Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Security Issues and Challenges in Big Data Analysis

[1]Subaira.A.S*, [2]Gayathri.R, [3]Sindhujaa.N

[1, 3] Assistant Professor/CSE, Mahendra College of Engineering, India

[2]Assistant Professor/IT, Mahendra College of Engineering, India

Abstract*— Data has become an essential part of every Economy, Production, Organization, Business function and individual. The amount of data in world is growing day by day because of use of internet, Smartphone, social network, fine tuning of ubiquitous computing and many other technological advancements. Big Data is a term used to identify the datasets that whose size is beyond the ability of typical database software tools to store, manage and analyses. Generally size of the data is Petabyte and Exabyte. Most of the data is partly structured, unstructured or semi structured and it is heterogeneous in nature. Due to its specific nature, Big Data is stored in distributed file system architectures. Hadoop and HDFS by Apache are widely used for storing and managing Big Data. Analyzing it, is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. Traditional database systems are not able to capture, store and analyze this large amount of data. Security and privacy issues are magnified by velocity, volume, and variety of big data, such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition, and high volume inter-cloud migration. Therefore, traditional security mechanisms, which are tailored to securing small-scale static (as opposed to streaming) data, are inadequate. In this paper, we highlight the some important concept of big data-specific security and privacy challenges so that it will bring renewed focus on fortifying big data infrastructures.

Keywords*— Big data, Petabyte, Exabyte, Database, velocity, volume, Variety

I. INTRODUCTION

The term Big Data is now used almost everywhere in our daily life. The term _Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of Infra Stress". Many Researchers and organizations have tried to define Big Data in different ways. Gartner defines Big Data are high-volume, high-velocity and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization[1].Other areas of research where Big Data is of central importance area astronomy, oceanography, and engineering among many others.

The leap in computational and storage power enables the collection, storage and analysis of these Big Data sets and companies introducing innovative technological solutions to Big Data analytics are flourishing. In this article, we explore the term Big Data as it emerged from the peer reviewed Literature. As opposed to news items and social media articles, peer reviewed articles offer a glimpse into Big Data as a topic of study and the scientific problems methodologies and solutions that researchers are focusing on in relation to it. The purpose of this article, therefore, is to sketch the emergence of Big Data as a research topic from several points: (1)timeline, (2) geographic output, (3) disciplinary output, (4) types of published papers, and (5) thematic and conceptual development.

Fig1. Big Data

The 5Vs that define Big Data are Variety, Velocity and Volume, Variability and Veracity

1) Volume: There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.

2) Velocity: Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

3) Variety: Today, data comes in all types of formats. Structured, Numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. We need to find ways of governing, merging and managing these diverse forms of data.

4) Variability: Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved [2]

5) Veracity: Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. A data environment can lie along the extremes on any one of the following parameters, or a combination of them, or even all of them together.

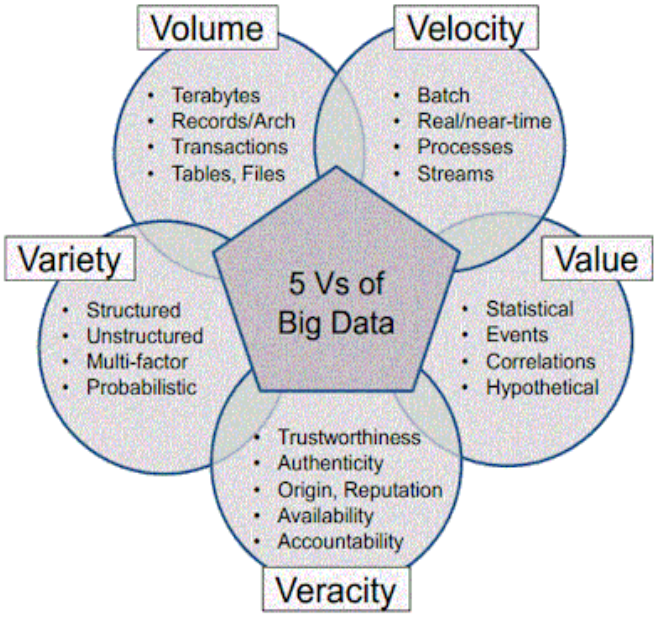

Fig 2.Characteristics of Big Data

Types of Big Data: There are two types of big data.

- Structured Data and Unstructured Data.

1. Structured Data: Structured Data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

2. Unstructured Data: Unstructured Data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

When making an attempt to understand the concept of Big Data, the words such as ―"maps reduce" and "Hadoop" cannot be avoided.

Hadoop

Hadoop, which is a free, Java-based programming frame work, supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure [6]. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, fault tolerant and flexible. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data.

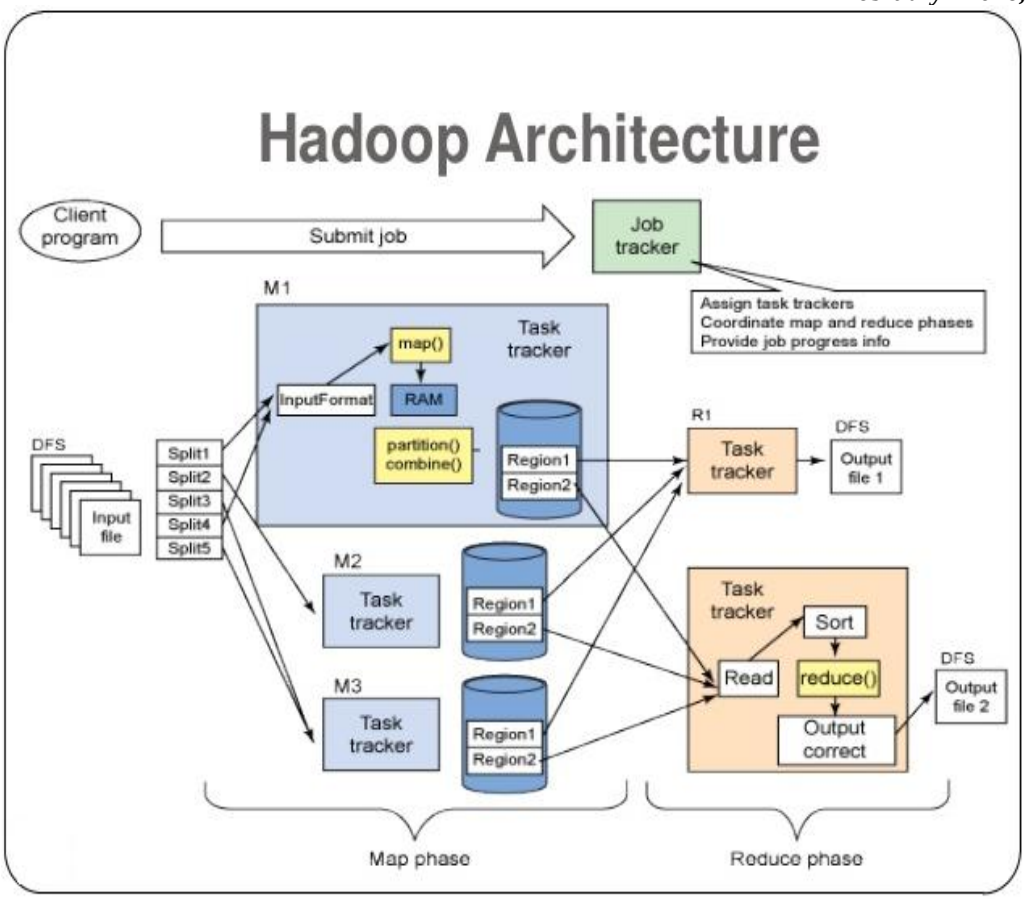

Fig 3.Hadoop Architecture

Hadoop has two main sub projects –Map Reduce & Hadoop Distributed File System (HDFS).

Map Reduce

Hadoop Map Reduce is a framework [7] used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file - system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

Hadoop Distributed File System (HDFS)

HDFS [8] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures

II. LITERATURE REVIEW

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics.The growth of big data has raised a number of eyebrows as far as the challenges are concerned. Several authors have discovered a plethora of challenges which include data storage and privacy. Sachchidanand Singh et al. explained the concept, characteristics & need of Big Data and different offerings available in the market to explore unstructured large data [13].

Changqing ji et al. discussed the scope of Big Data processing in cloud computing environment [14]. Dan Garlasu et al. proposed architecture for managing and processing Big Data using grid technologies [15]. Xiaoxue Zhang et al described the storage challenges of Big Data and they analyzed those using Social Networks as examples [16]. They further classified the related research issues into the following classifications: small files problem, load balancing, replica consistency and deduplication. Meiko Johnson also did some work on the privacy issues involved with Big Data. He classified these challenges into the following taxonomy: interaction with individuals, re-identification attacks, probable vs. provable results, targeted identification attacks and economics effects visualize and understand their algorithm results.

Kapil Bakshi et al [17] discussed the architectural considerations for big data are concluded that despite the different architectures and design decisions, the analytics systems aim for Scale-out, Elasticity and High availability. Chansup Byun et al. brought together the Big Data and Big compute by combining Hadoop clusters and MPI clusters [18]-[19].Tyson Condie et al. discussed the machine learning computational models for Big Data [20]. Xindong Wu et

al.presented a HACE theorem that characterizes the features of the Big Data revolution and proposed a Big Data processing model from the data mining perspective [21]. Dr. Sun-Yuan Kung proposed cost-effective design on kernelbased machine learning and classification for Big Data learning applications [22].

III. STAGES INVOLVED IN BIG DATA

1. Data Acquisition: The first step in Big Data is acquiring the data itself. With the growing medium the rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. The Large Haudron Collider in Switzerland produces petabytes of data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it's merged with other valuable data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.

2. Data Extraction: All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user's movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use. The challenges presented in data extraction are twofold: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the data was initially generated. For instance, footage of a security camera with the same frames may be discarded however it is important not to discard similar data in a case where it is being generated by a heart-rate sensor.Secondly, a lack of a common platform presents its own set of challenges. Due to wide variety of data that exists, bringing them under a common platform to standardize data extraction is a major challenge.

3. Data Collation: Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user.Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bigger picture is often considered a very important part of processing.

4. Data Structuring: Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis. A major issue with big data is providing real time results and therefore structuring of aggregated data needs to be done at a rapid pace.

5. Data Visualization: Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. For instance, data containing average temperatures are shown alongside water consumption rates to calculate a relation in between them. This analysis and presentation of data makes it ready for consumption for users. Raw data cannot be used to gain insights or for judging patterns, therefore "humanizing" the data becomes all the more important.

6. Data Interpretation: The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types:

i)Retrospective Analysis includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us judge the popularity of the show in those areas.

ii)Prospective Analysis includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is an example of prospective analysis.Problems accruing from such interpretations pertains to fallacious and misleading trends being predicted. This is particularly dangerous due to an increasing reliance on data for key decisions. For example, if a particular symptom is plotted against the likelihood of being diagnosed with a particular disease, it might lead to misinformation about the symptom being caused due to the particular disease itself. Insights gained from data interpretation are therefore very important and the primary reason for processing big data as well.All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

IV. BIG DATA SECURITY AND PRIVACY CHALLENGES

1. Secure Computations in Distributed Programming Framework:

Distributed programming framework utilize parallelism in computations and storage to process massive amounts of the data .A popular example is map reduce framework, which splits an input file into multiple chunks in the first phase of map reduce, a mapper for each chunk reads the data, perform some computation, and outputs a list of key/value pairs.In the next phase, a reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper.

2. Security Best Practices for Non-Relational Data Stores:

Non-relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure.For instance, robust solutions to NoSQL injection are still not mature each NoSQL DBs were built to tackle different challenges posed by the analytics world and hence security was never part of the model at any point of its design stage.Developers using NoSQL databases usually embed security in the middleware .NoSQL databases do not provide

any Support for Enforcing it explicitly in the database. However, clustering aspect of NoSQL databases poses additional challenges to the robustness of such security practices.

3. Secure Data Storage and Transaction Logs:

Data and transaction logs are stored in multi-tiered storage media manually moving data between tiers gives the it manager direct control over exactly what data is moved and when. However as the size of data set has been and continues to be, growing exponentially, scalability and availability have necessitated auto-tiring for big data storage management. Auto-tiering solutions do not keep track of where the data is stored, which poses new challenges to secure data storage.

4. End Point Input Validation/Filtering:

Many big data use cases in Enterprise settings require data collection from many sources, such as end point devices for example, a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software application in an enterprise network. A key challenge in the data collection process is input validation: how can we trust the data? How can we validate that a source of input data is not malicious and how can we filter malicious input from our collection? Input validation and filtering is a daunting challenge posed by untrusted input sources, especially with the Bring Your Own Device (BYOD) model.

5. Real –Time Security/Compliance Monitoring:

Real time security monitoring has always been a challenge, given the number of alerts generated by (Security) devices.These alerts (correlated or not) lead to many false positive, which are mostly ignored or simply"clicked away", as humans cannot cope with the shear amount. This problem might even increase with the bid data given the volume and velocity of data streams however, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data.

6. Scalable and Composable Privacy-Preserving Data Mining and Analytics:

Big data can be seen as a troubling manifestation of big brother by potentially enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control. A recent analysis of how companies are leveraging data analytics for marketing purpose identified an example of how a retailer was able to identify that teenager was pregnant before her father knew. Similarly anonym zing data for analytics is not enough to maintain user privacy. For example AOL released anonymized search logs for academic purposes ,but users were easily identified by their searchers .Netflix faced a similar problem when users of their anonymized data set were identified by correlating their Netflix movie scores with IMDB scores. Therefore, it is important to establish guidelines‟ and recommendations for preventing inadvertent privacy disclosures.

7. Cryptographically Enforced Access Control And Secure Communication:

To ensure that the most sensitive private data is end to end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption (ABE) has to be made richer, more efficient, and scalable. To ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented.

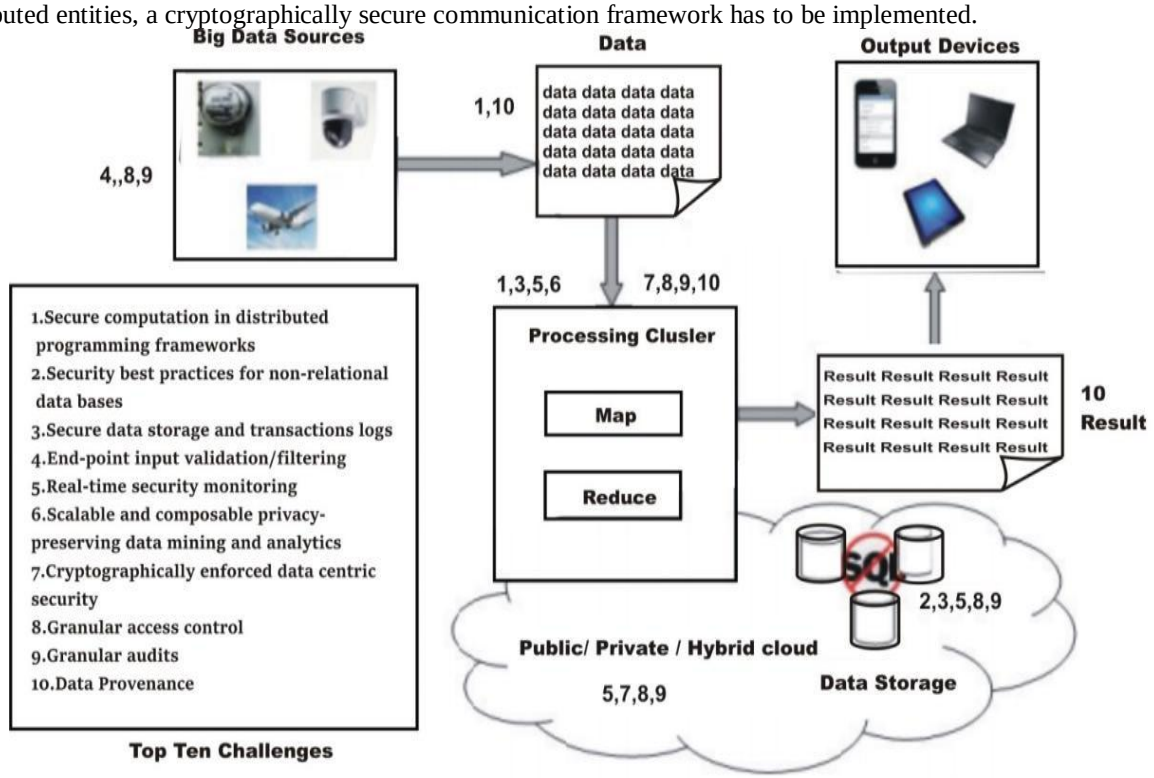

Fig 4.Top Ten Challenges in Big Data

8. Granular Access Control:
The security Property that matters from the perspective of access control is secrecy-preventing access to data by people that should not have access .The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy.

9. Granular Audits:
With real time security monitoring, we try to be notified at the moment an attack takes place. In reality, this will not always be the case (e.g., new attacks, missed true positives). In order to get to the bottom of the missed attack, we need audit information. This is not only relevant because we want to understand what happened and what went wrong ,but also because compliance, regulation and forensics reasons .in that regard ,auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily) distributed.

10. Data Provenance:
Provenance metadata will grow in complexity due to large provenance graphs generated from provenance- enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.

V. CHALLENGES IN BIG DATA ANALYSIS

1. Heterogeneity and Incompleteness:
When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes.

2. Scale:
Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, 9 there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

3. Timeliness:
The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

V. CONCLUSION

The paper is a systematic study of various security issues and challenges of Big Data analytics. Big Data is a very challenging research area. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFERENCES

[1] "Big Data: science in the petabyte era", Nature 455 (7209):1, 2008
[2] Douglas and Laney, "The importance of _Big Data: A definition" ,2008
[3] Agrawal, Amr El Abbadi et al.,"Big data and cloud computing: current state and future opportunities", Proceedings of the 14th International Conference on Extending Database Technology, ACM, Sweden, March 21-24, 2011

[4] http://dashburst.com/infographic/Big-data-volume-variety-velocity/
[5] http://www.wired.com/insights/2013/05/the-missing-vs-in-Big-data-viability-and-value/
[6] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications", Hangzhou, China: 2012,pp. 59 – 63, 21-24 Oct. 2012.
[7] Wie, Jiang, Ravi V.T and Agrawal G., "A Map-Reduce System with an Alternate API for Multi-core Environments", Melbourne, VIC: 2010,pp. 84-93, 17-20 May. 2010.International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014
[8] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions", JeJu Island: 2013, pp. 132-137, 11-12 Apr.2013.
[9] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications", IEEE International Conference, Silicon Valley, CA, pp. 32 – 37, Oct 6-9, 2013.
[10] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the Map Reduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, pp. 35 – 39, Apr 14-19, 2013.
[11] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." Open Cirrus Summit (OCS), 2012 Seventh,Beijing, pp. 48 – 52, Jun 19-20, 2012.
[12] Venkata Narasimha Inukollu , Sailaja Arsi1 and Srinivasa Rao Ravuri, "SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUDCOMPUTING", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, pp. 45-56, May 2014.
[13] Sachchidanand Singh, Nirmala Singh,"Big Data Analytics", International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India,2012.
[14] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li,"Big Data Processing in Cloud Computing Environments", 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), IEEE, pp. 17–23, 2012.
[15] Dan Garlasu, Virginia Sandulescu, Ionela Halcu,Giorgian Neculoiu, Oana Grigoriu, Mariana Marinescu and Viorel Marinescu, "A big data implementation based on Grid computing", 11th RoEduNet international Conference, pp. 1-4, 2013
[16] Xiaoxue Zhang, Feng Xu, "Survey of Research on Big Data Storage",12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES), IEEE, pp.76-80, SEPT.2013.
[17] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE Aerospace conference, pp. 1-7, March 2012.
[18] C. Byun, W. Arcand, D. Bestor, B. Bergeron, M. Hubbell, J. Kepner, A. McCabe, P. Michaleas, J. Mullen, D. O'Gwynn, A. Prout, A. Reuther,A. Rosa & C. Yee, "Driving Big Data With Big Compute", IEEE High Performance Extreme Computing (HPEC), Sep 10-12, 2012
[19] "Top Ten Big Data Security And Privacy Challenges", CLOUD SECURITY ALLIANCE, NOV. 2012, https://cloudsecurityalliance.org/
[20] Tyson Condie , Paul Mineiro , Neoklis Polyzotis , Markus Weimer,"Machine Learning on Big Data", 29TH IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE Conference), pp. 1242-1244, 2013
[21] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,"Data mining with Big Data", IEEE Transactions on Knowledge & Data Engineering, vol.26, Issue No.01, Jan 2014.
[22] Dr. Sun-Yuan Kung,"From Green Computing to Big-Data Learning: A Kernel Learning Perspective"IEEE 24th International Conference on Application-Specific Systems, Architectures and Processors (ASAP), USA, JUNE 2013

