Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Explorative Relational Analysis of Big Data with Cloud Computing - New Research Direction

**Vikram Deep Aulakh***
A.P., Computer Science and Engineering, G.G.G.I, Dinarpur, Ambala, Kurukshetra University, Haryana, India

***Abstract*— In *the present scenario the cessation users are more concern about how to organize their data, how to label different kinds of it (structured, unstructured, semi-structured, internal and external), which technologies one uses to store it and retrieve. Sizably voluminous Data is a broader concept then just data that transpire to be sizably voluminous. Massive magnification in the scale of data or sizably voluminous data engendered through cloud computing has been observed. Addressing astronomically immense data is a challenging and time- authoritatively mandating task that requires an astronomically immense computational infrastructure to ascertain prosperous data processing and analysis. The elevate of immensely colossal data in cloud computing is reviewed in this study. The definition, characteristics, and relegation of sizably voluminous data along with some discussions on cloud computing are introduced. The relationship between immensely colossal data and cloud computing, astronomically immense data storage systems, and Hadoop technology are additionally discussed. Furthermore, research challenges are investigated, with fixate on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, licit and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized.. On the other hand, many cloud applications, i.e. distributed systems, have an abundance of expeditious data to track, typically as transforming immutable event streams and pre-aggregating them into OLAP stores for later query. Anyhow, cloud applications need to scale in their database back end as well, and that is where NewSql, BigSql, NoSql and Expeditious Data (a variation of Sizably Voluminous Data) are in play. This paper fixates on such an immensely colossal intersection between cloud applications and sizably voluminous data. As a conclusion, Astronomically immense Data represents content and Cloud Computing is infrastructure.***

***Keywords*— Big data, Cloud computing, Hadoop, distributed computing, Bigtable, mapreduce***

## I. INTRODUCTION

The perpetual increase in the volume and detail of data captured by organizations, such as the elevate of gregarious media, Internet of Things (IoT), and multimedia, has engendered an inundating flow of data in either structured or unstructured format. Data engenderment is occurring at a record rate [1], referred to herein as sizably voluminous data, and has emerged as a widely apperceived trend. Astronomically immense data is eliciting attention from the academia, regime, and industry. Immensely Colossal data are characterized by three aspects: (a) data are numerous, (b) data cannot be categorized into conventional relational databases, and (c) data are engendered, captured, and processed rapidly. Moreover, sizably voluminous data is transforming healthcare, science, engineering, finance, business, and eventually, the society. The advancements in data storage and mining technologies sanction for the preservation of incrementing amounts of data described by a transmutation in the nature of data held by organizations [2]. The rate at which incipient data are being engendered is staggering [3]. A major challenge for researchers and practitioners is that this magnification rate exceeds their competency to design opportune cloud computing platforms for data analysis and update intensive workloads. Cloud computing can not only minimize the cost and restriction for automation and computerization by individuals and enter- prises but can withal provide reduced infrastructure maintenance cost, efficient management, and utilizer access.

## II. LITERATURE SURVEY

Many researchers have suggested that commercial DBMSs are not congruous for processing profoundly and astronomically immense scale data. Classic architecture's potential bottleneck is the database server while faced with peak workloads. One database server has restriction of scalability and cost, which are two consequential goals of sizably voluminous data processing. In order to acclimate sundry immensely colossal data processing models, D. Kossmann et al. presented four different architectures predicated on classic multi-tier database application architecture which are partitioning, replication, distributed control and caching architecture [4]. It is pellucid that the alternative providers have different business models and target different kinds of applications: Google seems to be more fascinated with diminutive applications with light workloads whereas Azure is currently the most affordable accommodation for medium to sizably voluminous accommodations. Most of recent cloud accommodation providers are utilizing hybrid architecture that is capable of gratifying their authentic accommodation requisites. In this section, we mainly discuss astronomically

 

immense data architecture from three key aspects: distributed file system, non-structural and semi-structured data storage and open source cloud platform Google File System (GFS)[5] is a chunk-predicated distributed file system that fortifies fault-tolerance by data partitioning and replication. As an underlying storage layer of Google's cloud computing platform, it is utilized to read input and store output of MapReduce[6]. Similarly, Hadoop additionally has a distributed file system as its data storage layer called Hadoop Distributed File System (HDFS)[7], which is an open-source obverse of GFS. GFS and HDFS are userlevel filesystems that do not implement POSIX semantics and heavily optimized for the case of immensely colossal files (quantified in gigabytes)[8]. Amazon Simple Storage Accommodation (S3) [9] is an online public storage web accommodation offered by Amazon Web Accommodations. This file system is targeted at clusters hosted on the Amazon Elastic Compute Cloud server-on-demand infrastructure. S3 aims to provide scalability, high availability and low latency at commodity costs. ES2[10] is an elastic storage system of epiC[6], which is designed to fortify both functionalities within the same storage. The system provides efficient data loading from different sources, flexible data partitioning scheme, index and parallel sequential scan. In integration, there are general file systems that have not to be addressed such as Moose File System (MFS)[7], Kosmos Distributed Filesystem (KFS)[8].

## III. RELATIONAL ANALYSIS

Here are some of the reasons why Sizably voluminous Data and cloud are bundled together:i) Anywhere access - if your data sources are spread around the world you can utilize (public) cloud to sanction those sources more expeditious access to your storageii) Elasticity - if you require more storage to store the data a cloud platform can dynamically expand to accommodate your storage needs. If you don't require the storage anymore (which is authentically astute thing to do once you get the insights from the data) you can shrink it and don't pay for it anymore iii)Scalability as Gary mentions below, that sanctions you to process the data more expeditious than the traditional way. At the terminus of the day it always comes to cost. If you process the data on a traditional platform you require to provision the storage and compute upfront and keep them up and running aeonianly while with the cloud you provision only what you require (storage) and when you require it (processing). The fig 1 shows relationship of their existence.

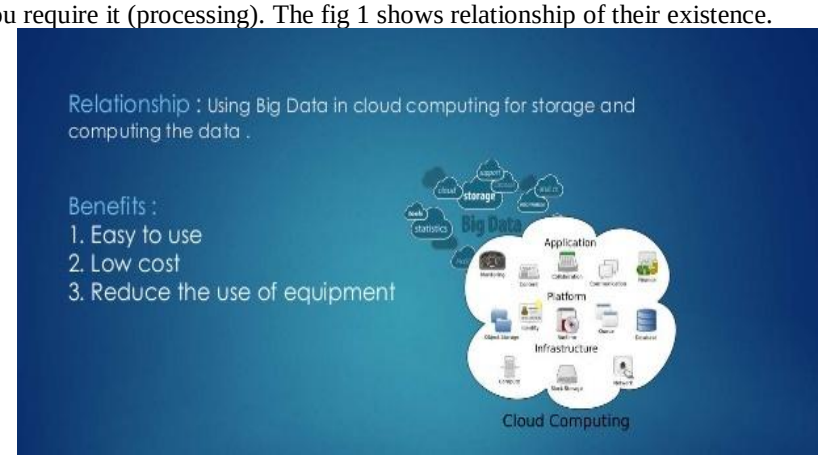


Fig 1: Relationship between Big Data and Cloud Computing

Big Data is all about extracting VALUE out of "Variety, Velocity and Volume" (3V) from the Information Assets available, while Cloud focuses on On-Demand, Elastic, Scalable, Pay-Per use Self Service models. The question often asked is then what is the relationship between Cloud and Big Data. Why are these two entirely different areas discussed together?

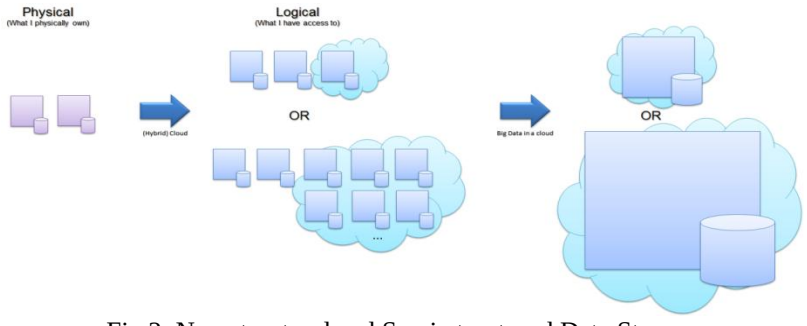


Fig 2. Non-structural and Semi-structured Data Storage

With the prosperity of the Web 2.0, more and more IT companies have incrementing needs to store and analyze the ever growing data, such as search logs, crawled web content, and click streams, customarily in the range of petabytes, amassed from a variety of web accommodations. However, web data sets are conventionally non-relational or less structured and processing such semi-structured data sets at scale poses another challenge likewise the above Fig 2 depicts the same scenario. Moreover, simple distributed file systems mentioned above cannot gratify accommodation providers like Google, Yahoo!, Microsoft and Amazon. All providers have their purport to accommodate potential users and own their pertinent state of- the-art of immensely colossal data management systems in the cloud environments. Bigtable [11]

 

is a distributed storage system of Google for managing structured data that is designed to scale to a prodigiously and sizably voluminous size (petabytes of data) across thousands of commodity servers. Bigtable does not fortify a full relational data model. However, it provides clients with a simple data model that fortifies dynamic control over data layout and format. PNUTS[12] is a massive scale hosted database system designed to fortify Yahoo! 'web applications. The main focus of the system is on data accommodating for web applications, rather than intricate queries. Upon PNUTS, incipient applications can be built very facilely and the overhead of engendering and maintaining these applications is nothing much. The Dynamo[13] is a highly available and scalable distributed key/value predicated data store built for fortifying internal Amazon's applications. It provides asimple primary-key only interface to meet the requisites of these applications. However, it differs from key-value storage system. Facebook proposed the design of an incipient cluster-predicated data warehouse system, Llama[14], a hybrid data management system which amalgamates the features of row-sagacious and column-sagacious database systems. They withal describe an incipient column-sapient file format for Hadoop called CFile, which provides better performance than other file formats in data analysis.

## IV. APPLICATION AREA

(The Apache Hadoop Framework and MapReduce) Incipient technologies are emerging to magnify data analytics possible and cost-efficacious .The Apache Hadoop* framework as shown in Fig.3 is evolving as the best incipient approach. The Hadoop framework redefines the way data is managed and analysed by leveraging the puissance of a distributed grid of computing resources. The Hadoop open-source framework [16] [17] [18] [19] utilizes a simple programming model to enable distributed processing of immensely colossal data sets on clusters of computers. The consummate technology stack includes prevalent utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management. In additament to offering high availability, the Hadoop framework is more cost- efficacious for handling astronomically immense, involute, or unstructured data sets than conventional approaches, and it offers massive scalability and celerity.

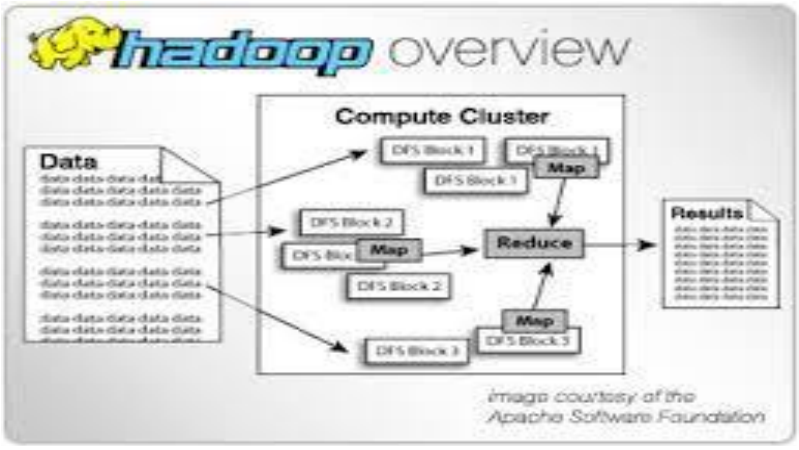


Fig: 3. The Apache Hadoop Framework

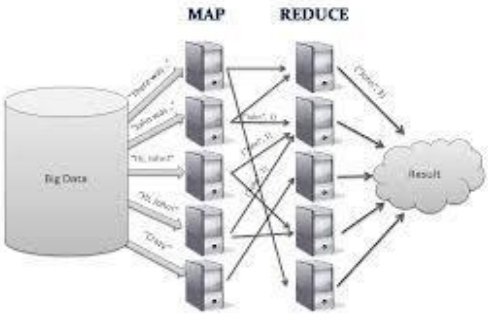


Fig: 4. MapReduce

MapReduce in Fig 4 withal is a popular platform in which the dataflow takes the form of a directed acyclic graph of operators. However, it requires lots of I/Os and dispensable computations while solving the quandary of iterations with MapReduce. Twister[20] proposed by J. Ekanayake et al. is an enhanced MapReduce runtime that fortifies iterative MapReduce computations efficiently, which integrates an extra Cumulate stage after Reduce stage. Thus, data output from coalesce stage flows to the next iteration's Map stage. It evades instantiating workers perpetually during iterations and antecedently instantiated workers are reused for the next iteration with different inputs. HaLoop[21] is kindred to Twister, which is a modified version of the MapReduce framework that fortifies for iterative applications by integrating a Loop control. It additionally sanctions to cache both stages' input and output to preserve more I/Os during iterations. There subsist lots of iterations during graph data processing. Pregel[22] implements a programming model incentivized by the Bulk Synchronous Parallel(BSP) model, in which each node has its own input and transfers only some messages which are required for the next iteration to other nodes.On a deeper look, the other patterns of this relationship emerge.

 

Cloud has glorified the "As-a-Accommodation" Model by obnubilating the intricacy and challenges involving in building a scalable elastic self-accommodation application. The same is the requisite for Sizably voluminous Data Processing. Hadoop in a kindred way obnubilates the intricacy of the astronomically immense scale distributed processing from the terminus utilizer perspective. The utilizer indite "Map-Reduce" programs or familiar kenned constructs with "Hive" or "Pig" and are able to seamless do the sizably voluminous data crunching without worrying about the involution of node failures, linear scalability, replication, fault-tolerance elasticity etc., where Hadoop silently provides the astronomically immense scale distributed capabilities abaft the scene. Thus the simplification provided by Cloud and Sizably Voluminous data is the prime reason for the mass adoption of Immensely colossal Data and Cloud. Amazon has just demonstrated how this simplification provided by the cumulation of Cloud and Immensely Colossal Data can increment the adoption of a ostensibly intricate quandary of sizably voluminous scale distributed processing. The key here is simplification only. Both Cloud and Astronomically immense Data is about distributing value to enterprise by lowering the cost of ownership. Cloud brings this through the Pay-per utilizer model turning CAPEX to OPEX while Apache open source has brought down the licensing cost of such a sophisticated solution ideally which would have cost millions to build and buy. Both Astronomically Immense Data and Cloud has been driving the cost down for the enterprise and bringing VALUE to enterprise. We have witnessed the early adopters of the Sizably voluminous Data moving away from the Traditional Licensing Models to a more open-sourced model and thus lowering the overall Cost per Terabyte (TB) processing. Both Cloud and Astronomically Immense Data distributes value and the key is how supple the enterprises get to break the hurdles of enterprise open source adoption and jump into the Astronomically immense Data Journey. Cloud and Sizably Voluminous Data brings in data security and privacy concerns. This is where System Integrators has been building solutions that espouse Cloud and Sizably voluminous Data within the Enterprise to build Elastic Scalable Private Cloud Solution to bring in the same value which enterprises can leverage to bring a Scalable Distributed Processing in action within the enterprise. Again we could visually perceive the kindred attribute between Cloud and Astronomically Immense Data with deference to Security Concerns and how innovative solutions could drive these adoptions within the enterprise.

## V. THE BIG SOLUTION

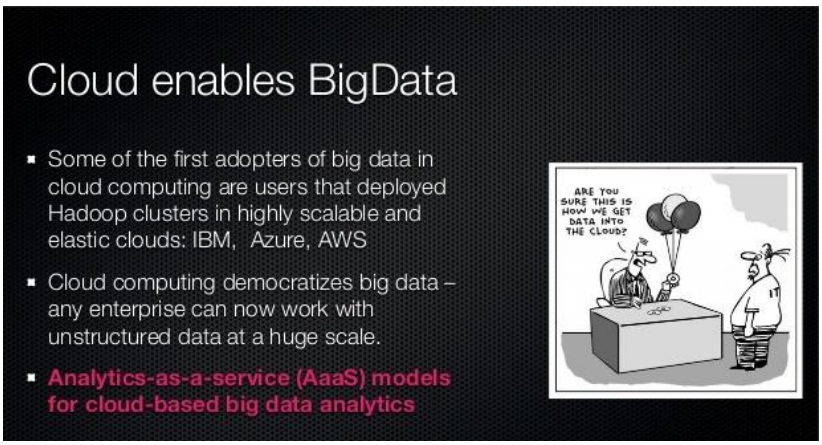


Fig 5: Cloud processing Big Data

A Hadoop cluster is a special type of computational cluster designed concretely for storing and analyzing immensely colossal amounts of unstructured data in a distributed computing environment.Analytics as a Accommodation (AaaS) refers to the provision of analytics software and operations through Web-distributed technologies. These types of solutions as shown in Fig 5 offer businesses an alternative to developing internal hardware setups just to perform business analytics. To put Analytics as a Accommodation in context, this type of accommodation is a component of a much wider range of accommodations with kindred names and homogeneous conceptions, including: i)Software as a Accommodation (SaaS) ii)Platform as a Accommodation (PaaS) iii) Infrastructure as a Accommodation (IaaS)

What these all have in prevalence is that the accommodation model supersedes internal systems with Web-distributed accommodations. In the example of Analytics as a Accommodation, a provider might offer access to a remote analytics platform for a monthly fee. This would sanction a client to utilize that particular analytics software for as long as it is needed, and to stop utilizing it and stop paying for it at a future time. Analytics as a Accommodation is becoming a valuable option for businesses because establishing analytics processes can be a work-intensive process. Businesses that need to do more analytics may need more servers and other kinds of hardware, and they may need more IT staff to implement and maintain these programs. If the business can utilize Analytics as a Accommodation instead, it may be able to bypass these incipient costs and incipient business process requisites.

## VI. FUTURE DIRECTION

We are awash in a flood of data today. In a broad range of application areas, data is being accumulated at unprecedented scale. Decisions that anteriorly were predicated on guesswork, or on painstakingly constructed models of authenticity, can now be made predicated on the data itself. Such Sizably voluminous Data analysis now drives

 

proximately every aspect of our modern society, including mobile accommodations, retail, manufacturing, financial accommodations, life sciences, and physical sciences. Scientific research has been revolutionized by Immensely Colossal Data [23]. The Sloan Digital Firmament Survey [24] has today become a central resource for astronomers the world over. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and withal of engendering public databases for use by other scientists. Fortuitously, subsisting computational techniques can be applied, either as is or with some extensions, to at least some aspects of the Sizably voluminous Data quandary. For example, relational databases rely on the notion of logical data independence: users can cogitate what they optate to compute, while the system (with adept engineers designing those systems) determines how to compute it efficiently. Similarly, the SQL standard and the relational data model provide a uniform, potent language to express many query needs and, in principle, sanctions customers to optate between vendors, incrementing competition. The challenge ahead of us is to cumulate these salubrious features of prior systems as we devise novel solutions to the many incipient challenges of Immensely Colossal Data.

***A) Data Acquisition and Recording***

Immensely colossal Data does not arise out of a vacuum: it is recorded from some data engendering source. For example, consider our faculty to sense and observe the world around us, from the heart rate of an elderly denizen, and presence of toxins in the air we breathe, to the orchestrated square kilometer array telescope, which will engender up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can facilely engender petabytes of data today.

The second astronomically immense challenge is to automatically engender the right metadata to describe what data is recorded and how it is recorded and quantified. The second sizably voluminous challenge is to automatically engender the right metadata to describe what data is recorded and how it is recorded and analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form felicitous for analysis. Doing this correctly and plenarily is a perpetuating technical challenge.

***B) Information Extraction and Cleaning***

Frequently, the information accumulated will not be in a format yare for analysis. For example, consider the amassment of electronic health records in a hospital, comprising transcribed dictations from several medicos, structured data from sensors and quantifications (possibly with some associated dubiousness), and image data such as x-rays. We cannot leave the data in this form and still efficaciously

analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form congruous for analysis. Doing this correctly and thoroughly is a perpetuating technical challenge

***C) Data Integration, Aggregation, and Representation***

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure.

We must enable other professionals, such as domain scientists, to engender efficacious database designs, either through devising implements to avail them in the design process or through forgoing the design process plenarily and developing techniques so that databases can be used efficaciously in the absence of perspicacious database design.

***D) Query Processing, Data Modeling, and Analysis***

Methods for querying and mining Immensely colossal Data are fundamentally different from traditional statistical analysis on diminutive samples. Immensely Colossal Data is often strepitous, dynamic, heterogeneous, inter-cognate and untrustworthy. Nevertheless, even strepitous Immensely colossal Data could be more valuable than minute samples because general statistics obtained from frequent patterns and correlation analysis conventionally inundate individual fluctuations and often disclose more reliable obnubilated patterns and cognizance

***E) Interpretation***

Having the facility to analyze Immensely Colossal Data is of constrained value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This

interpretation cannot transpire in a vacuum. Customarily, it involves examining all the posits made and retracing the analysis. Furthermore, as we visually perceived above, there are many possible sources of error: computer systems can have bugs, models virtually always have posits, and results can be predicated on erroneous data. For all of these reasons, no responsible utilizer will cede ascendancy to the computer system. Rather she will endeavor to understand, and verify, the results engendered by the computer.

## VII. CHALLENGES IN RELATIONAL ANALYSIS

Having described the multiple phases in the Immensely colossal Data analysis pipeline, we now turn to some mundane challenges that underlie many, and sometimes all, of these phases. These are shown as five boxes in the second row of below Fig 6.

 

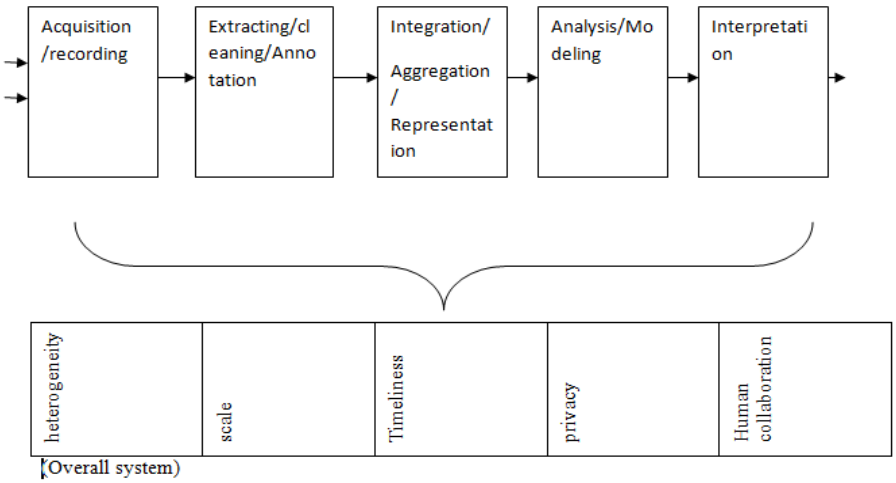


Fig 6: Relational analysis Pipeline

***A) Heterogeneity and Incompleteness***

When humans consume information, a great deal of heterogeneity is comfortably abode. 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 conscientiously structured as a first step in (or prior to) data analysis

***B) Scale***

Of course, the first thing anyone cerebrates of with Astronomically immense Data is its size. After all, the word "big" is there in the very denomination. Managing sizably voluminous and rapidly incrementing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting more expeditious, following Moore's law, to provide us with the resources needed to cope with incrementing volumes of data.

***C) Timeliness***

The flip side of size is celerity. The more astronomically immense the data set to be processed, the longer it will take to analyze. The design of a system that efficaciously deals with size is likely additionally to result in a system that can process a given size of data set more expeditious. However, it is not just this speed that is conventionally designated when one verbalizes of Velocity

***D) Privacy***

The privacy of data is another immensely colossal concern, and one that increments in the context of Immensely Colossal Data. For electronic health records, there are stringent laws governing what can and cannot be done. For other data, regulations, categorically in the US, are less forceful. However, there is great public fear regarding the incongruous utilization of personal data, concretely through linking of data from multiple sources. Managing privacy is efficaciously both a technical and a sociological quandary, which must be addressed jointly from both perspectives to realize the promise of immensely colossal data. In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can facilely detect but computer algorithms have an arduous time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs. Ideally, analytics for Immensely colossal Data will not be all computational – rather it will be designed explicitly to have a human in the loop. A popular incipient method of harnessing human ingenuity to solve quandaries is through crowd-sourcing. Wikipedia, the online encyclopedia, is perhaps the best kenned example of crowd-sourced data.

## VIII. CONCLUSION

This paper described a systematic flow of survey on the sizably voluminous data processing in the context of cloud computing. I have find out profound platform to discuss the key issues, including cloud storage and computing architecture, popular parallel processing framework, major applications and optimization of MapReduce. Immensely colossal Data is not an incipient concept but very arduous. It calls for scalable storage index and a distributed approach to retrieve required results near authentic-time. It is a fundamental fact that data is too astronomically immense to process conventionally. Nevertheless, immensely colossal data will be intricate and subsist perpetually during all sizably voluminous challenges, which are the astronomically immense opportunities for us. In the future, consequential challenges need to be tackled by industry and academia. It is an exigent need that computer philomaths and convivial sciences philomaths make close cooperation, in order to assure the long-term prosperity of cloud computing and collectively explore incipient territory.

**REFERENCES**

[1] R.L.Villars,C.W.Olofson,M.Eastwood,Bigdata:whatitisandwhy you shouldcare,WhitePaper,IDC,2011,MA,USA.
[2] R. Cumbley,P.Church,IsBigDatacreepy?Comput.LawSecur.Rev. 29 (2013)601–609.

 

[3] S.Kaisler,F.Armour,J.A.Espinosa,W.Money,BigData: Issues and Challeng Moving Forward, System Sciences(HICSS), 2013,in:Proceedings of the46th Hawaii International Conference on, IEEE, 2013,pp.995–1004.

[4] D. Kossmann, T. Kraska, and S. Loesing, "An evaluation of alternative architectures for transaction processing in the cloud," in Proceedings of the 2010 international conference on Management of data. ACM, 2010, pp. 579–590.

[5] S. Ghemawat, H. Gobioff, and S. Leung, "The google file system," in ACM SIGOPS Operating Systems Review, vol. 37, no. 5. ACM, 2003, pp. 29–43.

[6] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[7] D. Borthakur, "The hadoop distributed file system: Architecture and design," Hadoop Project Website, vol. 11, 2007.

[8] A. Rabkin and R. Katz, "Chukwa: A system for reliable large-scale log collection," in USENIX Conference on Large Installation System Administration, 2010, pp. 1–15.

[9] S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," Communications Surveys & Tutorials, IEEE, vol. 13, no. 3, pp. 311–336, 2011.

[10] Y. Cao, C. Chen, F. Guo, D. Jiang, Y. Lin, B. Ooi, H. Vo, S. Wu, and Q. Xu, "Es2: A cloud data storage system for supporting both oltp and olap," in Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, 2011, pp. 291–302.

[11] F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A distributed structured data storage system," in 7th OSDI, 2006, pp. 305–314.

[12] B. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H. Jacobsen, N. Puz, D. Weaver, and R. Yerneni, "Pnuts: Yahoo!'s hosted data serving platform," Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1277–1288, 2008.

[13] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin,S.Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available keyvalue store," in ACM SIGOPS Operating Systems Review, vol. 41, no. 6. ACM, 2007, pp. 205–220.

[14] Y. Lin, D. Agrawal, C. Chen, B. Ooi, and S. Wu, "Llama: leveraging columnar storage for scalable join processing in the mapreduce framework," in Proceedings of the 2011 international conference on Management of data. ACM, 2011, pp. 961–972.

[16] a b c d e f "Data, data everywhere". The Economist (25) 2010 Retrieved (9) 2012

[17] "E-Discovery Special Report: The Rising Tide of Nonlinear Review". Hudson Global Retrieved 2012.by Cat Casey and Alejandra Perez

[18] "What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology — Re-Humanizing Technology-Assisted Review" Forbes. Retrieved 2012

[19] Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner Retrieved 2001

[20] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. Bae, J. Qiu, and G. Fox, "Twister: a runtime for iterative mapreduce," in Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, 2010, pp. 810–818.

[21] Y. Bu, B. Howe, M. Balazinska, and M. Ernst, "Haloop: Efficient iterative data processing on large clusters," Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 285–296, 2010.

[22] G. Malewicz, M. Austern, A. Bik, J. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: a system for largescale graph processing," in Proceedings of the 2010 international conference on Management of data. ACM, 2010, pp. 135–146.

[23] Advancing Discovery in Science and Engineering. Computing CommunityConsortium. Spring 2011.

[24] SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. Jan. 2008. Available at http://www.sdss3.org/collaboration/description.pdf

 