Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

Emergence of Cloud Computing Technologies in Big Data: Challenges and Future Perspectives

Ritu [1st]
[1st] M. Tech. Scholar,
Department of Computer Science and Engineering
Bhiwani Institute of Technology and Science, Bhiwani
Haryana, India

Ashok [2nd]
[2nd] Assistant Professor,
Department of Computer Science and Engineering
Bhiwani Institute of Technology and Science, Bhiwani
Haryana, India

Abstract: **The amount of data generated by devices and other internet based sources regularly is huge, which is called big data. This data can be processed and analyzed to develop useful applications for specific domains. Several mathematical and data analytics techniques have found use in this sphere. This has given rise to the development of computing models and tools for big data computing. However, the storage and processing requirements are overwhelming for traditional systems and technologies. Therefore, there is a need for infrastructures that can adjust the storage and processing capability in accordance with the changing data dimensions. Cloud Computing serves as a potential solution to this problem. However, big data computing in the cloud has its own set of challenges and research issues. This chapter surveys the big data concept, discusses the mathematical and data analytics techniques that can be used for big data and gives taxonomy of the existing tools, frameworks and platforms available for different big data computing models. Besides this, it also evaluates the viability of cloud-based big data computing, examines existing challenges and opportunities, and provides future research directions in this field.**

INTRODUCTION*:*

The advent of Internet and rapidly increasing popularity of mobile and sensor technologies have led to an outburst of data in the systems and web world. This data explosion has posed several challenges to systems, traditionally used for data storage and processing. In fact, the challenges are so grave that it would not be wrong to state that traditional systems can no longer fulfill the growing needs of data-intensive computing.The two main requirements of big data analytics solutions are (1) scalable storage that can accommodate the growing data (2) high processing ability that can run complex analytical tasks in finite and allowable time. Among many others, the Cloud Computing technology is considered an apt solution to the requirements of big data analytics solutions considering the scalable, flexible and elastic resources that it offers (Philip Chen and Zhang 2014). Firstly, the Cloud offers commodity machines that provide scalable yet cost-effective storage solutions. Besides this, the processing ability of the system can be improved by adding more systems dynamically to the cluster. Therefore, the flexibility and elasticity of the Cloud are favorable characteristics for big data computing.Cloud-based big data analytics technology finds a place in future networks owing to the innumerable 'traditionally unmanageable abilities and services' that this technology offers. The general definition of future networks describe it as a network that possesses the capabilities to provide services and facilities that existing network technologies are unable to deliver. Therefore, a component network, an enhanced version of an existing network or a federation of new and existing networks that fulfill the above-mentioned requirements can be referred to as future networks.This technology finds applications in diverse fields and areas.

2.0 DEFINING BIG DATA

Several definitions for big data exist owing to the varied perspectives and perceptions with which this concept is viewed and understood.The most accepted definition of big data describes it as huge volumes of exponentially growing, heterogeneous data. Doug Laney of Gartner, in the form of the 3V Model, gave the first definition of big data (Gartner 2016). The fundamental big data characteristics included in this classification are volume, variety and velocity. However, big data, as a technology, picked up recently after the availability of open source technologies like NoSQL (Salminen 2012) and Hadoop[1], which have proven to be effective and efficient solutions for big data storage and processing. Accordingly, the definition of big data was modified to data that cannot be stored, managed and processed by traditional systems and technologies.Besides the above mentioned, there have been several other viewpoints and perspectives on big data. Scholars like Matt Aslett sees big data as an opportunity to explore the potential of data that was previously ignored due to limited capabilities of traditional systems while some others call it a new term for old applications and technologies like Business Intelligence (Abu-Mostafa, Magdon-Ismail, and Lin 2012). Regardless of the big data definition one chooses to follow, nothing can take away the fact that big data opens doors to unlimited opportunities and in order to make use of this reserve, we need to develop new or modify the existing tools and technologies.

2.1 Multi-V Model

The initial 3-V model, given by Doug Laney in the year 2001, includes volume, variety and velocity, as the three fundamental big data characteristics (Gartner 2016). The amount of data included in a dataset indicates the volume of data, which is the most obvious characteristic of big data. The data concerned may come from different sources and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

can be of diverse types. For instance, data coming from a social media portal includes textual data, audio and video clips and metadata. In order to accommodate for these different types of data, big data is said to include structured data, semi-structured and unstructured data. Lastly, data may be produced in batches, near time or real-time. The speed at which the concerned data is being generated denotes the velocity characteristic of big data. Figure 1 depicts the scope of interest for the three V's mentioned above.

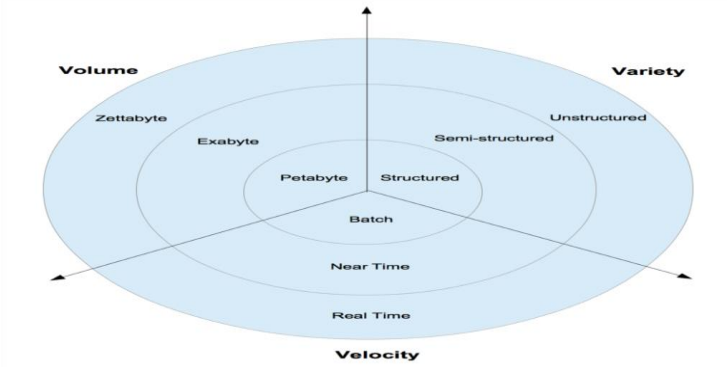

Fig. 1 – Volume, Variety and Velocity of Big Data

2.2 HACE Theorem

HACE stands for Heterogeneous, Autonomous, Complex and Evolving (Xindong et al. 2014) and describes big data as a large volume of heterogeneous data that comes from autonomous sources. These sources are distributed in nature and the control is essentially decentralized. This data can be used for exploration of complex and evolving relationships. These characteristics make identification and extraction of useful information from this data, excessively challenging.

3.0 BIG DATA LIFECYCLE

The lifecycle of big data includes several phases, which include data generation, acquisition, storage and processing of data. These four phases have been explained below and illustrated in Fig. 2.

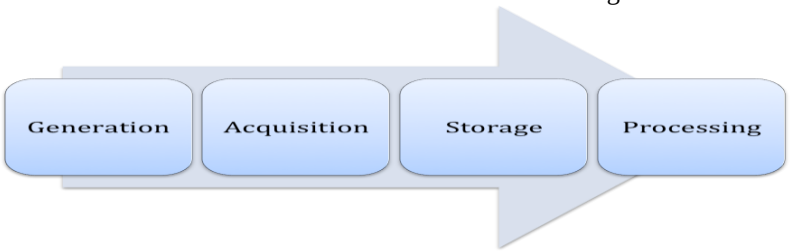

Fig. 2 – Big Data Lifecycle

1 Generation

Data is the most rapidly increasing resource in the world. Perhaps, the reason for this staggering rise in its generation is the diverse types of devices, entities and systems involved. With the rapid advancement in technology, devices like sensors, online portals, social networking websites and online systems like online trading and banking, in addition to many others, have come into existence. All these systems, portals and devices generate data at a periodic basis, contributing to the volume, variety and velocity of big data.

2 Acquisition

Now that we know that big data is being generated by diverse sources, this data needs to be acquired by big data systems for analysis. Therefore, during this stage of the big data lifecycle, the raw data generated in the world is collected and given to the next stage for further processing.

3 Storage

From the first distributed file system, Google File System (Ghemawat, Gobioff, and Leung 2003), to Hadoop Distributed File System or HDFS (Shvachko et al. 2010), there is a range of solutions available in this category. One of the latest and most popular additions to this category is NoSQL database solution like MongoDB[2] and platforms like Cassandra[3].

4 Processing

Like traditional data analysis, the objective of big data analysis is extraction of useful information from the available data. Common methods used for this purpose include clustering, classification and data analysis techniques, besides many others.

4.0 TECHNIQUES FOR BIG DATA PROCESSING

An elaborative description of the techniques used for big data processing has been given below.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

4.1 Mathematical Analysis Techniques

4.1.1 Mathematical Techniques

Factor analysis is mostly used for analysis of relationships between different elements that constitute big data. As a result, it can be used for revealing the most important information. Taking the relationships analysis a step further, correlation analysis can be used for extracting strong and weak dependencies (Ginsberg et al. 2009).

4.1.2 Statistical Methods

Statistical methods are mathematical techniques that are used for collection, organization and interpretation of data. Therefore, they are commonly used for studying causal relationships and co-relationships. It is also the preferred category of techniques used for deriving numerical descriptions. With that said, the standard techniques cannot be directly implemented for big data. In order to adapt the classical techniques for big data usage, parallelization has been attempted.

4.1.3 Optimization Methods

Core fields of study like physics, biology and economics involves a lot of quantitative problems. In order to solve these problems, optimization methods are used. Some of these methods that have found wide-ranging use, because of the ease with which they can be parallelized include Genetic Algorithm, Simulated Annealing, Quantum Annealing and Adaptive Simulated Annealing (Sahimi and Hamzehpour 2010).

4.2 Data Analytics Techniques

4.2.1 Data Mining

Data mining allows extraction of useful information from raw datasets and visualization of the same in a manner that is helpful for making decisions. Commonly used data mining techniques include classification, regression analysis, clustering, machine learning and outlier detection. In order to analyze different variables and how they are dependent on one another, regression analysis may be used.

4.2.2 Machine Learning

A sub-field of artificial intelligence, machine learning allows systems to learn and evolve using empirical data. As a result, intelligent decision-making is fundamental to any system that implements machine learning. However, in the big data context, standard machine learning algorithms need to be scaled up to cope with big data requirements. Several Hadoop-based frameworks like Mahout are available for scaling up machine learning algorithms.

4.2.3 Signal Processing

The introduction of Internet and mobile technologies has made the use of social networking portals and devices like mobiles and sensors, excessively common. As a result, data is being generated at a never-seen-before rate. The massive scale of data available, presence of anomalies, need for real-time analytics and relevance of distributed systems gives rise to several signal processing opportunities in big data (Bo-Wei, Ji, and Rho 2016).

4.2.4 Neural Networks

Image analysis and pattern recognition are established applications of Artificial Neural Networks (ANN). It is a well-known fact that as the number of nodes increase; the accuracy of the result gets better. However, the increase in node number elevates the complexity of the neural network, both in terms of memory consumption and computing requirements. In order to combat these challenges, the neural network needs to be scaled using distributed and parallel methods (Mikolov et al. 2011). Parallel training implementation techniques can be used with deep learning for processing big data.

4.2.5 Visualization Methods

In order to make the analysis usable for the end user, analytics results need to be visualized in an understandable and clear manner. The high volume and excessive rate of generation of data makes visualization of big data a daunting challenge. Evidently, it is not possible to use traditional visualization methods for this purpose.

5.0 BIG DATA COMPUTING MODELS

This section gives taxonomy (see Fig. 3) of big data computing models and discusses the different tools belonging to each of the categories.

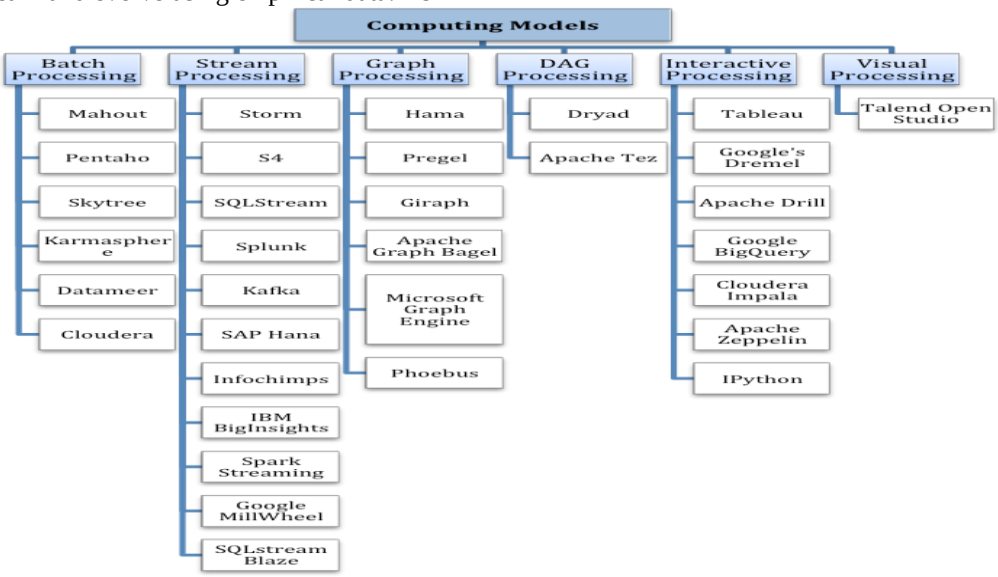

Fig. 3 – Big Data Computing Models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

5.1 Batch Processing

Theoretically, batch processing is a processing mode in which a series of jobs are performed on a batch of inputs. MapReduce programming paradigm is the most effective and efficient solution for batch processing of big data. Hadoop, a MapReduce implementation, is identified as the most popular big data processing platform.Various tools are metioned below.

Mahout : Used for scalable machine learning in the parallel environment

Pentaho : Hadoop-based software platform for business reports generation

SkyTree : A general-purpose server used for machine learning and advanced analytics, which enables optimized machine learning implementation for real-time analytics.

Karmasphere: Hadoop-based platform for analysis of business big data

Datameer Analytic Solution (DAS): HadoopService based Platform(PAAS)

Cloudera: Apache Hadoop distribution system

5.2 Stream Processing

Stream processing is considered to be the next-generation computing paradigm for big data. The tools available for this purpose have been described below:

Storm: Open-source,scalable, faulttolerant,distributed system for real-time computation.

S4: A scalable, pluggable, distributed, fault-tolerant computing platform.

SQLStream : A platform that supports intelligent, automatic operations, for processing unbounded large-scale data streams.

Splunk: A platform for analyzing machine-generated data streams.

Kafka: Developed for LinkedIn for in-memory management and processing of messaging and stream data.

SAP Hana: A tool for in-memory processing of data streams.

Infochimps: A cloud suite that provides Infrastructure-as-a-Service (IaaS) provisioning.

BigInsights: A stream-based tool by IBM, used in Infosphere platform for big data analytics.

Spark Streaming: Stream processing component of Spark.

Google MillWheel : Framework for fault-tolerant stream processing.

SQLstream Blaze: Stream processing suite.

5.3 Graph Processing

Big data graph processing techniques work in accordance with the Bulk Synchronous Parallel (BSP) computing paradigm (Cheatham et al. 1996). This computing paradigm is commonly used in Cloud Computing.Graph Processing Systems are metioned below:-

Hama (Seo et al. 2010): BSP-inspired computing paradigm that runs on top of Hadoop.

Pregel (Malewicz et al. 2010): A graph computation model that approaches problems using BSP programming model.

Giraph (http://giraph.apache.org/): Scalable, iterative graph processing system.

Apache Spark Bagel (Spark 2016): A Pregel implementation, which is a component of Spark.

Microsoft Graph Engine (Microsoft 2016): In-memory and distributed graph processing engine.

Phoebus (XSLogic 2016): Large-scale graph processing framework.

5.4 DAG Processing

DAG processing is considered a stepup from the MapReduce programming model as it avoids the scheduling overhead prevalent in MapReduce and provides developers with a convenient paradigm for modeling complex applications that require multiple execution steps. Dryad (Isard et al. 2007) is based on dataflow graph processing-based programming model that supports scalable and distributed programming applications. Another application framework that allows execution of a complex DAG created from tasks is Apache Tez[4]. This framework is built on top of YARN.

5.5 Interactive Processing

There needs to be a system that sits between big data applications and users to facilitate smooth communication between the two entities, in order to make big data applications usable. Some of the tools that are developed for this purpose are Tableau, Google Dremel (Melnik et al. 2011), Apache Drill, Google BigQuery, Cloudera Impala, Apache Zeppelin, IPython.

5.6 Visual Processing

One of the most popular tools for visual big data analytics is Talend Open Studio[5]. It can be integrated with Hadoop and user interfaces can be easily created using its drag-and-drop functionality. Moreover, it also offers RSS feed functionality.

6.0 CLOUD COMPUTING FOR BIG DATA ANALYTICS

Cloud Computing has come a long way to become the technology that transforms McCarthy's ideas into reality. The introduction of solutions like Amazon's Elastic Compute Cloud (Amazon EC2[6]) and Google App Engine[7] have been historic milestones in the history of Cloud Computing. Cloud Computing is a technology that allows on-demand, convenient and ubiquitous network access to computing resources that can be configured with minimal requirement of management and interaction with the service provider. The NIST definition also mentioned the five key characteristics, deployment models and delivery models for Cloud Computing. An overview of the NIST definition of Cloud Computing is illustrated in Fig. 4. There are three main components of the Cloud Computing Ecosystem namely, end-user or consumer, distributed server and data center. The cloud provider provisions the IT resources to the end-user with the help of distributed server and data center.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

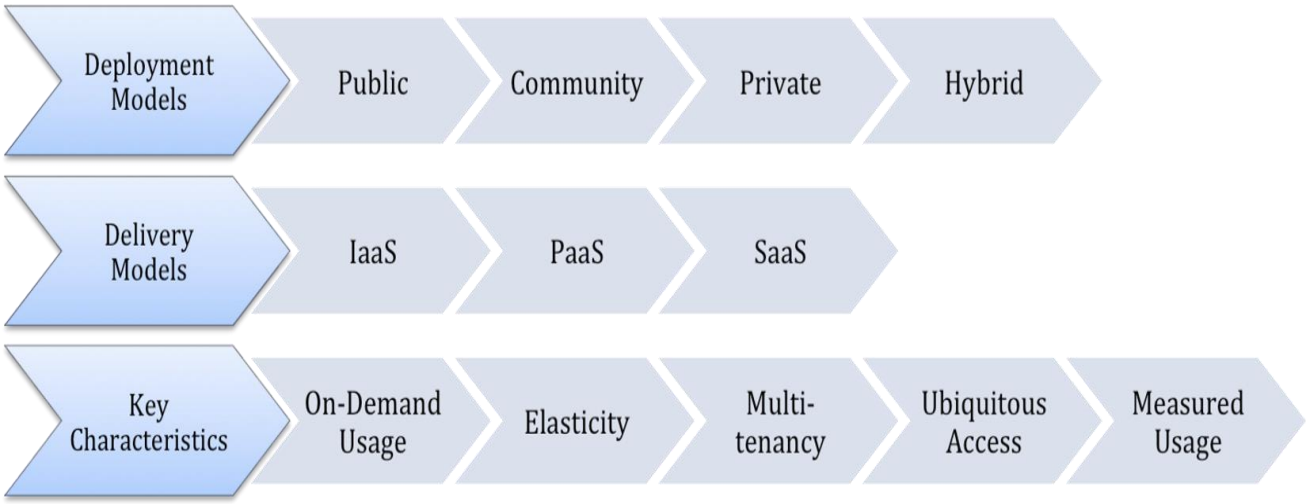

Fig. 4 – Cloud Computing Model Defined By NIST

6.2.1 Infrastructure-as-a-Service (IaaS)
Popular cloud providers of IaaS are Rackspace[8] and Amazon EC[2].
6.2.2 Platform-as-a-Service (PaaS)
While infrastructure forms the base layer and requirement for any kind of development or usage, developers may also require pre-deployed and pre-configured IT resources.Some of the most popular products in this category are .NET-based environment provided by Microsoft Azure[9] and a Python-based environment by the Google App Engine.
6.2.3 Software-as-a-Service (SaaS)
The shared cloud service can also host software solutions that can directly be used by consumers on need basis. Products like Google Docs[10], which is a shared service and provisions documentation software and storage to consumers, are examples of SaaS.
6.2.4 Hybrid Models
Apart from the three formal models used for delivery of cloud solutions, the user also has the option to use any combinations of these models.
6.3 Deployment Models
The cloud environment has three main characteristics namely, size, ownership and access. On the basis of these three characteristics, the deployment model for the cloud environment is determined.
6.3.1 Public Cloud
The creation and maintenance of the cloud environment is the sole responsibility of the cloud provider.
6.3.2 Community Cloud
The Community Cloud is an adaptation of the Public Cloud. The only difference between these two types of deployments is that the community cloud restricts access to the cloud services. Only a small community of cloud consumers can use the services of this cloud.
6.3.3 Private Cloud
When an individual or organization owns a cloud and limits the access of the cloud to the members of the organization only, the deployment model is known as Private Cloud.
6.3.4 Hybrid Cloud
A deployment model created using two or more deployment models is called a hybrid cloud.
6.3.5 Other Deployment Models
Some other cloud deployment models like External Cloud, Virtual Private Cloud or Dedicated Cloud and Inter-cloud also exist.

6.4 What makes Cloud Computing an Ideal Match for Big Data?

Big data solutions have two fundamental requirements. The size of the data is 'big'. Therefore, in order to store this data, a large and scalable storage space is required. Moreover, the standard analytics algorithms are computing-intensive. Therefore, an infrastructural solution that can support this level of computation is needed. The cloud meets both these requirements well.

Firstly, there are several low-cost storage solutions available with the cloud. Besides this, the user pays for the services he or she uses, which makes the solution all the more cost effective. Secondly, cloud solutions offer commodity hardware, which allows effective and efficient processing of large datasets. It is because of these two reasons that Cloud Computing is considered an ideal infrastructural solution for big data analytics.

6.5 Hadoop on the Cloud

One of the most popular frameworks used for big data computing is Hadoop. It is an implementation of MapReduce that allows distributed processing of large, heterogeneous datasets. There are many solutions that allow moving of Hadoop to the cloud. Some of the popular

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

solutions are the ones provided by Amazon's Elastic MapReduce[11] and Rackspace. There are several reasons why running Hadoop on the cloud is gaining immense popularity.

6.5.1 Deploying Hadoop in Public Cloud

Providers like Hortonworks[12], Cloudera[13] and BigInsights[14] offer Hadoop distributions, which can be deployed and run on public clouds provided by Rackspace, Microsoft Azure and Amazon Web Services[15]. Such a configuration is typically referred to as 'Hadoop-as-a-Service'. The issue with such solutions is that they use Infrastructure-as-a-Service (IaaS) provided by the cloud providers. In other words, the IT resources being used are shared between many customers. This gives the user little control over the configuration of the cluster. Besides this, the availability and performance of the cluster are also dependent on the VM (Virtual Machine) that is being used. The advantages of using Hadoop on a private cloud are as follows:

- Better control and visibility of the cluster
- Better mitigation of data privacy and security concerns

6.5.3 Key Considerations for Deployment

There are obvious advantages of running Hadoop on the Cloud. However, it is important to understand that this does not come without problems and potential issues. Some of the things that must be paid heed to before using Hadoop on the cloud are given below.

- The security provided by the Hadoop cluster is very limited in its capability. Therefore, the security requirements and criticality of data being shared with the Hadoop cluster need to be carefully examined, in advance.
- Typically, Hadoop runs on Linux. However, Hortonworks also provides a Hadoop distribution that works with Windows and is available on Microsoft's Azure Cloud. It is important to identify the operating system requirements and preferences before choosing a Cloud-based Hadoop solution.
- Hadoop can never be viewed as a standalone solution. When it comes to designing big data analytics applications, you will need to look beyond the Hadoop cluster and see if the cloud solution supports visualization tools like Tableau and R, to serve your purpose in totality.
- An important consideration that is usually overlooked is data transmission. Is the data already on the cloud or will it have to be loaded from an internal system? If the application needs transferring of data from one public cloud to another, some transmission fees may apply.
- Using the VM-based Hadoop cluster may suffer from performance issues. These arrangements are good solutions only for development and testing purposes or unless performance is not an issue.

7.0 CHALLENGES AND OPPORTUNITIES:

Big data computing requires the use of several techniques and technologies. MapReduce and Hadoop are certainly the most popular and useful frameworks for this purpose. Apart from Cloud Computing, it has also been proposed that bio-inspired computing; quantum computing and granular computing are potential technologies for big data computing (Philip Chen and Zhang 2014). However, each of these technologies needs to be adapted for this purpose and is not free from potential challenges.

Owing to the elasticity and scalability of cloud solutions, this technology is one of the frontrunners for big data computing (Talia 2013). With that said, the feasibility and viability of using a synergistic model is yet to be explored. NESSI presented challenges specific to implementation of existing machine learning techniques for big data computing and development of analytics solutions, mentioning the following requirements as fundamental (NESSI 2012).

- There is a need for development of solid scientific foundation, to facilitate selection of method or technique that needs to be chosen.
- There is a need for development of scalable and efficient algorithms that can be used.
- The developed algorithms cannot be implemented unless appropriate technological platforms have been selected.
- Lastly, the solution's business viability must be explored.

There are several identified challenges related to the use of Cloud Computing for big data analytics (Hashem et al. 2015, Assunção et al. 2015). Big data in the cloud suffers from several trials, both technical as well as non-technical. Technical challenges associated with cloud-based big data analytics can further be divided into three categories namely, big data management, application modeling and visualization.

Challenges associated with characteristics, storage and processing of big data are included in big data management. Management of big data is a challenging task considering the fact that data is continuously increasing in volume. Moreover, aggregation and integration of unstructured data, collected from diverse sources is also under research consideration. There are two aspects of data acquisition and integration. Firstly, data needs to be collected from different sources.

Besides this, the collected data may be structured, unstructured or semi-structured, in type. Integration of a variety of data types into an aggregation that can be further used for analytics is an even bigger problem.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

CONCLUSION:

The amount of data generated by devices and other internet based sources regularly is huge, which is called big data. This data can be processed and analyzed to develop useful applications for specific domains. Several mathematical and data analytics techniques have found use in this sphere. This has given rise to the development of computing models and tools for big data computing. This chapter illustrate the big data problem, giving useful insights in the tools, techniques and technologies that are currently being used in this domain, with specific reference to Cloud Computing, as the infrastructural solution for the storage and processing requirements of big data. Although, the big data problem can model any data-intensive system, there are some established practical applications that have gained popularity amongst the research community and governing authorities. These applications include smart cities (Khan, Anjum, and Kiani 2013), analytics for healthcare sector (Raghupathi and Raghupathi 2014), asset management system for railways (Thaduri, Galar, and Kumar 2015), social media analytics (Burnap et al. 2014), geospatial data analytics (Lu et al. 2011), customer analytics for the banking sector (Sun et al. 2014), e-commerce recommender systems (Hammond and Varde 2013) and Intelligent Systems for transport (Chandio, Tziritas, and Xu 2015), in addition to several others.

7.0 FUTURE RESEARCH DIRECTIONS

There is a need for a Cloud-based framework to facilitate advanced analytics. The analytical workflow is composed of several steps, which include data acquisition, storage, processing, analytics and visualization. Besides this, each of these steps is composed of many sub-steps. For instance, data acquisition is composed of sub-steps like data collection, pre-processing and transformation.

With specific focus on processing and analytics, existing solutions in the field are not generic. There is tight coupling between field-specific analytical solutions and data model used for the same. In addition to this, data model diversity also exists as a fundamental issue for generic framework development. Lastly, the data characteristics used to classify data as big data varies substantially with time and changes from one application to another. Research needs to be directed towards development of a generic analytical framework and cloud-based big data stack that addresses the complexities of issues mentioned above.

Big data technology applies and appeals to every walk of human life. There is no technology enabled system that cannot make use of the big data-powered solutions for enhanced decision making and industry-specific applications. However, in order to make this technology commercially viable, research groups need to identify potential ‘big’ datasets and possible analytical applications for the field concerned. With that said, the feasibility and commercial viability of such analytical applications need to aligned with business and customer requirements.

REFERENCES

[1] Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning From Data*. United States: AMLBook.com.
[2] Akidau, Tyler, Sam Whittle, Alex Balikov, Kaya Bekiroğlu, Slava Chernyak, Josh Haberman, Reuven Lax, Sam McVeety, Daniel Mills, and Paul Nordstrom. 2013. "MillWheel." *Proceedings of the VLDB Endowment* 6 (11):1033-1044. doi: 10.14778/2536222.2536229.
[3] Assunção, Marcos D., Rodrigo N. Calheiros, Silvia Bianchi, Marco A. S. Netto, and Rajkumar Buyya. 2015. "Big Data computing and clouds: Trends and future directions." *Journal of Parallel and Distributed Computing* 79-80:3-15. doi: 10.1016/j.jpdc.2014.08.003.
[4] Buyya, Rajkumar. 2016. "Big Data Analytics = Machine Learning + Cloud Computing." In *Big Data*, 7-9. Massachusetts, USA: Morgan Kaufmann Publisher.
[5] Chandio, Aftab Ahmed, Nikos Tziritas, and Cheng-Zhong Xu. 2015. "Big-data processing techniques and their challenges in transport domain." *ZTE Communications*.
[6] Cheatham, Thomas, Amr Fahmy, Dan Stefanescu, and Leslie Valiant. 1996. "Bulk Synchronous Parallel Computing — A Paradigm for Transportable Software."61-76. doi: 10.1007/9781-4615-4123-3_4.
[7] Gartner. 2016. "Gartner IT Glossary." Gartner Inc.
[8] Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. "Detecting influenza epidemics using search engine query data." *Nature* 457 (7232):1012-4. doi: 10.1038/nature07634.
[9] Hammond, Klavdiya, and Aparna S. Varde. 2013. "Cloud Based Predictive Analytics: Text Classification, Recommender Systems and Decision Support."607-612. doi: 10.1109/icdmw.2013.95.
[10] Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. 2015. "The rise of “big data” on cloud computing: Review and open research issues." *Information Systems* 47:98-115. doi: 10.1016/j.is.2014.07.006.
[11] Isard, Michael, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly. 2007. "Dryad." *ACM SIGOPS Operating Systems Review* 41 (3):59. doi: 10.1145/1272998.1273005.
[12] Khan, Zaheer, Ashiq Anjum, and Saad Liaquat Kiani. 2013. "Cloud Based Big Data Analytics for Smart Future Cities."381-386. doi: 10.1109/ucc.2013.77.
[13] Qian, Ling, Zhiguo Luo, Yujian Du, and Leitao Guo. 2009. "Cloud Computing: An Overview." 5931:626-631. doi: 10.1007/978-3-642-10665-1_63.
[14] Raghupathi, W., and V. Raghupathi. 2014. "Big data analytics in healthcare: promise and potential." *Health Inf Sci Syst* 2:3. doi: 10.1186/2047-2501-2-3.
[15] Ratner, Bruce. 2003. *Statistical Modeling And Analysis For Database Marketing*. Boca Raton, Fla.: Chapman & Hall/CRC.
[16] Sahimi, Muhammad, and Hossein Hamzehpour. 2010. "Efficient Computational Strategies for Solving Global Optimization Problems." *Computing in Science & Engineering* 12 (4):74-83. doi: 10.1109/mcse.2010.85.
[17] Salminen, A. 2012. "Introduction to NOSQL." NoSQL Seminar 2012.
[18] Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3):210-229. doi: 10.1147/rd.33.0210.
[19] Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. "The Hadoop Distributed File System."1-10. doi: 10.1109/msst.2010.5496972.
[20] Spark. 2016. "Spark Documentation." Spark.Apache.Org.
[21] Sun, N., J. G. Morris, J. Xu, X. Zhu, and M. Xie. 2014. "iCARE: A framework for big databased banking customer analytics." *IBM Journal of Research and Development* 58 (5/6):4:1-4:9. doi: 10.1147/jrd.2014.2337118.
[22] Talia, Domenico. 2013. "Clouds for Scalable Big Data Analytics." *Computer* 46 (5):98-101. doi: 10.1109/mc.2013.162.
[23] Xiaodong, Li, and Yao Xin. 2012. "Cooperatively Coevolving Particle Swarms for Large Scale Optimization." *IEEE Transactions on Evolutionary Computation* 16 (2):210-224. doi: 10.1109/tevc.2011.2112662.