High Performance Computing and Big Data Analytics – Paradigms and Challenges

Tulasi.B
PhD Scholar, Computer Science Jain University, Bangalore

Rupali Sunil Wagh
PhD Scholar, Computer Science Jain University, Bangalore

Balaji S
Center for Emerging Technologies
Jain University, Bangalore

ABSTRACT
The advent of technology has led to rise in data being captured, stored and analyzed. The requirement of improving the computational models along with managing the voluminous data is a primary concern. The transition of the High Performance Computing from catering to traditional problems to the newer domains like finance, healthcare etc. necessitates the joint analytical model to include Big Data. The rise of Big Data and subsequently Big Data analytics has changed the entire perspective of data and data handling. Ever growing analytical needs for Big Data can be satisfied with extremely high performance computing models. As a result of enormous research in this field, recent years have seen the emergence diverse paradigms for Big Data analytics. With the spread of Big Data analytics in varied domains, newer concerns regarding the effectiveness of analytical paradigms are also observed. This paper highlights the major analytical models and concerns and challenges in High Performance Data Analytics.

Keywords
High Performance Analytics, Big Data Analytics, Stream Analytics, Converging Paradigms, Computational Models.

1. INTRODUCTION
Technology has played a vital role in increasing the growth of digital data. It has also eased the means of communication. With the penetration of technology into all sectors, there has been a surge of data available in digital form. The resources from which data can be obtained have dramatically increased with the advent of technology. Sensors, log files, mobile devices, remote sensing technologies wireless sensor networks, and social networking are few resources from which data is being captured. The by-product of internet, "exhaust data" is seen predominantly increasing at an exponential rate. This data is not directly derived data rather is obtained as a consequence of various activities performed on the network. Advances in mobile technology are the vital triggers in the generation "Digital Exhaust Data" in larger capacity [1]. Platforms like social networking have provided trails of digital data which cannot be put into the traditional structure of storage. The varied nature of data is prevalent and adds a new dimension to data storage. The various forms of the huge data along with seasonal and increased flow cannot be handled by traditional models. This data which is voluminous and varied, when analyzed provides deeper insights, adds value to the data itself. This data which is identified by multiple dimensions is termed as "Big Data".

The growth of data has been on steady increase and according to IDC's Digital Universe Study sponsored by EMC in June 2011, 130 Exabyte's of data were created and stored in 2005. The amount grew to 1,227 Exabyte's in 2010 and is projected to grow at 45.2% to 7,910 Exabyte's in 2015 [2]. If growth of the digital data is on ascend, the type of the data also is a concern. The digital data is in multiple forms. A major segment of this digital data tends to fall under the category of unstructured data, data which cannot be represented by location in a computer record. Around 95% of this digital data is unstructured which cannot be handled by the traditional approaches [3]. The metadata associated with unstructured data facilitates data handling and processing. The universe of digital data is expanding and evolving. The characteristics of the digital data would be largely governed by the unstructured data. The tools, applications required for storing, processing and managing data need to be enhanced to cater to the diverse need of the digital data. Structured data can be handled using relational structures but the large chunks of data which are now available cannot be handled by the conventional methods. The significant challenges faced due the explosion of data is on data processing applications and management tools.

Data can drive continuous improvement in all the process of an organization. Hence Data Analytics is seen as the driving force in transforming the working of organizations in various sectors. Data driven decisions are facilitating organizations to adapt to the dynamic needs of industry, optimize resources and handle various other issues effectively. Penetration of technology has created enormous data resources which when appropriately used would alleviate the "data driven" decision making process. The form of analytics called "digital analytics" has evolved to analyze both qualitative and quantitative forms of digital data to drive better outcomes [4]. With the availability of large and varied data sets, computation of them is a subject of concern. The traditional computing approaches would not be able to meet the requirements of this large volume of data. Large scale data analytics has been synonymous to High Performance Computing (HPC) has been the driving force for expansions in HPC. But the technical advancement implores the need to deploy new forms of HPC to mine deeper insights in the large volumes of data available. With boundaries between HPC and Big Data becoming hazed, a new term has been coined "High Performance Data Analytics" (HPDA) [5]. HPC has been traditionally employed to solve large computational problems related to science and technology like astronomy, climatic modelling among others. HPDA moves to integrate HPC with Large scale data analytics and open source platforms like Hadoop. High performance computing is characterized by high speed networks, diskless nodes and parallel file systems. Large scale analytics are moving out from local file systems and disk nodes. To bridge the gap, there is a requirement to enhance Hadoop like systems to allow data processing from HPC clusters. One of the biggest challenges faced by HPDA is to obtain actionable patterns or intelligence which would

coerce data driven decisions into a higher level. This paper identifies the various dimensions of Big Data Analytics along with high Performance Computing. The next section talks about the evolution of Big Data analytics covering the aspects of evolution like deep learning analytics and stream analytics. The third section discusses the convergence of the two paradigms and the last section focuses on few of the challenges.

2. BIG DATA AND EVOLUTIONS IN DATA ANALYTICS

Today Big data has become the buzzword in the IT industry. In contrast to older data systems where transactional database is the primary source of data, in today's business and computing scenario, data can be obtained from heterogeneous sources like sensor data, location data, data from internet etc. As mentioned in the previous sections, most of the data to be analyzed is unstructured and complex. With emerging databases technologies like spatial database and mobile database, the traditional SQL required transformations to adapt to newer type of data representation. The representation of data and its management has radically changed to incorporate the newer type of data [6]. Databases such as NoSQL, MongoDB provide suitable ways of representation of unstructured and more complex data and its management and attempt to cater to the "variety" dimension of big data. Challenges with respect to many types of data such as sensor data, network data, web generated data and text data are still a concern in spite of continuous advancements in data analytics for big data.

2.1 Large Scale Data Analytics

Data analytics and Data science, the branches of applied mathematics are catering to the information and knowledge needs of all the domains. Analytics primarily can be classified into three different levels descriptive, predictive and prescriptive. While descriptive analytics provides details of data distribution and visualization, predictive analysis mainly emphases on predicting trends and possibilities. Prescriptive analytics is aimed at providing insights from data that could be used in decision making. Descriptive analysis is based on visualization and is widely used in business intelligence. Predictive analytics uses statistical models whereas sophisticated machine learning techniques and simulations form the basis for prescriptive analysis to provide optimal solutions.

Traditional analytics paradigm of "store today analyse later" is clearly not enough for today's big data. Any typical big data application has following generic phases – data generation, data acquisition, data storage and data analytics [7]. Abundant availability of data from varied and distributed data sources makes processes in these phases very complex. The three Vs, volume, velocity and verity of data demand for additional approaches for effective analytics.

2.2 Batch Processing vs. Stream Processing

Traditional batch processing approach with scaling out of computing resources though is considered as most natural but it comes at the cost of dormancy [8]. Data is transmitted across the web very fast. High velocity is one of the key characteristic of big data which require distinctive computing paradigms and a supporting infrastructure. The pace of data introduces the concept of freshness of data and its importance in the context of the business value of the data. Data applications that require real time processing and analytics are based on the stream based approach. This approach is aimed at online data analytics where the computations are performed on continuous arriving data streams. Batch processing gives precise computational Analytics results at the cost of latency whereas stream processing is based on approximation but preserves the freshness of data.

2.3 Deep Learning analytics for Big Data

Big data application architectures are becoming increasingly complex. Applications of machine learning algorithms to extract patterns for prescriptive analytics on such complex data models is the challenging fields of large scale data analytics [9]. Deep belief networks and convolution network neural network are the examples of deep learning paradigms for big data. Obviously training becomes very complex due to the significant increase in the number required layers in the network. High performing computing framework is one of the most apparent requirements for deep learning of big data.

2.4 Domain Dependent Data Analytics

Advancements in data capturing technologies coupled with emerging trends in networking has resulted into emergence of many newer domains as probable beneficiaries of big data analytics. Mobile data analytics, sensor data analytics, web analytics, network data analytics, scientific data analytics are seen as the subfield of large scale data analytics and require specialized analytics perspectives. Since this data is not conventional structured data, different analytical paradigms and efficient computing requirements are essential.

Analytical models are undergoing continuous progression to satisfy changing needs of large scale data analytics. Efficient storage and retrieval systems along with other high performance computing infrastructure play a crucial role in the effectiveness of large scale data analytics. The need for distinctive and high performance computing environments is very much apparent from the above discussion. Today, when most of the business decisions are data driven, High Performance Computing solutions for big data analytics have become more conspicuous than ever before.

3. HIGH PERFORMANCE COMPUTING AND DATA INTENSIVE APPLICATIONS

High Performance Computing has been associated with applications of tremendous computational needs. The capabilities of High Performance Computing (HPC) in computing and storage have not been explored to its fullest extent for predictive analytics. The design of high performance computing provides maximum scalability and performance which can be applied to problems which involve large volumes of varied data.

3.1 Spear Head in Large Scale Data Management - Big Table

Storage has been one of the focal points when large scale computations are concerned. Big table has been one of the earliest models of storage being considered for the distributed storage system of structured data. Data which is voluminous and the required of real time data handling, big table has been able to provide a high performance solution to them [10]. Big Table is used in clusters by various products of Google like Google Analytics. Though Big Table is compared to parallel databases and main memory databases, it provides divergent framework as it is not relational database. It provides a simple data model which allows dynamic control on data. It allows indexing using column and row names. It considers data as

uninterpreted strings. The schema of Big Table allows clients to serve data from disk or memory.

Data is organized in three dimensions rows, columns and timestamps. Rows with subsequent row keys are grouped into tablets. These tablets form the unit of load balancing. The columns are grouped into families for access control. Timestamps are for the different values stored in a cell. It is associated with the real time or by any other logic. Big table uses Goole File System (GFS), a distributed file which provides greater transparency about the multiple replicas of data and reliability through it.[11] Chubby, an efficient distributed lock service helps in storing the bootstrap details of the Big Table, schema. Though Big Table was initially conceptualized for storing structured data the possibility of storing unstructured and semi structured does exist. The uninterpreted string form of data can be explored for it.

3.2 Convergence of Paradigms

With rise in data intensive applications, there is also a requirement for adopting different approaches and solutions for the effective data management and computational needs. The two of the existing approaches: Big Data Stack built on Hadoop and High Performance Computing can be converged to provide a feasible solution to the challenges faced by these applications. There are two important dimensions of this congregation of paradigms (1) The issue of different architectures associated and (2) The relevance of data with different, varied features. This is depicted in the Figure1.

. Both HPC and Hadoop were designed to support different types of data and workflow. Due to the rise in the data intensive applications and the varied needs of them, the evolution of both the paradigms is required. More compute demanding workloads are being put on Hadoop clusters which are again of heterogeneous in nature. This has led to improvement in Hadoop with introduction of YARN and Mesos. Similarly to handle large scale data parallel file systems need to be evolved to support the data management issue. Parallel file systems like Lustre file system can be integrated with Hadoop for providing an efficient solution to the data intensive applications which are increasing in number. [5] This would move the local storage scenario of Hadoop towards a distributed file system. It would add on the leverage to the paradigm where in the features of distributed file systems are embedded. Increased throughput, sustainable storage performance and distributed shared memory would provide a definite benefit to the processing of the applications.

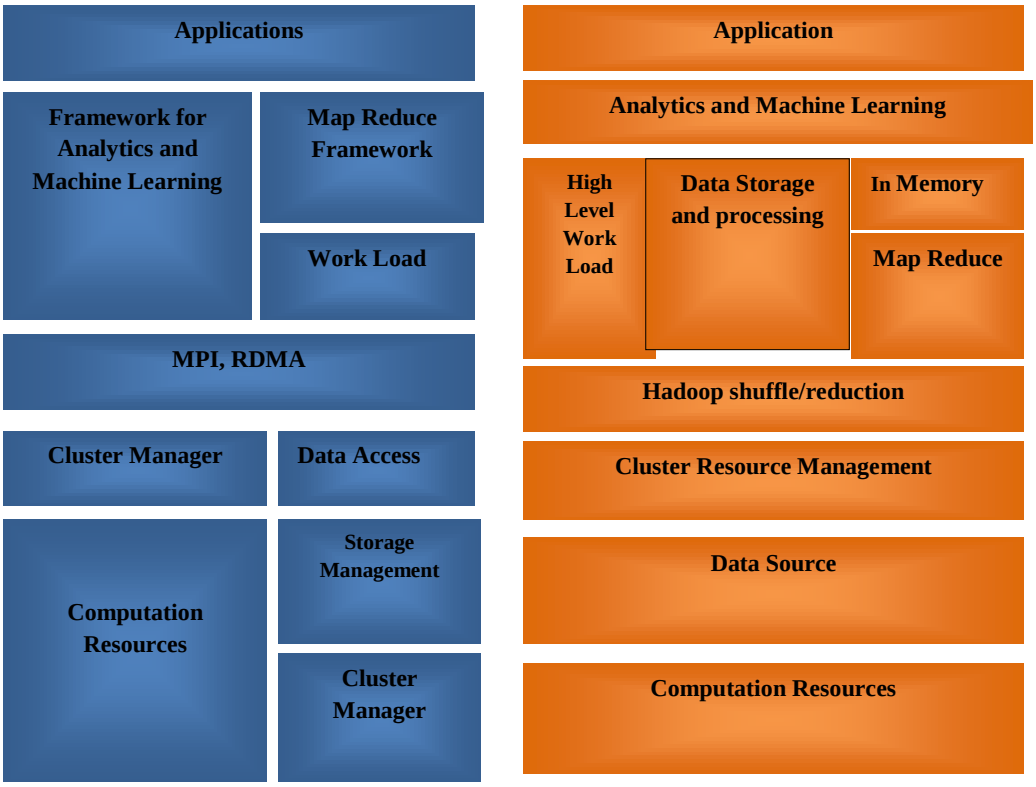

Figure 1: Architectural Comparison of HPC and Big Data

3.3 Towards Resource Efficient Big Data Analytics - Analytics as a service

Along with big data analytics, another technology which has revolutionized the industry is cloud computing. Cloud technology, which has its roots in grid computing, was essentially developed to provide variety of services with an objective of efficient resource sharing. Big data and big data analytics has been leveraging upon high performance computing infrastructures. With the emergence of cloud era, data center aggregation and extensive sharing of computing resources have contributed to significant improvements in computation time and cost reduction [12]. With the evolving paradigms in big data analytics the dominance of cloud technology has increased profusely. The success of IaaS, infrastructure as a service, Saas, software as a service and Paas, platform as a service has opened up radically different approaches to data handling. There is a shift in the way organizations are dealing with their data. Data intensive domains like genomics, scientific computing, and weather

forecast which require HPC for their analytics have been moving to cloud computing paradigms. High Performance Computing as a service, HPCaaS provide domain specific customizations of cloud environments with the help of various plugins. [13]. Cloud computing is emerging as a tool to exploit the capabilities of HPC to the fullest. Distributed and parallel computation frameworks such as Hadoop, and the scalable big data storage, management and query tools make the cloud an excellent platform for analytics. More and more organizations are adapting cloud platforms for more efficient business intelligence which accelerates the concept of having well-defined analytical service interfaces as well. Analytics as a service is seen as the upcoming cloud technology aimed at providing comprehensive big analytics solution encompassing data integration, data cleaning, query processing and even application of machine leaning based predictive analytics [14]. Examples of analytics as a service could be Risk analytics for financial transactions or the likelihood that lacking inventory will delay production given forthcoming weather patterns Conceptual framework for analytics as a service is shown in figure 2. Analytics as a service can be seen as convergence of big data with the benefits of cloud technology for efficient utilization of high end analytical capability of huge heterogeneous data to get deeper insights. Analytics as a Service is a confluence of the cloud computing which facilitates effective on demand features and big analytics to provide meaningful insights.

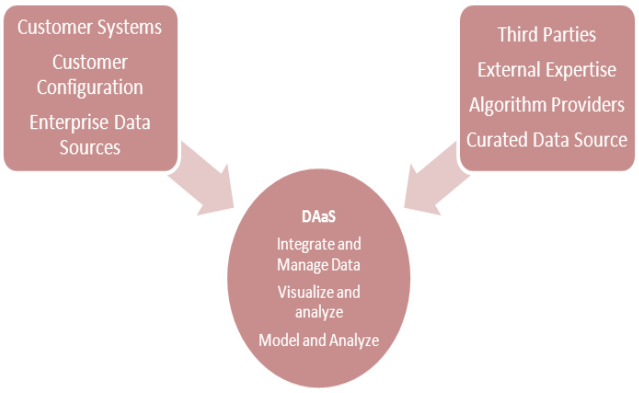

Figure 2: Framework for Analytics as a Service

With the freshness of data becoming one of the decisive factors for its business significance, real time analytics of data without any need to transfer it across locations for further analysis is looked for by organizations. Continuous analytics as a service is a variation of analytics as a service which caters to continuous or stream data analytics [15].

4. CHALLENGES

Big data analytics essentially requires very high computing capabilities in order to drive data into meaningful insights. High performance data analytics, HPDA, seeks to widen the HPC and big data analytics application domains by augmenting with other related technologies. However, varied and complex requirements of big data analytics pose many challenges at micro as well as macro level. At micro level, there are unusual and specific issues about statistical modelling of big data. At macro level big data analytics is challenged by the complexities in effective computational prototypes.

4.1 Challenges in Statistical Modeling of Big Data

Appropriate statistical model is the foundation for effectiveness of computational approach. Statistical modelling is the very first step in designing of any analytical solution. The specific shortcomings of traditional and very successful statistical data models for big data analytics are discussed in [16]. The complexity of mixture population model is considered to be a challenge considering the heterogeneity of data. Calculation of appropriate statistical metrics in and across the subpopulation of mixture model remains a challenge. High dimensionality and sparse nature of big data analytics adds on to further complications like noise accumulation due to large number of dimensions. High dimensionality also results in other undesirable conditions like spurious correlations and incidental endogeneity which direct negative impacts on results of analytics. Most of the predictive analytics algorithms are originated in statistical models. Above mentioned issues related to statistical big data modeling pose challenges with respect to the accuracy of the analytics.

4.2 Effective Computational and analytics Models

Due to the convergence of various technologies, computational models for big data analytics are continuously evolving. For the optimum utilization of the HPC made available through resource effective cloud environment, stronger computational models are required. The penetration of varied domains in big data analytics, there is an increasing need for most generic to very specific analytical paradigms. Two major features which contribute to the efficiency and effectiveness of big data analytical solution are briefly discussed below.

4.2.1 Parallel Computational Models

Parallel computing paradigm can be considered as the backbone for all big data analytics which targets at maximum resource utilization thereby delivering significant time efficiency improvements. Map reduce is considered to be the de facto solution big data handling and has been dominating the big data world. Researchers are exploring the possibilities of enhancing this most popular parallel programming model [17]. Massively parallel programming frameworks for big data analytics can be can be instrumental in providing solutions to storage and communication bottlenecks. Heterogeneity and un-structuredness of the data requires appropriate parallel computing models. While there are efforts to bring in parallelism while processing totally unstructured data like text, cost and time efficient parallel computing for big data analytics is still far from maturity.

Deep learning of big data is an emerging field which aims at application of machine learning algorithms for getting meaningful insights from big data. These applications are inherently parallel in nature. As mentioned earlier the sheer volume of data poses many challenges in application of deep big data analytics [9]. Visualization support is one of the primary objectives of analytics which can also be achieved efficiently only with the help of appropriate parallel processing. Design of effective massively parallel systems for high volume, variety and high velocity data is extremely complex and remains a challenge in big data analytics.

4.2.2 Computational Support for Continuous Analytics

High velocity of data has given rise to the concept of "freshness" of data. To analyse continuously and real time is becoming increasingly important in competitive business environment. Due to the paradigm shift from "store today analyse later" to "stream or real time analytics", computational models are undergoing revolutions. Continuous or stream analytics is complicated due to following requirements

- Continuous analytics does not work in isolation; the results have to be integrated with stored data analytics.
- Continuous analytics is sliding window based data analytics which should be supported by appropriate data staging.
- Scaling out continuous data analytics with parallel and distributed infrastructure.
- Real time visual analytics.

Continuous analytics does not come as default component but a separate real time analysis engine is required for this special need. The framework of real time analytics is shown in figure 3.

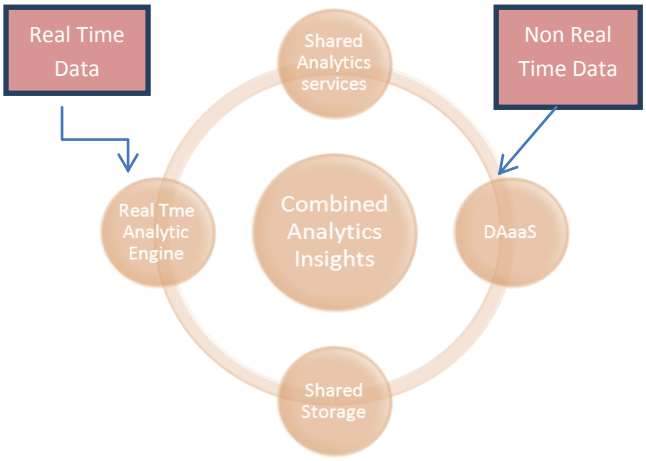

Figure 3: Integrated Real Time Analytics Framework

4.3 Data Capture and Storage

Present computing applications have unique high volume data processing and high computing requirements. The degree of "data intensive" applications is on ascent along with the "real data computing" need. The need for effective storage and retrieval methodology is principal in nature. It is the bottleneck and also the imperative factor which would enhance the computing systems to handle the evolving problems. Big Data has revolutionized the storage architecture and access mechanism. Data availability is the main concern in knowledge Discovery process. Though storage technologies like storage state device (SSD) and Phase Change Memory (PCM) are being developed they have not been able to cater to the requirement effectively. Technologies need to be redesigned to cater to the high performance requirement of data intensive applications.

Storage Area Network (SAN), Direct-Attached Storage (DAS), Network-Attached Storage (NAS) are the architectures which are looked upon for these challenges. The concurrency coupled with throughput is the issue which becomes the limitation to these architectures. To improve data intensive computing, optimization of data access is a significant factor to be considered. Distributed data centric storage is seen as the viable solution to the large scale data rich applications [18]. It is one of the canonical forms of data storage retrieval being used in distributed systems. The mapping of storage and searching techniques in distributed data centric storage are efficient spatial temporal similarity search schemes which can be utilized for data centric applications.

5. CONCLUSION

Information driven economy relies on the actionable insights extracted from data analytics. The era of data revolution has paved way to the need of convergence of paradigms like High Performance Computing and Big Data Analytics. The amalgamation of these paradigms is a herculean task involving various aspects like data management and computing efficiency. This has given rise to evolution of the data storage technologies and computing models. The transformation of traditional analytical paradigms to cater to the requirement of the data intense applications and High Performance Computing is the need of hour. The convergence of the paradigms "High Performance Computing" and "Big Data Analytics" can lead to a sustainable solution for the data driven applications. The continuous flow of "real" data which is the predominant type of data seen in data intense applications needs to be handled by a different architectural platform termed as "Real Time Analytical Framework". The computational requirements of these newer models are different from the traditional models and hence the evolution of the models becomes the key challenge of High Performance Data Analytics.

6. REFERENCES

[1] Big Data: The next frontier for innovation, competition and productivity. James Maniyka, Executive summary ,McKinsey Global Institute ,May 2011, http://www.mckinsey.com/mgi/publication/big.data/MGI_big_data_exec_summary.pdf>.

[2] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf [Accessed on 2nd January 2015]

[3] Beyond the hype: Big data concepts, methods, and analytics,Amir Gandomi , Murtaza Haide, International Journal of Information Management 35 (2015) 137–144

[4] https://analyticsacademy.withgoogle.com/course01/assets/pdf/DigitalAnalyticsFundamentals-Lesson2.1TheimportanceofdigitalanalyticsText.pdf [Acesses on 27th December 2014]

[5] Big Data Meets High Performance Computing Intel® Enterprise Edition for Lustre* software and Hadoop combine to bring big data analytics to high performance computing configurations. http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-meets-high-performance-computing-white-paper.pdf

[6] Dilpreet Singh **and** Chandan K Reddy**,** *A survey on platforms for big data analytics*, *Journal of Big Data* 2014, **2**:8 doi:10.1186/s40537-014-0008-6

[7] Han Hu, Yonggang Wen, Tat T-Seng Chua, and Xuel Ong Li, *Toward Scalable Systems for Big Data Analytics:A Technology Tutorial*, *Digital Object Identifier 10.1109/ACCESS.2014.2332453*

[8] Thibaud Chardonnens, Philippe Cudre-Mauroux and Martin Grund ,*Big Data Analytics on High Velocity Streams: A Case Study*, Benoit Perroud, 2013 IEEE International Conference on Big Data, 978-1-4799-1293-3/13/$31.00 ©2013 IEEE

[9] Xue-Wen Chen and Xiaotong Lin, *Big Data Deep Learning: Challenges and Perspectives*, *Digital Object Identifier 10.1109/ACCESS.2014.2325029*

[10] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Bigtable: *A distributed storage system for structured data, on Seventh Symposium on Operating System Design and Implementation*, 2006

[11] http://highscalaibility.com [Accessed on 6th February 2015]

[12] Thomas Sandholm and Dongman Lee, Notes on Cloud computing principles, Journal of Cloud Computing: Advances, Systems and Applications 2014

[13] Philip C. Church and Andrzej Goscinski, A Survey of Approches and Frameworks to Carry out Genomic Data Analysis on the cloud, 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2014 IEEE, DOI 10.1109/CCGrid.2014.127

[14] Farhana Zulkernine, Michael Bauer, Ashraf Aboulnaga, Patrick Martin and Ying Zou, Femida Gwadry-Sridhar, *Towards Cloud-based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud*, 2013 IEEE International Congress on Big Data

[15] Ascent / Data Analytics as a Service: unleashing the power of Cloud and Big Data, Published in March 2013

[16] Jianqing Fan, Fang Han and Han Liu, *Challenges of Big Data analysis*, *National Science Review* 1: 293–314, 2014, doi: 10.1093/nsr/nwt032

[17] *Parallel Data Processing with MapReduce: A Survey*, SIGMOD Record, December 2011 (Vol. 40, No. 4)

[18] M. Pavlovic, Y. Etsion, and A. Ramirez, *On the memory system requirements of future scientific applications: Four case-studies*, in Workload Characterization (IISWC), 2011 IEEE International Symposium on, 2011, pp. 159-170.

[19] Techniques and Challenges of Data Centric Storage Scheme in Wireless Sensor Network Khandakar Ahmed and Mark A. Gregory, Journla oof Sensor Netwroks,volume 1, Issue 1,59-85