Scientific Journal of Impact Factor (SJIF): 5.71
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

Cloud-Based Big Data Analytics

Dr. Dineshkumar B. Vaghela[1], Prof. Arvind D. Meniya[2]

[1](Information Technology Department, Shantilal Shah Engineering College Bhavnagar, Gujarat)
[2](Information Technology Department, Shantilal Shah Engineering College Bhavnagar, Gujarat)

ABSTRACT: *The continual increment in the volume and detail of information gathered by organizations has created immense stream of information in either organized or unstructured arrangement, which is known as Big Data. Cloud computing provides a good platform for big data storage, processing and analysis. It is latest innovation to perform huge scale and complex computing distributing the need to keep up costly processing equipment, space, and programming. Few issues should be solved before this model can be famously utilized. This paper investigates the current research, challenges, open issues and future research heading for this field of study.*

KEYWORDS - *Big Data, Cloud Computing, Big Data Analytics, Big Data Analytics in Cloud Environment, Cloud-based Big Data Analytics*

I. INTRODUCTION

As the dawn of digital age, data being generated, stores and shared are increased day by day. These data comes from various sources like data warehouses, web pages and blogs to audio/video streams, social media etc. The result of this explosion is the generation of massive amounts of complex data. This data should be efficiently created, stored, shared and analyzed to extract useful information. The requirement of an efficiently and effective analytics service, applications, programming tools and frameworks has given birth to the concept of Big Data Processing and Analytics.

There are many applications of Big Data Analytics like medical research, solutions for the transportation and logistics sector, global security and prediction and management of is-sues concerning the socio-economic and environmental sector, scientific research [1].

Applications like Facebook, LinkedIn, Twitter, Amazon, eBay and Google+ requires storage and processing of data in the cloud environment. Moreover, the data mining algorithms used for Big Data analytics require high performance processors. All of the above mentioned requirements can be fulfilled using cloud.

A. Characteristics of Big Data [15]

Big data means not only just the size of data. It covers other characteristics as follows:

1) Volume : It is the amount of data generated from different sources
2) Variety: It specifies that Big Data can be of any type either structured or unstructured like text, audio, video, log files etc.
3) Velocity: It is referred as the speed of generation or transfer of data.
4) Veracity: Establishing trust and reliability in Big Data can be referred as veracity.

The cloud computing environment offers development, installation and implementation of software and data applications 'as a service' [16]. Cloud providers usually offer three different services: Infrastructure as a Service (IaaS); Platform as a Service (PaaS); and Software as a Service (SaaS), which are the basic cloud services:

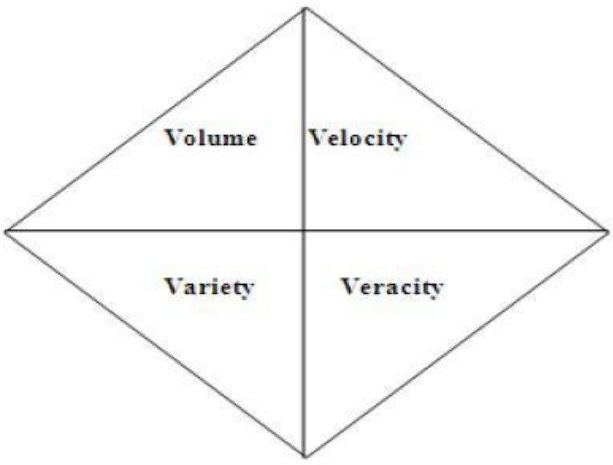

Fig 1. Big Data Characteristics [15]

@IJAERD-2018, All rights Reserved

II. CLOUD COMPUTING

- IaaS provides infrastructure, which means storage, processing power, and virtual machines. The cloud provider satisfies the needs of the client by virtualizing resources according to the need.

- PaaS is built on top of IaaS, it allows users to deploy cloud applications created using the programming and run-time environments supported by the provider. It is at this level where big data DBMS are implemented.
- SaaS consists of applications running directly in the cloud provider [20]. It is the most popular model of cloud services.

These three basic services are closely related: SaaS is developed over PaaS and ultimately PaaS is built atop of IaaS. One more service is DaaS. It is the Data as a Service which provides the public data sets to perform analytics. The cost of storage has remarkably reduced with the use of cloud. Moreover, the 'pay-as-you-go' model permit absorbing and convenient handling of extensive data, offering rise to the idea of big data as a service. A case of one such stage is Google Big Query. It provides the Big Data in the Cloud environment [3].

Cloud computing and Big Data are adjoined. Big Data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner. Cloud computing provides the underlying engine through the use of Hadoop, a class of distributed data-processing platforms [6]. Hadoop is an open-source programming structure for processing and storing big data information in an appropriated style on huge groups of item equipment.

The Hadoop contains a wide range of tools. Two of them are centre parts of Hadoop:

1) Hadoop Distributed File System (HDFS) is a virtual document framework that resembles some other record framework with the exception of than when you move a record on HDFS, this document is part into numerous little documents, each of those records is reproduced and put away on (ordinarily, might be altered) three servers for adaptation to internal failure requirements.

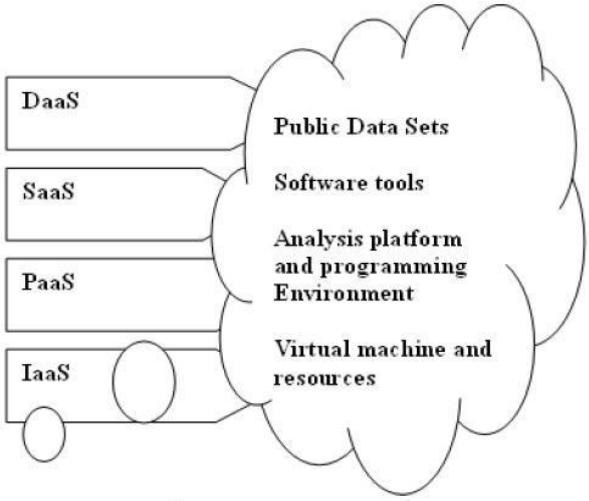

Fig 2. Cloud Services for Big Data [15]

2) Hadoop MapReduce is an approach to part every solicitation into littler solicitations which are sent to numerous little servers, permitting a really adaptable utilization of CPU power [18].

There are not many hands-on applications of big data analytics that utilizes the cloud. This has headed to an increasing shift of research emphasis towards cloud-based big data analytics. An issue that is evident in this arrangement is information security and data privacy.

III. BIG DATA ANALYTICS IN CLOUD ENVIRONMENT

Big Data Analytics could not be performed on the traditional data management tools or data mining techniques because of the large volume and capacity of the datasets.

Artificial intelligence-based algorithms were developed for data mining in 1980s. Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand and Steinberg [3] mention the ten most influential data mining algorithms k-means, C4.5, Apriori, Expectation Maximization (EM), Page Rank, SVM, Ada Boost, CART, Naïve Bayes and kNN (k-nearest neighbors). Most of these algorithms have been used commercially as well. Alam and Shakil [4] propose architecture for management of data through cloud techniques.

Apache's Hadoop Distributed File System (HDFS) is emerging as a widespread programming fragment for cloud computing joined alongside incorporating parts, for example, Map Reduce. Hadoop and cloud computing has a

@IJAERD-2018, All rights Reserved

significant cost benefits. It is also helping faster and better decision making as the future depends mostly on data driven decisions and also a lot of space in improvement either with new applications or services [18].

For processing data on the cluster of computers the most popular model used is MapReduce. Jackson, Vijayakumar, Quadir and Bharathi [5] provide a survey on the programming models that support big data analytics. It identifies Map-Reduce/Hadoop as the most productive model for Big Data Analytics. The frameworks are used for storing and processing of data. In order to store this data, which may be of any structure databases like HBase, BigTable and Hadoop DB may be used. When it comes to data processing, the Pig and Hive technologies can be used.

Map-Reduce accelerate the processing of large amounts of data in cloud; thus the Map-Reduce is the preferred computation model of cloud providers. It is the popular platform of the cloud computing framework that robotically performs scalable distributed applications and provides an interface that allows for parallelization and distributed computing in a cluster of servers [6].

IV. RELATED WORK

Data stored in a cloud-based database can support businesses with their decision-making processes. Using cloud-based big data, analysts have more data to work with and also have the processing power to handle large numbers of records with many attributes which increases predictability. The combination of big data and cloud computing also lets analysts discover new behavioral data such as websites visited or location on a daily basis.

Research efforts have been made to create a big data management framework for the cloud. Khan, Naqvi, Alam and Rizvi propose a data model and provides a schema for big data in the cloud and attempts to ease the process of querying data for theuser.

Balachandran, Bala M., and Shivika Prasad [16] presents Cloud-based big data analytics service model. In this model elements of the big data analytics process are provided through a public or private cloud. Cloud-based data analytics applications and services are typically offered under a subscription-based or utility (pay-per-use) pricing model. This service model is called Cloud Analytics as a Service (CL AaaS) [19]. In this model, analytics is voluntarily accessible through a cloud computing platform. Such cloud-based data analytics service will facilitate businesses to automate processes on an anytime, anywhere basis.

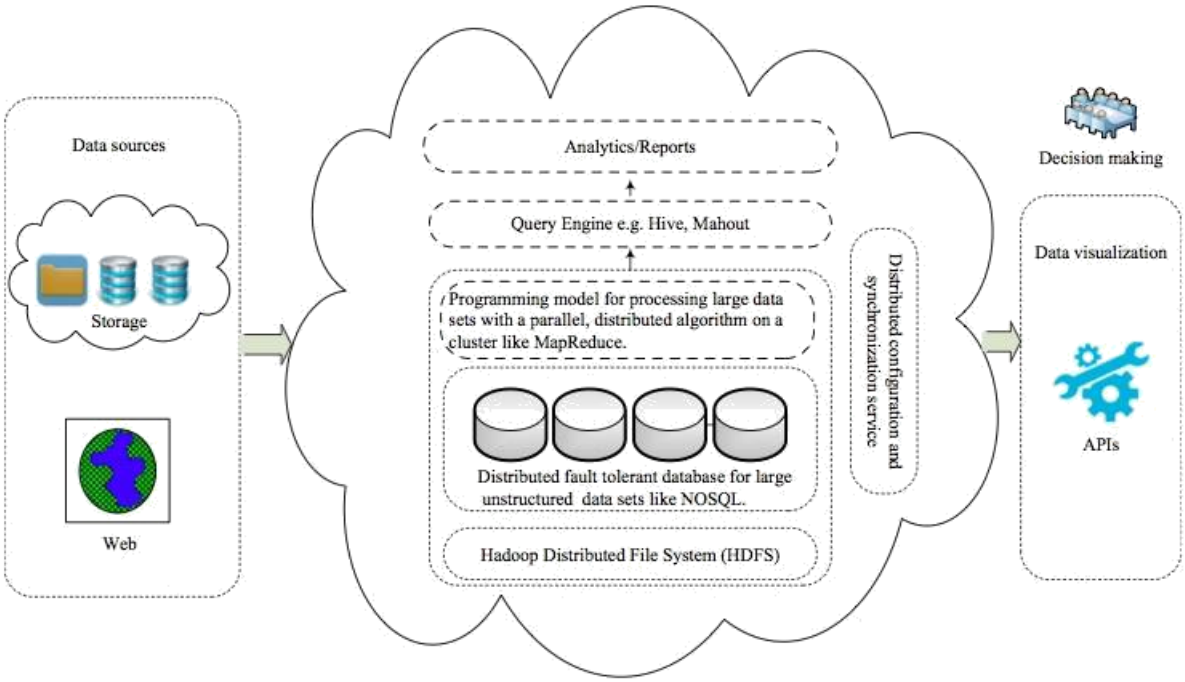

Fig 3. Big Data Analytics with Cloud [20]

Jain, Vinay Kumar, and Shishir Kumar [15] says that Cloud computing provides the support for Big Data deployment. Bandwidth and integration issues can become major obstacles for the use of Big Data on the cloud. They also have listed the various problems with Big Data like security, privacy, widespread deployment. They have discussed the various techniques of computation of Big Data in cloud environment along with their benefits.

Real time Big Data Analysis has gathered the attention of the research community. Many commercial cloud service providers are providing solutions for real-time analysis. AWS based-solutions for real-time stream processing is AWS Kinesis [9] Apache S4 [10], IBM InfoSphere Streams [11] and Storm [12] are some of the frameworks.

@IJAERD-2018, All rights Reserved

A G-Hadoop based security framework is suggested by Zhao, Wang, Tao, Chen, Sun, Ranjan, Kolodziej, Streit and Georgakopoulos [13], which makes use of solutions like SSL and public key cryptography for ensuring security of big data resident on distributed cloud data centers. This framework is used to simplify the processes of submitting job and authenticating users. Talia [14] suggests further research and development in the areas like programming abstracts or scalable high-level models and tools, solutions for data and computing inter-operability issues, integration of different big data analytics frameworks and techniques for mining provenance data.

V. CONCLUSION

With data increasing on a daily base, big data systems and analytic tools have become a major strength of innovation that provides a way to store, process and get information over peta byte datasets. Cloud environments strongly control big data solutions by providing fault-tolerant, scalable and available environments to big data systems.

Considering the rate at which data is being created in the digital world, big data analytics and analysis have become all the more relevant. Moreover, most of this data is already on the cloud. Therefore, shifting big data analytics to the cloud framework is a viable option. While big data systems are powerful systems that enable both enterprises and science to get insights over data, there are some concerns that need further exploration.

In cloud-based big data analytics, challenges like selection and implementation of effective big data solutions using cloud architecture and alleviating the security and privacy risks also exist.

VI. FUTURE SCOPE

One of the greatest concerns while using big data analytics and cloud computing in an integrated model is security. This is the reason why this aspect of cloud-based big data analytics and its practical usage and implementation has pulled in huge consideration.

Additional effort must be employed in developing security mechanisms and standardizing data types. Security and Privacy can be resolved using data encryption. The system must ensure that data is accessed quickly and that encryption does not affect processing times so badly. New and secure QoS (quality of service) based data uploading mechanisms can be used to ease data uploading onto the cloud. The major concern relies upon developing fully automatic reactive and proactive systems that are capable of dealing with load requirements automatically.

The fundamental reason why cloud-based analytics are such a big thing is their easy accessibility, cost effectiveness and ease of setting up and testing. Cloud-based big data analytics solutions must provide a provision for the availability of these data analytics on the cloud so that cost effective and efficient services can be provided.

Some of the main research directions include evolution of analytics and information management with respect to cloud-based analytics, alteration and advancement of techniques and strategies to improve efficiency, analysis and adaptation of legal and ethical practices with respect to the changing viewpoint, impact and effects of technological advances.

REFERENCES

[1] Chen, C. L. P. and Zhang, C. Y. , Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275 (2014) 314-347.

[2] Google Cloud Platform. (n.d.). Big Query. Retrieved from:https://cloud.google.com/bigquery/

[3] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A, Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D. (2008) , Top 10 algorithms in Data Mining. Knowl Inf Syst. 14:1-37. DOI: 10.1007/s10115-007-0114-2. Retrieved from: http://www.cs.umd.edu/ samir/498/10Algorithms-08.pdf.

[4] Alam, M., Shakil, K. A. (2013), Cloud Database Management System Architecture. UACEE International Journal of Computer Science and its Applications. 3(1), 27-31.

[5] Jackson, J. C., Vijayakumar, V., Quadir, M. A. and Bharathi, C., Survey on Programming Models and Environments for Cluster, Cloud and Grid Computing that defends Big Data. 2nd International Symposium on Big Data and Cloud Computing (ISBCC 2015) procedia Computer Science 50 (2015) 517-523.

[6] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. and Khan, S. U., The rise of âĂIJbig dataâĂİ on cloud com-puting: Review and open research issues. Information Systems 47 (2015) 98-115.

[7] Khan, I., Naqvi, S.K. Alam, M. Rizvi, S.N.A. Data model for Big Data in cloud environment. Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference. pp. 582-585.

@IJAERD-2018, All rights Reserved

[8] Liu, C., Yang, C., Zhang, X. and Chen, J. External Integrity Verification for Outsourced Big Data in cloud and IoT: A Big Picture. Future Generation Computer System 49 (2015), pp. 58-67.

[9] Amazon Kinesis. (n.d.). Developer Resources. Retrieved from: http://aws.amazon.com /kinesis/developer-resources/.

[10] Apache S4. (n.d.). Distributed Stream Computing Platform. Retrieved from: http:// incubator.apache.org/s4/

[11] IBM InfoSphere Streams. (n.d.). InfoSphere Streams. Retrieved from: http:// www.ibm. com/ software/products/en/ infosphere-streams.

[12] Storm. (n.d.). Apache Storm: Distributed and fault tolerant real time computation.Retrieved from:http://storm.incubator.apache.org.

[13] Zhao, J., Wang, L., Tao, J., Chen, J., Sun, W., Ranjan, R., KoÅĆodziej, J., Streit, A. and Georgakopoulos, D. A security framework in G-Hadoop for big data computing across distributed Cloud data centers. Journal of Computer and System Sciences 80 (2014), 994-1007.

[14] Talia, D.Clouds for Scalable Big Data Analytics. Published by IEEE Computer Society. (2013)Retrieved from: http://scholar.google.co.in

[15] Jain, Vinay Kumar, and Shishir Kumar Big Data Analytic Using Cloud Computing. Advances in Computing and Communica-tion Engineering (ICACCE), Second International Conference on IEEE, 2015.

[16] Balachandran, Bala M., and Shivika Prasad. Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Busi-ness Intelligence. Procedia Computer Science 112 (2017) 1112-1122.

[17] Khan, Samiya, Kashish Ara Shakil, and Mansaf Alam. Cloud-Based Big Data Analytics - A Survey of Current Research and Future Directions. Big Data Analytics. Springer, Singapore 595-604.

[18] Yunus Yetis, Ruthvik Goud Sara, Berat A. Erol, Halid Kaplan, Abdurrahman Akuzum and Mo Jamshidi Ph.D. Application of Big Data Analytics via Cloud Computing.

[19] Zulkernine, F. Bauer, M. and Aboulnaga,A. Towards Cloud-based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud. 2013 IEEE International Congress on Big Data.

[20] Neves, Pedro Caldeira, et al Big Data in Cloud Computing: features and issues. 2013 IEEE International Congress on Big Data.

@IJAERD-2018, All rights Reserved