop.apache.org.
[37] D. Xu, D. Wu, X. Xu, L. Zhu, and L. Bass. Making real time data analytics available as a service. In Proceedings of the 11thInternational ACM SIGSOFT Conference on Quality of Software Architectures, QoSA '15, pages 73– 82, New York, NY, USA, 2015.ACM.
[38] A. O'Driscoll, J. Daugelaite, and R. D. Sleator. 'Big Data', Hadoop and cloud computing in genomics. Journal of biomedical informatics,46(5):774–781, 2013.
[39] Z. Prekopcsa k, G. Makrai, T. Henk, and C. Gaspar- Papanek. Radoop: Analyzing Big Data with rapidminer and Hadoop. In Proceedings of the 2nd RapidMiner community meeting and conference (RCOMM 2011), pages 865–874. Citeseer, 2011.

 