ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

An Exclusive Survey on Big Data Analytics, Applications and Tools

Renu Dwivedi[1], Prof. Satpal Singh[2], Prof. Sumit Nema[3]

M.Tech. Student, Department of Computer Science Engineering, Global Engineering College, Jabalpur, Madhya Pradesh, India[1]

Assistant Professor, Department of Computer Science Engineering, Global Engineering College, Jabalpur, Madhya Pradesh, India[2]

Assistant Professor and Head of The Department, Department of Computer Science Engineering, Global Engineering College ,Jabalpur, Madhya Pradesh, India[3]

ABSTRACT: In this paper, we review the background and state-of-the-art of big data. We first introduce the general background of big data and review related technologies, such as could computing, Internet of Things, data centers, and Hadoop. We then focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, we introduce the general background, discuss the technical challenges, and review the latest advances. We finally examine the several representative applications of big data, including enterprise management, Internet of Things, online social networks, applications, collective intelligence, and smart grid. These discussions aim to provide a comprehensive overview and big-picture to readers of this exciting area. This survey is concluded with a discussion of open problems and future directions.

KEYWORDS: Real-time analytics; Big Data; Hadoop

I. INTRODUCTION

Huge amount of information generated every day and everywhere. The flow of data arriving in such a fast and complex way needs to be managed, stored, and analysed. Sensors, log data of machines, data storages, public web, social media, business apps, media, archives, and numerous other types of technologies are creating and capturing data continuously in large quantities. Huge data offers a great opportunity to manipulate and use it in beneficial applications. However, we face new technical challenges 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 performance using Big Data analytics and they can manipulate knowledge to upgrade the process of decision making and achieving higher performance [2]. Analysing 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.

This paper aims to present a comprehensive review on real- time Big Data analytics applications and frameworks, and addresses the importance and necessity of real-time data analytics. The paper categorizes areas of application of Big Data and real-time data analytics and then discusses some of the methodologies and tools that are being used in these fields. In addition, it provides some comparison of these methodologies and their best fit based on different requirements of real-time Big Data applications. This is accomplished by investigating numerous research papers and significant contributions to the field and categorizing them.
First, a background of Big Data, data analytics, and real-time processing will be discussed to introduce the concept of data science in recent years. Then the various methodologies that researchers have contributed to various fields of real-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

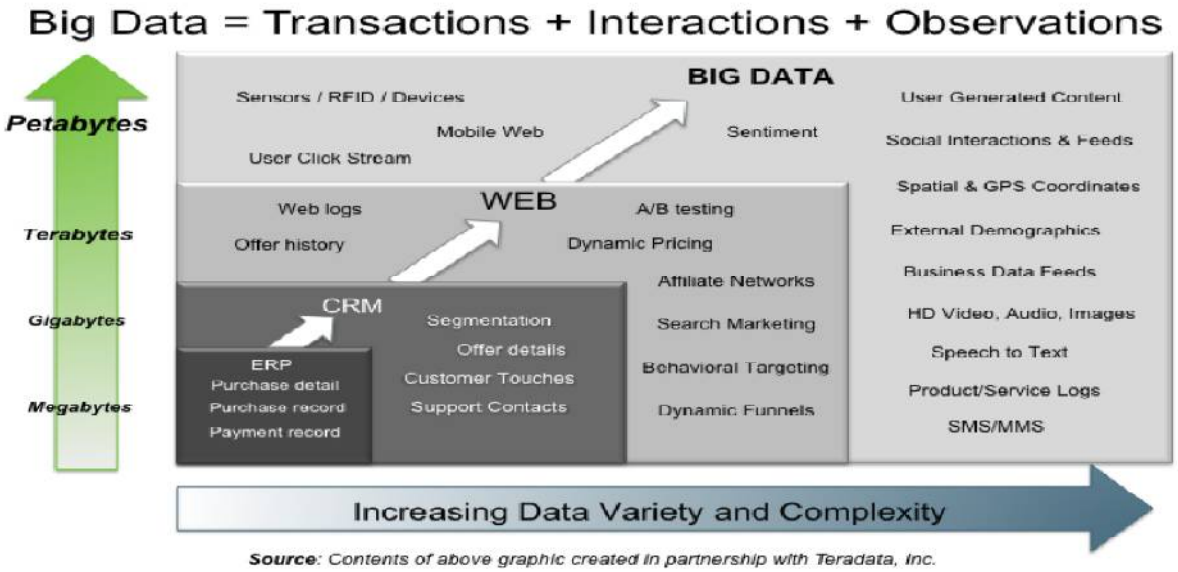
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

time data analytics within the past five years will be addressed, and Section IV presents the tools and technologies that have been used to accomplish their sought out tasks.
Paper is organized as follows. Section II describes automatic text detection using morphological operations, connected component analysis and set of selection or rejection criteria. The flow diagram represents the step of the algorithm. After detection of text, how text region is filled using an Imprinting technique that is given in Section III. Section IV presents experimental results showing results of images tested. Finally, Section V presents conclusion.



Source: *Contents of above graphic created in partnership with Teradata, Inc.*

Paper is organized as follows. Section II describes automatic text detection using morphological operations, connected component analysis and set of selection or rejection criteria. The flow diagram represents the step of the algorithm. After detection of text, how text region is filled using an Inpainting technique that is given in Section III. Section IV presents experimental results showing results of images tested. Finally, Section V presents conclusion.

II. BACKGROUND

Big Data differs from regular data in few characteristics known as the 3 V's: Big Data volume, velocity, and variety. Also, some other characteristics of Big Data are recently introduced as new V's such as value and veracity.

Volume: The size of digital data in 2011 was estimated as 1.8 Zettabytes , and it seems we have to expect to deal with 50 times more information by year 2020 [3]. The Internet and smart phones contributes remarkably in generating this data. Velocity: Huge amounts of data are produced and Big Data sets are generated rapidly every second. Organizing, accessing and processing the data as it is collected to be included in the decision making in real-time applications is usually the most important technical challenge [4].

Variety: There are a lot of data types like messages, images, and videos, sensors data, business transactions, and economic and political news. Collected Big Data beyond the structured data, includes unstructured data of all varieties [5].

Value: There is usually unknown valuable information in the huge amount of data stored and unused. The characteristic of Big Data is by using technologies, value can be extracted from underdeveloped data [5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

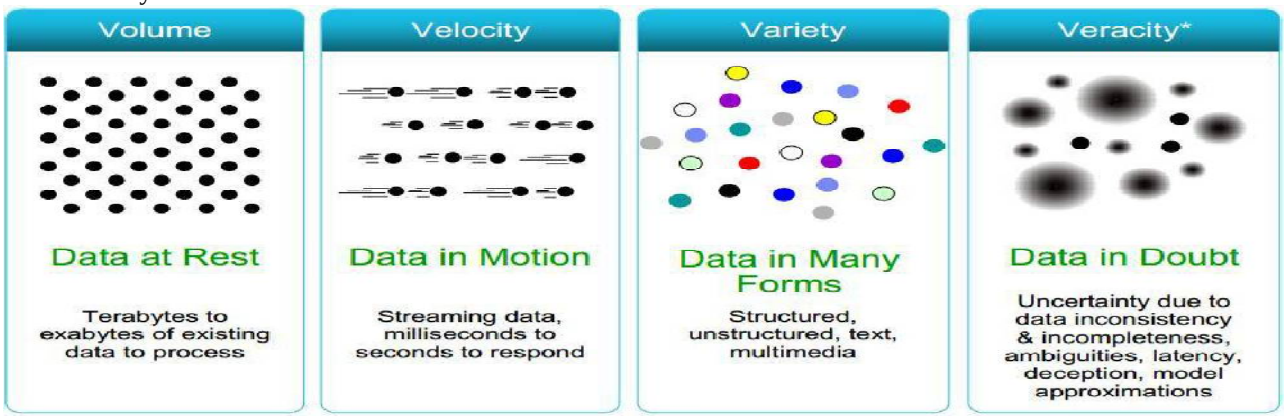
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Veracity: The important aspect of Big Data is the quality of captured data, which can vary greatly depending on accurate analysis.



Challenges in Big Data analytics are that traditional data management methods cannot handle it, and there are difficulties in facing unstructured data. For this reason, NoSQL databases were introduced to handle non structured data. There are various ways to extract information for different types of unstructured data. [6]. Real-time applications rely on instantaneous input and fast analysis to arrive at a decision or action within a short and very specific time line. In many cases, if a decision cannot be made within that timeline, it becomes useless [4]. Originally, data analytics have been performed after storing data on hard disks which eventually have a fair amount of access latency. Dealing with large amount of data makes hard disks not suitable for performing real-time data analytics, especially when most part of data is unstructured. In-memory processing significantly decreases the amount of access latency, this will have a crucial role when real-time analytics is done.

III. APPLICATION

Among the numerous papers accumulated and analysed through this survey, a wide variety of methodologies were utilized. A vast majority of the papers published in the field of real-time data analytics make some contribution to a specific category in daily life. This study categorizes each paper into specific areas of application of the different proposed real- time data analytics methodologies.

A. Surveillance

Existing surveillance systems often suffer from the extremely slow identification process. Hua, Jiang, and Feng propose a new method. In a crowded 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 system that monitors student behavior. From analyzing this data, a conclusion will be made predicting the likelihood of some students being subjected and prone to terroristic ideologies [8]. Shi and Abdel-Aty aim to monitor and reduce crash risk by implementing real-time congestion measurement with the help of Big Data. Statistical techniques are used to identify factors that contribute to congestion and crashes. By analyzing that data in real-time, the work is able to trigger safety warnings at relevant times [9].

B. Internet of Things

The data from different sources and sensors are becoming more and more available. Public and private data can be used to drive innovations and new solutions to various problems. The Internet of Things (IoT) is particularly promising in real- time predictive data analytics for effective decision support. Big Data Processing is involved in our daily life such as mobile devices and wireless sensors, which are aimed at dealing with billions of users' interactive data. At the same time, real-time processing is eagerly needed in integrated systems [10]. The authors have considered Hadoop with

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Flume for a large sensor network being used for a Gas Sensor Application [11]. The data retrieved from web or from IoT sensors can be used for effective decision making. This can happen in the context of smart environments which are developing each day [12].

C. Environment

One of the major applications of real-time data analytics is environmental status analysis. A two-step solution of data management and data analytics has been proposed to discover trends that can predict energy usage and its effects on the environment [13]. Huang, Cai, and Yu propose a distributed neuron network based on machine learning to implement real- time data analytics. [14]. Bilal et al. propose an architecture regarding construction waste management and analytics. Having a data analytics framework to monitor the behavior of waste disposal, the environment can be significantly benefited [15]. Loughalam, Akbarian, and Ulm, by using real-time data analytics, locate parts of the highway where CO2 is most prevalent [16]. A situation that certainly requires real-time data analytics is environmental forecasting whether it is for weather, water quality, greenhouse gas emissions, or any number of other environmental issues. Wang, Zhang, and Babovic aim to improve water quality indicators to qualify the changes that occur in aquatic ecosystems more accurately. The quality of water has numerous determinants and can constantly change as a result, thus real-time data analytics is crucial to the analysis and improvement of the environment [17]. Rathore, Ahmad, Paul, and Daniel propose a framework for analyzing data of land and sea taken from satellites and other sensors [18].

D. Social Media

Another important and growing application of real-time data analytics is the area of social media. Analyzing posts on sites such as Facebook and Twitter can prove quite useful for drawing conclusions and making predictions about activities that occur in specific areas of the world at certain times [19]. Authors propose 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 making of its messaging application [20]. Social media platforms can be quite informative through a crowdsourcing standpoint. Nguyen and Jung offer a method of event detection through the behavioral analysis of Twitter users. By utilizing real-time data analytics on big social data, important events, even emergencies, can be predicted and detected [21]. Preotiuc-Pietro et al. present an architecture for analyzing social media text. By filtering keywords, languages, and more to large datasets of tweets in real-time, data can be processed in numerous more ways to draw more conclusions from more organized data [22]. Jones searches for trends in tweets by use of complex event-processing [23]. Twitter services themselves make use real-time data analytics query suggestion and spelling correction [24].

E. 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 [25, 26]. 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 telehealth system that provides real-time information updates .

F. Business Intelligence

Business applications today are much more data-intensive than they used to be. Extracting meaning from all of the data produced by these applications is one of the most pressing challenges in the business industry. Vera- Baquero, Colomo-Palacios, and Molloy propose a solution that monitors business activities, events, and state changes in real- time . The amount of real-time data analytics methods in business, tells us that it is one of the most prominent fields in Big Data analytics and one that requires significant real-time application . Kolomvatsos, Anagnostopoulos, and Hadjiefthymiades propose an additional mechanism to the query controller. The query is terminated when it has enough data to report a conclusion to the user. This mechanism allows for faster data retrieval and analysis . As

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

decisions support systems become more popular, Demirkan and Delen propose a framework to make DSS and decision making with Big Data in real-time available as a service .

G. Marketing

Deng, Gao and Vupplapati address the challenge by developing Big Data Mobile Marketing analytics and advertising recommendation framework . Their framework supports both offline and online advertising operations in which the selected analytics techniques are used to provide advertising recommendations based on collected Big Data on mobile user's profiles, access behaviors, and mobility patterns. The project of Deng and Gao provides a real-time and static on-demand service for advertisers and publishers. Their project requires solutions to analyze the collected big advertising data, discover customers' behavior patterns, and establish an innovative model for advertising recommendation .

H. Visualization

He, Huynh, and Mong are addressing the problem in visualizing ever changing data in real-time. To introduce the challenge at stake, they propose a tutorial for GPGPU research for processing real-time data analytics . Chopade, Zhan, Roy, and Flurchick propose a visualization architecture for Big Data networks analytics. The visual representation, X-SimViz, is interactive in real-time and fully dynamic and is a simple transfer of information between creator and user . Kitchin writes on a popular controversy among data scientists relating to IoT the smart city. While this is a broad paper discussing pros and cons of having a data sensor-driven city, Kitchin introduces the idea of city dashboards to visualize the area's analyzed data regarding numerous subjects such as environment, traffic, economy, and more. Liu, Jiang, Heer present a method for visualizing Big Data in real-time through imMens. Their method involves data reduction methods such as filtering, random sampling, and binned aggregation .

I. Cybersecurity

With the very fast growth of the Internet, web-based systems are facing malicious and suspicious files threatening their security. Mahmood and Afzal review security analytics which can help monitoring streams in real time and detect and counteract these attacks .

Listed above are the major fields of application for data analytics in real-time. There, however, remain numerous miscellaneous areas in which real-time data analytics can be beneficial. Shi and Abdel-Aty propose that this kind of study is useful for monitoring and managing traffic congestion on urban expressways. By using real-time data analytics, congestion can be immediately detected and solved as quickly as possible [9]. Cha and Wachowicz. propose that analyzing data in real-time requires data ingestion and processing of the stream of data before the data storage step .

IV. TOOLS AND TECHNOLOGY

Part of how Big Data got the distinction as Big is that it became too much for traditional systems to handle. In this part, we are reviewing some of the dominant tools and technologies that are being used in Big Data and real-time analytics. Hadoop, as a base for data storage and distributed processing, is the most important tool being used in this area, however, for stream processing, Spark is a powerful tool for in-memory computing which analyze data in real-time. Also, for real-time data analytics there will be a need for data ingestion tools like Kafka, Storm, and Flume which are using patterns to import data in a way that is ready to be analyzed on clusters.

A. Hadoop

The Apache Hadoop software library allows for the distribution processing of large data sets across clusters of computers while using simple programming models. It can scale up from single server to many machines, each of them having their own storage and processing. Here there is no need to depend on hardware to achieve high-availability. The library itself is designed to detect and handle failures at the application layer; it delivers a highly-available service on top of a cluster of computers, each of which may be prone to failures. Hadoop is a distributed processing framework based on Java effective in data-intensive analytics . All of the methods proposed in [7] can be implemented by utilizing Hadoop for data processing, analytics, and storage. Specifically, they make use of Hadoop for the processing of millions of images but utilizes different storage systems that are searchable.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Hadoop MapReduce is limited to batch processing of one job at a time. Additionally, it offers parallel computing that contributes to high performance and efficiency for large data analytics projects [18]. Xu, Wu, Xu, Zhu, and Bass propose the implementation of real-time analytics as a provided service. The architecture is made up of three service components: a backend training system, a service wrapper for easy machine learning accessibility, and service user interfaces to make implementation of real-time analytics simple for non- programmers . Baig and Jabeen [8] chose to collect, analyze, and store students' behavior with the data that the school collect using Hadoop. While Batarseh and Latif are sure to make use of several health care specific data analytics tools, Hadoop is implemented in their framework to handle storage and querying of health care data received [28].

There are some sources that do not explicitly use Hadoop for their data analytics, but some simply evaluate the performances of their systems using Hadoop framework [21]. Hadoop's MapReduce feature is appealing to multiple researchers aiming to efficiently process large datasets and attain meaningful information from them [23]. Many argue that an optimal solution to a Big Data problem can be achieved through the use of Hadoop and its several provided methods. Patel, Birla, and Nair do this by performing a couple different experiments on different Big Data sets .
Borthakur et al., the creators of Facebook Messaging, utilize Hadoop for their application because of its elasticity, fault isolation, low latency, and consistency semantics [20]. There are researchers, however, that make use of only part Hadoop's functionality. Driscoll, Daugelaite, and Sleator apply MapReduce for the processing of data but the cloud for data storage. There are some that argue that while Hadoop is well-suited for experimentation in data analytics, it is not ideal when aiming for real-time processing. In fact, some of the architects behind Twitter attempted to supply query suggestion and spell checking services with Hadoop's processing, but changed their method because the access latency was too high [24]. Duan et al. introduce a method for real-time image retrieval. An in- memory vocabulary tree makes use of MapReduce for the training and retrieval of images in real-time.

B. Spark

Spark is a framework for parallel processing of Big Data. Spark is designed to use the basis of Hadoop MapReduce with some modifications that enables it to perform more efficiently than Hadoop MapReduce. Spark has its own streaming API and independent processes for continuous batch processing across varying short time intervals. Spark runs up to 100 times faster than Hadoop in certain circumstances, however it still uses Hadoop distributed file system. This is the reason why most of the Big Data projects install Spark on Hadoop so that the advanced Big Data applications can be run on Spark by using the data stored in Hadoop distributed file system. So we can consider Spark as an extension of Hadoop, which has some features 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 service. It is able to support both stream and batch processing while Hadoop is made mostly for batch processing. Spark provides many real-time processing and evaluation options that Hadoop alone cannot. Therefore, to manage the data for their architecture, they utilize Spark specifically . Though Bilal et al. are making use of a graph database, Neo4J, to store datasets, Spark is the graph processing system being used. Their use of Spark will allow them to process the waste data and analyze it efficiently [15]. The research on distributed computing engines shows that Spark has consistent scalability for large datasets . Yan, Huang, and Yi show Spark is scalable to process seismic data with its in-memory computation and data locality features. They have used some seismic data processing algorithms to study the performance and productivity of Spark.

C. Storm

Storm is another real-time computation system. It is a task parallel distributed computing system which can reliably process unbounded streams of importing data. 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. Many of the explored resources make use of Storm with their new contributions to real-time data analytics [10]. Storm, unlike Hadoop alone, can continue to analyze data as it arrives. As Storm is a complex event processing system that has the ability to detect important event occurrences, it is the processing system that Jones utilizes to detect crucial events through the processing of Twitter feeds [23].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

D. Flume

Data ingestion tools have a very important role in real-time analytics. Flume is one of the tools that prepare a distributed, reliable, and available service for efficiently importing data. Collecting, aggregating, and bringing in huge amount of data with its flexible architecture based on streaming data flows, makes it possible for Big Data frameworks to ingest data in a way that makes it easy for processing tools to reach data. Flume is one of the data processing frameworks that has the ability to be applied to real-time data analytics acts similarly to that of Storm . Makeshwar et al. proposed a framework to receive and store huge data from a sensor network and to analyze the received data [11]. They explored the suitability of Flume and Mahout to deliver high performance computational scalability of Hadoop Distributed File System.

E. Kafka

The other powerful tool for data ingestion is Kafka. It is a distributed streaming platform and a message broker project which provides a unified, high-throughput data feeds. Another characteristic of Kafka, which makes it different, is low-latency platform for handling real-time streaming data. It is, in its essence, a scalable message queue which has designed to distribute transaction log making it highly valuable for infrastructures to process streaming data [6]. Ta et al. have used multiple streams of messages that are generated from Kafka's producers and processed at Storm, then stored in a distributed storage NoSQL Cassandra system.

There are other tools being used in data analytics. S4 is a popular framework for in-memory stream processing. It is able to handle data continuously as it arrives without terminating . The sources discussed in this paper make use of common tools, but each framework utilizes the tools in various ways. As an example, Prekopcsak et al. introduce Radoop as a new method for data analytics. Radoop is meant to be a Hadoop and RapidMiner hybrid. It makes it possible to analyze data beyond the boundaries of main memory.

F. Comparing Processing Tools

Hadoop, Spark and Storm are open source processing frameworks and can be used for real-time Big Data analytics. They all provide fault tolerance and scalability and have a simple implementation methodology. Hadoop, Spark and Storm are implemented in JVM based programming languages- Java, Scala and Clojure respectively. However, there are differences between their processing models and their performance. Hadoop MapReduce is best suited for batch processing. For Big Data applications that require real-time options, other platforms must be used. Spark can make use of existing machine learning libraries and process graphs. Thanks to the high performance of Spark, it can be used for both batch processing and real-time processing. Micro-batching is a special kind of batch processing wherein the batch size is orders smaller. Storm is a complete stream processing engine that supports micro-batching. Spark processes in-memory data whereas Hadoop MapReduce persists back to the disk after a map action or a reduce action thereby Hadoop MapReduce lags behind when compared to Spark in this aspect. Spark requires huge memory just like any other database, as it loads the process into the memory and stores it for caching. However, if Spark runs on top of YARN with various other resources demanding services, then there is a possibility of performance deprivation for Spark. In the case of Hadoop MapReduce, the process is killed as soon as the job is completed, making it possible to run along with other resource demanding services with a slight difference in performance. Spark and Storm, both provide fault tolerance and scalability but differ in the processing model. Spark streams events in small batches that come in short time window before it processes them, whereas Storm processes the events one at a time. Thus, Spark has a latency of few seconds, but can be used in stateful computations to ensure that the event is just processed once, whereas Storm processes an event with just millisecond latency without data loss. Spark has good performance on dedicated clusters when the entire data can fit in the memory, whereas Hadoop can perform well along other services when data does not fit in memory.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Table I categorizes different studies based on the technology they used and area of application of Big Data analytics of each of them.

	Tools					
	Hadoop	Spark	Storm	Kafka	Flume	Other
Surveillance	[8],[7]					[9]
Visualization	[40]					[40],[42]
Environment	[18]	[15]				
Social media	[20],[44],[24],[21],[22]	[54]	[44],[23],[54]	[6]	[52]	[54]
Health care	[28],[47]		[53]	[53]		
Business Intelligence	[30],[32],[36]					[35]
Marketing	[37]	[37]				[38]
Cyber security	[43]					
Internet of things	[11]		[10]		[11]	
General	[49],[48],[58]	[58]				[49]

V. CONCLUSION

The significance and necessity of real-time Big Data analytics is completely clear to the manipulation of data generated, and in improvement of technology and in turn the facilitation of everyday life. There is a growing need for access 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 real- time data analytics, a literature survey has been conducted. By classifying proposed 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 benefit of data analytics in real-time while using appropriate tools for their circumstances. Furthermore, researchers will be able to understand 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)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

[6] Datadog Engineering Blog. Monitoring Kafka performance metrics. 23 May 2016.
[7] Apache Hadoop [Online]. Available: http://hadoop.apache.org/
[8] Chansler, R., Kuang, H., Radia, S., Shvachko, K. "The Hadoop Distributed File System," in Proc. IEEE Conf. Mass Storage Systems and Technologies (MSST), Incline Village, NV, 2010, pp. 1 – 10
[9] Dean, J., Ghemawat, S. "MapReduce: Simplified DataProcessing on Large Clusters," Mag. Commun. ACM 50thanniversary, vol. 51, issue 1, 2008, pp.107-113
[10] HIVE-1644 [Online]. Available:https://issues.apache.org/jira/browse/HIVE-1644
[11] HIVE-1694[Online]. Available:https://issues.apache.org/jira/browse/HIVE-1694
[12] An, M., Wang, W., Wang, Y., "Using Index in the MapReduce Framework, ", 12th Intl. Asia Pacific Web Conf. (APWEB),Beijing, China, 2010, pp. 52-58
[13] Antony, S., Chakka, P., Jain, N., J., Liu, Murthy, R., Sarma, J.S., Thusoo, A., Zhang, N "Hive – A Petabyte Scale DataWarehouse Using Hadoop," IEEE 26th Intl. Conf. DataEngineering (ICDE), Long Beach, CA, 2010, pp. 996 – 1005
[14] H. Huang, Y. Cai, and H. Yu. Distributed-neuronnetwork based machine learning on smart-gateway net- work towards real-time indoor data analytics. 2016 De- sign, Automation & Test in Europe Conference & Exhibition (DATE), pages 720–725, 2016.
[15] M. Bilal, L. O. Oyedele, O. O. Akinade, S. O. Ajayi, H. A. Alaka, H.A. Owolabi, J. Qadir, M. Pasha, and S. A. Bello. Big Data architecture for construction waste analytics (cwa): A conceptual framework. Journal of Building Engineering, 6:144–156, 2016.
[16] A. Louhghalam, M. Akbarian, and F.-J. Ulm. Carbon management of infrastructure performance: Integrated Big Data analytics and pavement-vehicle-interactions. Journal of Cleaner Production, 2016.
[17] X. Wang, J. Zhang, and V. Babovic. Improving real- time forecasting of water quality indicators with combination of process-based models and data assimilation technique. Ecological Indicators, 66:428–439,
2016.
[18] M. M. Rathore, A. Ahmad, A. Paul, and A. Daniel. Hadoop based real-time Big Data architecture for remote sensing earth observatory system. In 2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pages 1–7. IEEE, 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.
[21] D. T. Nguyen and J. E. Jung. Real-time event detection for online behavioral analysis of big social data. Future Generation Computer Systems, 2016.
[22] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M.Niranjan. Trendminer: An architecture for real time analysis of social media text. Proceedings of the workshop on real-time analysis and mining of social streams, 2012.
[23] M. T. Jones. Process real-time Big Data with twitter Storm. IBM Technical Library, 2013.
[24] G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin. Fast data in the era of Big Data: Twitter's real-time related query suggestion architecture. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pages 1147–1158. ACM, 2013.
[25] T. B. Murdoch and A. S. Detsky. The inevitable application of Big Data to health care. Jama, 309(13):1351– 1352, 2013.
[26] W. Raghupathi and V. Raghupathi. Big Data analytics in healthcare:promise and potential. Health Information Science and Systems,2(1):1, 2014.
[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.