CLOUD BASED BIG DATA ANALYTICS: A SURVEY OF CURRENT RESEARCH AND FUTURE DIRECTIONS

[1]R.Suganya, [2]M. Pavithra, [3]A.Rathika, [4]R.Ashwini

Assistant Professor,
Department of C.S.E,
Jansons Institute of Technology, Coimbatore, India

***Abstract*: The advent of the digital age has led to a rise in different types of data with every passing day. In fact, it is expected that half of the total data will be on the cloud by 2016 [1]. This data is complex and needs to be stored, processed and analyzed for information that can be used by organizations [2]. Cloud computing provides an apt platform for big data analytics in view of the storage and computing requirements of the latter. This makes cloud-based analytics a viable research field. However, several issues need to be addressed and risks need to be mitigated before practical applications of this synergistic model can be popularly used [3]. It explores the existing research, challenges, open issues and future research direction for this field of study. It discusses approaches and environments for carrying out analytics on Clouds for Big Data applications [4]. It revolves around four important areas of analytics and Big Data, namely (i) data management and supporting architectures; (ii) model development and scoring; (iii) visualization and user interaction; and (iv) business models. Through a detailed survey, we identify possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions [5]. A cloud framework refers to the aggregation of components like development tools, middleware and database services, needed for cloud computing, which aids in developing, deploying and managing cloud based applications strenuously, consequently making it an efficacious paradigm for massive scaling of dynamically allocated resources and their complex computing. Big Data Analytics (BDA) delivers data management solutions in the cloud architecture for storing, analyzing and processing a huge volume of data [6]. It presents a survey for performance based comparative analysis of cloud-based big data frameworks from leading enterprises like Amazon, Google, IBM, and Microsoft, which will assist researcher, IT analysts, reader and business user in picking the framework best suited for their work ensuring success in terms of favorable outcomes [7].**

***Keywords*: Cloud-based Big Data Analytics, Big Data, Big Data Analytics, Big Data Cloud Computing, Big Data, Cloud based Big Data Enterprise Solutions, Big Data Storage, Big Data Warehouse, Streaming Data, Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM Cloud, Microsoft Azure**

I. INTRODUCTION

With the advent of the digital age, the amount of data being generated, stored and shared has been on the rise. From data warehouses, webpages and blogs to audio/video streams, all of these are sources of massive amounts of data [8]. The result of this proliferation is the generation of massive amounts of pervasive and complex data, which needs to be efficiently created, stored, shared and analyzed to extract useful information. This data has huge potential, ever-increasing complexity, insecurity and risks, and irrelevance [9]. The benefits and limitations of accessing this data are arguable in view of the fact that this analysis may involve access and analysis of medical records, social media interactions, financial data, government records and genetic sequences [10]. The requirement of an efficient and effective analytics service, applications, programming tools and frameworks has given birth to the concept of Big Data Processing and Analytics [11].

Big data analytics has found application in several domains and fields. Some of these applications include medical research, solutions for the transportation and logistics sector, global security and prediction and management of issues concerning the socio-economic and environmental sector, to name a few. Apart from standard applications in business and commerce and society administration, scientific research is one of the most critical applications of big data in the real world [13]. Some of the identified high-impact areas include systems biology, structure and protein function prediction, personalized medicine and metagenomics. Besides this, one of the most relevant applications of big data analytics is to improve the existing business models for efficiency and customer satisfaction [12]. Big data, by definition, is a term used to describe a variety of data - structured, semi structured and unstructured, which makes it a complex data infrastructure [17]. The complexity of this infrastructure requires powerful management and technological solutions. One of the commonly used models for explaining big data is the multi-V model. Some of the Vs. used to characterize big data include variety, volume, velocity, veracity and value [14]. The different types of data available on a dataset determine variety while the rate at which data is produced determines velocity. Predictably, the size of data is called volume. The two additional characteristics, veracity and value, indicate data reliability and worth with respect to big data exploitation, respectively [16].

Data is the central element of communication and collaboration in Internet and all the applications that are built on this platform. The immense popularity of data intensive applications like Facebook, LinkedIn, Twitter, Amazon, eBay and Google+ contributes to increasing requirement of storage and processing of data in the cloud environment. Schouten [1] uses Gartner's estimation to predict that by the year 2016, half of the data will be on the cloud. Moreover, the data mining algorithms used for Big Data analytics possess high computing requirements [2]. Therefore, they require high performance processors to do the job. The cloud provides a

good platform for big data storage, processing and analysis, addressing two of the main requirements of big data analytics, high storage and high performance computing [3].

The cloud computing environment offers development, installation and implementation of software and data applications 'as a service'. Three multi-layered infrastructures namely, platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS), exist [2]. Infrastructure-as-a-service is a model that provides computing and storage resources as a service. On the other hand, in case of PaaS and SaaS, the cloud services provide software platform or software itself as a service to its clients [3]. The cost of storage has considerably reduced with the advent of cloud-based solutions. In addition, the 'pay-as-you-go' model and the concept of commodity hardware allow effective and timely processing of large data, giving rise to the concept of 'big data as a service' [4].

An example of one such platform is Google BigQuery, which provides real-time insights from big data in the cloud environment [12]. However, there have not been many practical applications of big data analytics that make use of the cloud. This has led to an increasing shift of research focus towards cloud-based big data analytics [5]. An issue that is evident in this arrangement is information security and data privacy. As part of the cloud services, trust in data is also defined as a service. There shall be a considerable decrease in trust in view of the fact that the chances of security breaches and privacy violation will significantly raise upon implementation of big data strategies in the cloud. In addition, another important issue of ownership and control will also exist [6].

II.REALTED WORK

Traditional data management tools and data processing or data mining techniques cannot be used for Big Data Analytics for the large volume and complexity of the datasets that it includes [7]. Conventional business intelligence applications make use of methods, which are based on traditional analytics methods and techniques and make use of OLAP, BPM, Mining and database systems like RDBMS [8].

One of the most popular models used for data processing on cluster of computers is MapReduce. It identifies MapReduce/Hadoop as the most productive model for Big Data Analytics yet mentions that languages and extensions like HiveQL, Latin and Pig have overpowering benefits for this use [9]. Hadoop is simply an open-source implementation of the MapReduce framework, which was originally created as a distributed file system. According to the paper, the evolution of Hadoop as a complete ecosystem or infrastructure that works alongside MapReduce components and includes a range of software systems like Hive and Pig languages, a coordination service called Zookeeper and a distributed table store called HBase [10]. For cloud-based big data analytics, several frameworks like Google MapReduce, Spark, Hadoop, Twister, Hadoop Reduce and Hadoop++ are available. It gives a pictorial representation of the use of cloud computing in big data analytics. These frameworks are used for storing and processing of data [11].

Some of the recent research breakthroughs and milestones in cloud-based big data analytics are discussed here. Lee [16] elaborates on the advantages and limitations of MapReduce in parallel data analytics. A Hadoop-based data analytics system, created by Starfish [13], improves the performance of the clusters throughout the cycle of data analytics. Moreover, the users are not required to understand the configuration details.

Research efforts have been made to create a big data management framework for the cloud. Khan, It propose a data model and provides a schema for big data in the cloud and attempts to ease the process of querying data for the user [8]. Moreover, an important subject of research has been performance and speed of operation. It explores the use of a proposed integrated Hadoop and MPI/OpenMP system and how the same can improve speed and performance [7].

In view of the fact that data needs to be transferred between data centers that are usually located distances apart, power consumption becomes a crucial parameter when it comes to analyzing efficiency of the system [10]. A network-based routing algorithm called GreeDi can be used for finding the most energy efficient path to the cloud data center during big data processing and storage [9].

Online risk analytics and the need for an infrastructure that can provide users the programming resources and infrastructure for carrying out the same have also appeared in the form of Aneka [6] and CloudComet [15]. Chen [7] investigates the concept of CAAAS or Continuous Analytics as a Service, which is used for predicting the behavior of a service or a user. The last topic under Big Data Analysis that has caught the attention of the research community is Real-time Big Data Analysis. Many commercial cloud service providers are providing solutions for real-time analysis. AWS based-solutions for real-time stream processing is AWS Kinesis [3]. Many frameworks and software systems have also been introduced for this purpose, some of which are Apache S4 53], IBM InfoSphere Streams [4] and Storm [2].

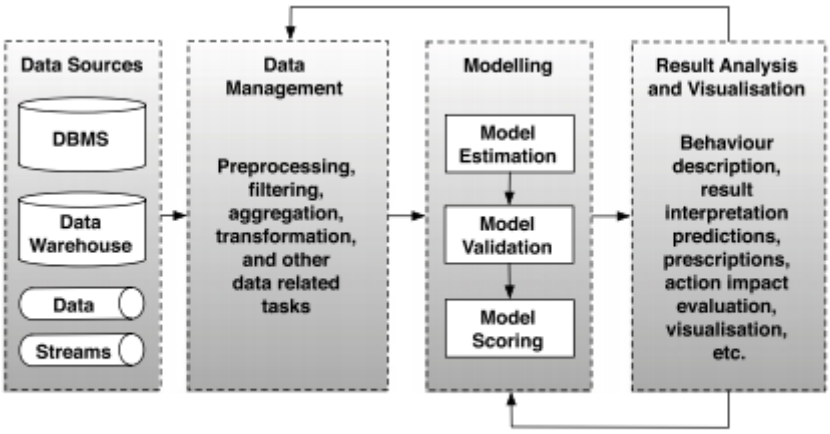


Fig. 1. Overview of the analytics workflow for Big Data.

III.CHALLENGES AND ISSUES

In order to move beyond the existing techniques and strategies used for machine learning and data analytics, some challenges need to be overcome. NESSI [2] identifies the following requirements as critical.

1. In order to select an adequate method or design, a solid scientific foundation needs to be developed. 2. New efficient and scalable algorithms need to be developed.

3. For proper implementation of devised solutions, appropriate development skills and technological platforms must be identified and developed.

4. Lastly, the business value of the solutions must be explored just as much as the data structure and its usability

In view of cloud-based big data analytics, additional challenges like adoption and implementation of effective big data solutions using cloud architecture and mitigating the security and privacy risks also exist [1]. One of the biggest concerns while using big data analytics and cloud computing in an integrated model is security. This is perhaps the reason why this aspect of cloud-based big data analytics and its practical usage and implementation has attracted immense attention [2]. It provides a summary, analysis and comparison of authenticator-based data integrity verification techniques on cloud and Internet-of-things data [3].

This paper suggests that any future developments in this area needs to look at three main aspects namely, efficiency, security and scalability/elasticity. A G-Hadoop based security framework is proposed [6], which makes use of solutions like SSL and public key cryptography for ensuring security of big data resident on distributed cloud data centers. In addition to several security mechanisms, this framework also aims to simplify the processes of submitting job and authenticating users [4].

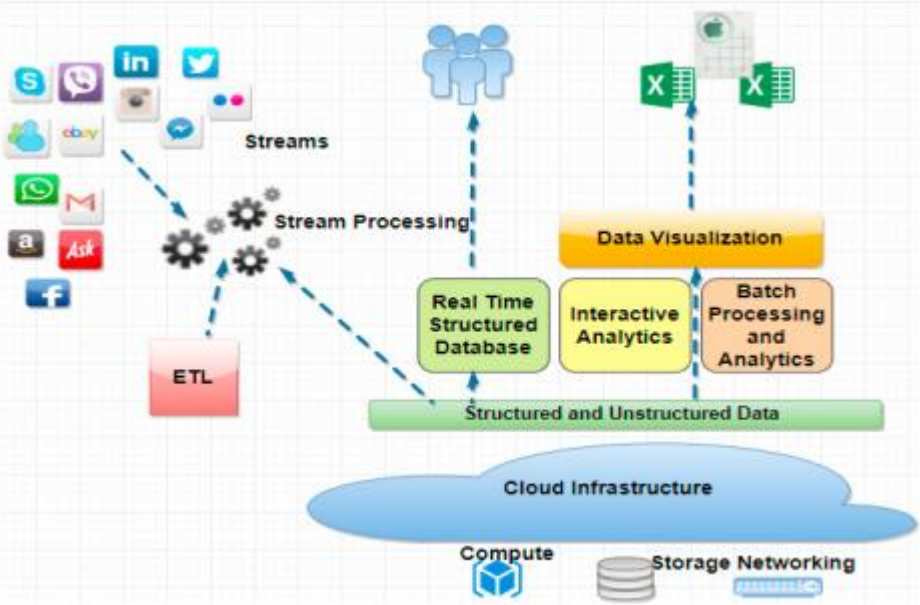
It suggests further research and development in the following areas:

1. Programming abstracts or scalable high-level models and tools.

2. Solutions for data and computing interoperability issues.

3. Integration of different big data analytics frameworks

4. Techniques for mining provenance data



IV.FUTURE RESEARCH DIRECTIONS

Several open source data mining techniques, resources and tools exist. Some of these include R, Gate, Rapid-Miner and Weka, in addition to many others [5]. Cloud-based big data analytics solutions must provide a provision for the availability of these affordable data analytics on the cloud so that cost-effective and efficient services can be provided [6]. 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. In view of this, some of the main research directions identified by Neaga and Hao [9] include:

1. Evolution of analytics and information management with respect to cloud-based analytics.

2. Adaptation and evolution of techniques and strategies to improve efficiency and mitigate risks.

3. Formulate strategies and techniques to deal with the privacy and security concerns.

4. Analysis and adaptation of legal and ethical practices to suit the changing viewpoint, impact and effects of technological advances in this regard. With this said, the research directions are not limited to the above-mentioned points [7]. The main goal is to transform the cloud from being a data management and infrastructure platform to a scalable data analytics platform.

Table1. Comparison of Cloud Based Big Data Enterprise Solution Frameworks

S No	Framework Type and Features	AWS	GCP	IBM Cloud	MS Azure
1.	Big Data Analytics	Amazon ElasticSearch Service	Google Cloud Dataproc	IBM Analytics Engine	Azure HDInsight
	Mode of Software	Open-Source	Open-Source	Open-Source	Open-Source
	Types of Data	Structured, semi-structured and unstructured	Structured, semi-structured and unstructured	Unstructured	Unstructured
	Data Sources	Amazon S3, Amazon Kinesis Firehose, and Amazon DynamoDB	Google Bigtable, Google Cloud Storage, and Google BigQuery	IBM Cloud Object Storage	Blob Storage
	Supported Operating System	CentOS, Ubuntu, and Amazon Linux	Debian 8	CentOS 7	Ubuntu 14, Ubuntu 16, and Windows Server 2012 R2
	Applications	Logs analytics, real-time applications monitoring, and clickstream analytics	Batch processing, querying, streaming, and machine learning	Data analytics, enterprise solution for various Big data problems, and analytics applications development and deployment	Stream and Batch data analytics
	Service Integration	Yes	Yes	Yes	Yes
	Deployment	Zonal	Zonal	Regional	Regional

Some of the research questions to be explored:

- Can we efficiently access data from a Big Data platform?

- Can we apply the existing data mining techniques to extract knowledge from the Big Data stored in distributed nodes?

- Are we missing any valuable data while processing and mining for information retrieval? Big Data systems have grown-up insignificance, and all types of private and public organizations are progressively more aware of the potential benefits of Big Data as an enabler to make use of their data [8].

The IT industry has reacted by exploring enormous efforts in Big Data systems; but, their limitations are becoming more and more evident. From a technical point of view, the prospect of Big Data will be formed by the new solutions that deal with these limitations [9]:

- New systems that permit the analysis of both structured and unstructured data to be joined, i.e., capable to unite multiple data sources (from social media to data warehouses) in an approach that is controllable, not only for the professionals but also more non-professional users and groups.

- New embedded analytics that exploits the streams of data in actual time under strict resource restrictions of computing capacity, storage, energy, and communication bandwidth.

- New paradigms that super-seed the pure batch and pure real-time method of current Big Data tools.

- New application frameworks capable to grasp all distributed computing resources, admitting to run diverse kinds of tasks (batch, stream analysis, interactive) virtualizing all the underlying infrastructure and scheduling usage based on the task needs.

- New database systems proficient to manage large datasets whilst holding the transactional semantics of data operations on hand in conventional relational databases.

- New Big Data tools that are guiding and controlling ethical, security and privacy issues in Big Data research.

V.ANALYSIS PIPELINE AND CHALLENGES WITH BIG DATA

The data type that rises most hurriedly is unstructured data. This data type is regarded as by "human information" like high-definition videos, movies, photos, scientific simulations, financial transactions, phone records, genomic datasets, seismic images,

geospatial maps, e-mail, tweets, Facebook data, call-center conversations, mobile phone calls, website clicks, documents, sensor data, telemetry, medical records and images, climatology and weather records, log files, and text [5]. As per statistics of the Computer World, unstructured information may well account for more than 70% to 80% of all data in organizations [7]. These data, which mainly initiate from social media, comprise 80% of the data worldwide and account for 90% of Big Data. At present, 84% of IT managers process unstructured data, and this percentage is expected to plummet by 44% in the near future [6]. The majorities of unstructured data is not modeled, is random, and are hard to analyze. For many organizations, suitable strategies are required to develop to deal with such data. Table 1 illustrates the quick production of data in different organizations further. As per statistics of the Industrial Development Corporation (IDC) and EMC Corporation, the quantity of data produces in 2020 will be 44 times larger [40 zettabytes (ZB)] than in 2009. This rate of rising is expected to continue at 50% to 60% annually [4].

VI.CHALLENGES IN BIG DATA ANALYSIS

i) Heterogeneity and incompleteness: When people get through information, a vast deal of heterogeneity is at ease tolerated. But, machine analysis algorithms look forward to homogeneous data, and cannot recognize nuance [2]. As a result, data must be with awareness structured as the first move in data analysis. An efficient demonstration, access, and analysis of semi-structured data entail additional effort. Some incompleteness and a number of errors in data are likely to remain even after data cleaning and error correction. There is a big challenge to correctly handle this incompleteness and these errors [3].

ii) Scale: Handling huge and quickly growing volumes of data has been a challenging matter for many decades. In previous years, this challenge was overcome by processors reaching faster following Moore's law, to present with the resources required to deal with escalating volumes of data [4].

iii) Timeliness: The turn over part of size is speed. The outsized the data set to be processed, the longer it will take to analyze. To effectively deal with the size of data, a new system is to be designed that also likely to process a given size of data set faster [5]. But, one should also come up with velocity along with speed. It is frequently essential to search elements in a big data set that meet up a particular principle. In terms of data analysis, this type of finding is expected to happen frequently. Scanning the whole data set to locate appropriate elements is noticeably unfeasible. To a certain extent, index structures are formed in proceeding to allow finding qualifying elements promptly [6].

iv) Privacy: The privacy of data is a further vast concern and one that rises in the circumstance of Big Data. Controlling privacy is effectively together with a technical and a sociological problem, which has to be addressed in cooperation from both perspectives to recognize the promise of Big Data [7]. To keep away a user location from others is a great deal of more challenging than hiding his/her individuality. This is due to with location-based services; the location of the user is essential for a winning data access or data collection, while the uniqueness of the user is not compulsory [8].

v) Human Collaboration: Despite the incredible advances made in computational analysis, there stay behind many patterns that humans can simply identify but computer algorithms have a tough time finding. Preferably, analytics for Big Data will not be all computational – somewhat it will be planned explicitly to have an individual in the loop [9]. A Big Data analysis system must allow input from multiple human experts and joint investigation of results. These manifold experts may be unconnected in space and time when it is too costly to bring together a whole team together in one room. The data system has to admit this distributed expert input and hold up their collaboration. When crowd-sourced data is acquired for hire much of the data formed may be a key objective of getting it done rapidly rather than accurately [10].

VII.SECURITY ISSUES AND CHALLENGES FOR BIG DATA ANALYTICS IN CLOUD

Security is becoming major issue for data storage in cloud based networks. Cloud computing technology comes with security issues which include networks, databases, operating systems, virtualization, resource scheduling and allocation, transaction management, load balancing and memory management [5]. The security issues associated with cloud computing environment can be categorized into several levels such as [3][4]:

- Network level

The issues and challenges associated with the network level includes network protocols and security in networks such as distributed nodes, distributed data etc.

- User Authentication level

The issues and challenges associated with this level includes encryption/decryption techniques, authentication methods which includes authentication of distributed applications, access rights for nodes, logging etc.

- Data level

The issues and challenges associated with this level include integrity of data and availability issues with data such as protection of data and distributed data.

- Generic level

The issues and challenges associated with this level includes different usage of security tools and usage of different technologies Cloud security alliance in 2013 identified top ten challenges for Big Data security such as

- Secure computation in distributed programming frameworks
- Security best practices for non-relational data bases
- Secure data storage and transactions logs
- End-point input validation/filtering
- Real-time security monitoring
- Scalable and composable privacy-preserving data mining and analytics
- Cryptographically enforced data centric security
- Granular access control
- Granular audits
- Data Provenance

VIII.CONCLUSION

This is an age of big data and the emergence of this field of study has attracted the attention of many practitioners and researchers. 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 [2]. Moreover, the cloud infrastructure suffices the storage and computing requirements of data analytics algorithms. On the other hand, open issues like security, privacy and the lack of ownership and control exist [3]. Research studies in the area of cloud-based big data analytics aim to create an effective and efficient system that addresses the identified risks and concerns. The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever increasing speed [4]. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimize its operation, and reduce its costs [5]. Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in alleviating these problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand [6].

Big Data has the perspective to modernize not just research, but also in education and other areas like urban planning (through fusion of high fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data), energy saving (through unveiling patterns of use), homeland security (through analysis of social networks and financial transactions of possible terrorists), and so on [7]. Thus, research like efficient access to Big Data and data mining on Big Data hold out significant roles in a variety of ways to the global economy. This paper presents various data processing platforms that are currently available and whether these are capable to handle Big Data or not. We also focus on some future research works to be completed for efficient access to the Big Data and in addition, a methodology has been proposed to solve the undiscovered issues for the Big Data [8]. It is believed that access methods for Big Data research has considerably broadened the scope of data analysis and will have a deep impact on Big Data access methodologies as well as data mining methodologies and applications in the long run. However, there are still some challenging research issues that need to be solved before access methods for Big Data can claim a keystone approach in data mining, text analytics, and related applications [9].

The area of Big Data Computing using Cloud resources is moving fast, and after surveying the current solutions we identified some key lessons:

• There are plenty of solutions for Big Data related to Cloud computing. Such a large number of solutions have been created because of the wide range of analytics requirements, but they may, sometimes, overwhelm non-experienced users [11]. Analytics can be descriptive, predictive, and prescriptive; Big Data can have various levels of variety, velocity, volume, and veracity. Therefore, it is important to understand the requirements in order to choose appropriate Big Data tools;

• It is also clear that analytics is a complex process that demands people with expertise in cleaning up data, understanding and selecting proper methods, and analyzing results. Tools are fundamental to help people perform these tasks [12]. In addition,

depending on the complexity and costs involved in carrying out these tasks, providers who offer Analytics as a Service or Big Data as a Service can be a promising alternative compared to performing these tasks in-house;

• Cloud computing plays a key role for Big Data; not only because it provides infrastructure and tools, but also because it is a business model that Big Data analytics can follow (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)). However, AaaS/BDaaS brings several challenges because the customer and provider's staff are much more involved in the loop than in traditional Cloud providers offering infrastructure/platform/software as a service [13].

IX.FUTURE WORK

Although Cloud infrastructure offers such elastic capacity to supply computational resources on demand, the area of Cloud supported analytics is still in its early days. In this paper, we discussed the key stages of analytics workflows, and surveyed the state-of-the-art of each stage in the context of Cloud-supported analytics. Surveyed work was classified in three key groups: Data Management (which encompasses data variety, data storage, data integration solutions, and data processing and resource management), Model Building and Scoring, and Visualization and User Interactions [2]. For each of these areas, ongoing work was analyzed and key open challenges were discussed. This survey concluded with an analysis of business models for Cloud-assisted data analytics and other non-technical challenges [3]. Recurrent themes among the observed future work include (i) the development of standards and APIs enabling users to easily switch among solutions and (ii) the ability of getting the most of the elasticity capacity of the Cloud infrastructure. The latter includes expressive languages that enable users to describe the problem in simple terms whilst decomposing such high-level description in highly concurrent subtasks and keeping good performance efficiency even for large numbers of computing resources [4]. If this can be achieved, the only limitations for an arbitrary short processing time would be market issues, namely the relation between the cost for running the analytics and the financial return brought for the obtained knowledge [5].

Future research should address how we should collect, understand and handle Big Data to be used for scientific purposes, correct quantitative research and representativeness [6]. In the literature, there are numerous classification and future prediction algorithms with greater accuracy but the most recent algorithms like cost-sensitive decision forest (an ensemble of decision trees) which are predicting class values more accurately than a single decision tree (used for both prediction and classification) and a decision forest can be seen as pool of logic rules with great potential for knowledge discovery [7].

REFERENCES

[1] Agarwal, D., Das, S. and Abbadi, A, "Big Data and Cloud Computing: Current State and Future Opportunities", ACM 978-1-4503-0528-0/11/0003, 2011.

[2] Assuncao, M. D., Calheiros, R. N., Bianchi, S. and Netto, M. A. S., "Big Data Computing and Clouds: Trends and Future Directions", J. Parallel Distrib. Computing, 79-80, 3-15, 2015.

[3] Borthakur, D., Gray, J., Sarma, J. S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., Ranganathan, K., Molkov, D., Menon, A., Rash, S., Schmidt, R. and Aiyer, A, "Apache Hadoop Goes Real-time at Facebook ", in: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2011), ACM, New York, USA, 2011, pp. 1071–1080, 2011.

[4] Calheiros, R. N., Vecchiola, C., Karunamoorthy, D. and Buyya, R, "The Aneka platform and QoS-driven resource provisioning for elastic applications on hybrid Clouds, Future Gener. Comput. Syst. 28 (6), 861–870, 2012.

[5] Chen, Q., Hsu, M. and Zeller, H, " Experience in Continuous analytics as a Service (CaaaS)", in: Proceedings of the 14th International Conference on Extending Database Technology, ACM, New York, USA, pp. 509–514, 2011.

[6] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. and Khan, S. U, "The rise of "big data" on cloud computing: Review and open research issues", Information Systems 47, 98-115, 2015.

[7] Ortiz, J. L. R., Oneto, L. and Anguita, D, "Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf", P2015 INNS Conference on Big Data. Published in Procedia Computer Science. Volume 53, pp. 121-130, 2015.

[8] Baker, T., Al-Dawsari, B., Tawfik, H., Reid, D. and Nyogo, Y, "GreeDi: An energy efficient routing algorithm for big data on cloud", Ad Hoc Networks 00, 1-14, 2015.

[9] 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, 2014.

[10] 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, pp. 58-67, 2015.

[11] O'Driscoll, A., Daugelaite, J. and Sleator, R. D, "Big Data, Hadoop and Cloud Computing in Genomics", Journal of Biomedical Informatics. Volume 46, Issue 5, October 2013, pp. 774-781, 2013.

[12] 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 '15). Procedia Computer Science 50, 517-523, 2015.

[13] Elragal, A, "ERP and Big Data: The Inept Couple", Procedia Technology, 16, 242- 249, 2014.

[14] 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, 2015.

[15] Zhao, J., Wang, L., Tao, J., Chen, J., Sun, W., Ranjan, R., Kołodziej, 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, 994-1007, 2014 .

[16] Shakil, K.A.; Sethi, S.; Alam, M.," An effective framework for managing university data using a cloud based environment", Computing for Sustainable Global Development (INDIACom), 2nd International Conference on , vol., no., pp.1262,1266, 11-13, 2015.

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