IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 06, Issue 01 (January. 2016), ||V4|| PP 31-39

Computation of Big Data in Hadoop and Cloud Environment

Sheikh Ikhlaq, Dr. Bright Keswani
Department of Computer Sciences Suresh Gyan Vihar University Jaipur, India
Department of Computer Sciences Suresh Gyan Vihar University Jaipur, India

Abstract: - Big Data is a new phenomenon that needs high attention due to the wealth it possesses only if it is addressed and if it is not addressed it will lead to a situation of Data Explosion and a huge amount of garbage will be formed. Various technologies like Hadoop, Map Reduce, and NoSQL etc. are used but since these technologies are very costly and cannot be afforded by the mid and small size organizations and most of the governments. Therefore there is a need of technology which is cheap and could be afforded by everyone. Cloud computing is a viable option that can prove as a boon for everyone .Implementing Cloud computing for big Data computation is very difficult task to do and has various challenges associated with it. In this review paper we will review various Big Data methods, Approaches and how cloud computing is implemented with challenges posed by it being addressed. Also we will try to know what else needs to be done.

Keywords- *BigData, Cloud Computing, Hadoop, Map Reduce, NoSQL*

I. INTRODUCTION

Since the rise of digitisation, organizations from various areas have amassed huge amounts of digital data, capturing trillions of bytes of information, ranging from their customers, to suppliers and operations. Volume of data is also growing exponentially due to machine-generated data (data records, web-log files, and sensor data) and from growing human engagement within the social networks. Variety of data has also increased from text to audio, images and videos. The velocity at with this data is growing is tremendous. All these three V's i.e. Volume, Variety and Velocity have led to collection of mammoth amount of data termed as Big Data. The growth of data can never stop or for that matter it can't be restricted. According to IDC Digital Universe Study published in 2011, 130 Exabyte's of data was created and stored in 2005. The amount grew drastically to 1,227 Exabyte's in 2010 and it is projected to grow at 45.2% to 7,910 Exabyte's in 2015. This data can be used to do wonders extracting the hidden wealth of information inside it [1]. On the other side if this data is left as such it will be nothing but garbage. With this wealth come various challenges which don't include only scale, but also heterogeneity of data, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, data acquisition, result interpretation, Data management and Security. In order to address many of the above listed problems we have a solution termed as Cloud Computing.

Cloud has the services with the ability to ingest, store and analyse data are available from sometime to handle most of the challenges posed by Big Data. Early adopters of Big Data on the cloud are probably the users deploying Hadoop clusters on the highly scalable and elastic environments provided by Infrastructure-as-a-Service (IaaS) provider like Amazon Web Services, for test and development, and for analysis of available datasets. It offers data storage and data back-up in a cost-effective manner. It delivers a low-cost and reliable environment that gives organisations the computing resources which they need to store their data in both the forms i.e. structured and unstructured. On the other hand at Software-as-a-Service (SaaS) level, embedded analytics engines helps to analyse the data stored on the cloud. The analytical output is then provided to the end users through a graphical interface. The only requirement here is that the development of queries and integration to the data source on the cloud are prerequisites that organisations need to perform before the usability can be delivered. Hence it is understood that, cloud computing is able to provide the support for BigData deployment. But, when it comes to the transfer of data to and from these two environments, Privacy, Data management, Retrieval, Access Speed, Security and Integration issues can become major issues for the use of Big Data on the cloud.

II. DATA, INFORMATION: A NEED

Data is said to be distinct pieces of information, usually formatted in a special way. All software is composed into two general categories: data and program. Program is a collection of instructions for manipulating data. Data may exist in a variety of forms, as numbers, text, bits and bytes, or as facts that are stored in a person's mind. Data by definition is the plural of datum, a single piece of information. Generally, data is used both the singular and plural form of the word. The term data is often used so as to distinguish binary machine-readable information from textual human-readable information. There are two categories of data:

A. Structured Data:

Data which resides in a fixed field within a record or file is called as structured data. Data contained Relational databases and Spread sheets. Structured data needs to create a data– a model of the type data that is or will be recorded and how it will be stored, processed and then accessed. All this includes defining what fields of data is to be be stored and how this data will be stored: data type (numeric, currency, alphabetic, name, date, address) and with any restrictions on the data input (number of characters; restricted to certain terms such as Mr, Ms or Dr; M or F).The advantage structured data is that, of being easily entered, stored, queried and analysed. At one time, because of certain condition's which include high cost and performance limitations of storage, memory and processing, relational databases (also Spread Sheets) using structured data were the only solution to effectively manage data.

B. Unstructured Data:

The term "unstructured data" often refers to the information that doesn't reside in a traditional row-column (Relational) database. It's quite the opposite of structured data (the data stored in fields in a database). Unstructured data files usually consists of text and multimedia content e.g. e-mail messages, videos, photos, audio files, presentations, webpages and various other kinds of business documents. It's worth noticing that even though these types of files may have an internal structure, yet they are still considered "unstructured" because the data contained by them doesn't fit neatly in a database. Experts have put an estimate that 80 to 90 per cent of the data in an organization is unstructured type. And the amount of unstructured data in organizations is growing significantly and drastically (usually many times faster than structured databases) [51].

Data which is, accurate and timely, specific and organized for a purpose, presented within a context that gives it meaning and relevance, and that can lead to an increase in understanding and decrease in uncertainty. Information is valuable asset as it can affect behaviour, a decision, or an outcome. A piece of information is valueless if, after obtaining it things don't change. In Today's world what we are considered is with the information and patterns that can help us in making decisions. What we can do with this data is that we can refine it to generate both the information and patterns which can be fruitful to us. To achieve this many methods and concepts were adopted and mostly widely accepted and Practised is Data- Ware housing.

III. DATA WAREHOUSING AND MINING:TO OBTAIN INFORMATION AND PATTERNS

It was in 1993, the father of data warehousing, Bill Inmon penned down the definition of data warehouse as: "A data warehouse is a subject oriented, integrated, time-variant, non-volatile collection of data in support of management decisions [2]." Recent advances in computer and networking technology have led to the development of hardware and software platforms that can help to collect, manage and distribute large amount of pertinent data. Data Warehousing is most interesting and dynamic among the new technological transitions available. It works as a repository of subjectively selected and adapted operational data, which can be successfullyused to answer any ad hoc, complex, statistical or analytical queries. Also, it provides a mechanism for implementing effective decision support system by utilizing data which scattered all over the organization. Data warehousing provides us with the information that is useful to us from summarized data. Now there was a need of something that could provide us with patterns or trends (Knowledge extraction) in the data that was not immediately apparent by just summarizing the data. This led to us a technology named Data Mining. Data mining is said to be originated from three branches of artificial intelligence i.e. a) neural networks, b) machine-learning and c) genetic algorithms that have brought us to great analytical advancement. It is method which we use to predict the future (predictive analytics) by providing the answer to "How" or "why" [3][4][5][6].

IV. BIGDATA HANDLING: SOLUTION AND APPROACHES

Big Data is understood as a data analysis methodology which is enabled by a new generation of technologies and architecture which supports high-velocity data capture, storage, and analysis. Data sources have now extended beyond the traditional corporate database so that they include e-mail, mobile device output, sensor-generated data, and social media output [7]. Data now is no longer restricted to structured database records but also unstructured data is included [8].

Big Data needs huge amounts of storage space. As the price of storage continues to decline, the resources needed to deal with big data can still create financial difficulties for small to medium sized businesses. Basic and typical big data storage and analysis infrastructure can be developed on clustered network-attached storage (NAS). Clustered NAS infrastructure needs configuration of many NAS "pods" with each NAS "pod" comprising of several storage devices connected to an actual NAS device. The series of NAS devices can then be interconnected to allow massive searching and sharing of data [9]. "Big Data is referred to as datasets whose size is beyond the ability of typical database software tools and methods to capture, store, manage and analyse. "As of now there is no explicit definition of how big a dataset must be in order to be considered as Big Data.

New technologies are to be placed to manage this Big Data phenomenon [10].Big Data technologies as a combination of new generation technologies and architectures which are designed so as to extract value economically from very huge volumes heterogeneous data" .This is done by enabling high velocity capture, discovery and analysis [11]."Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it [12]."

Analysis of big data is done using a programming paradigm called MapReduce [13]. In this Map Reduce paradigm, a query is made and data are mapped to find key values which are considered to relate to the query; the results after that reduced to a dataset which provides answer to the query. Most enterprise data management tools present today (database management systems) are designed to execute basic queries run quickly. Data is indexed so that only very small portions of the data need to be examined in order to answer a query. This solution is not applicable to data that cannot be indexed, i.e., in semi-structured form (text files) or unstructured form (media files). To reply and execute a query successfully in this case, all the data has to be examined [14].

To explain Map Reduce Further, Map Reduce is a data processing algorithm which uses a parallel programming implementation. In basic terms, Map Reduce is a programming paradigm which involves distributing a work across multiple nodes executing a "map" function. The map function accepts the problems, splits it into sub-parts and then sends them to different machines so that the entire sub-parts can executed concurrently. The results from these parallel map functions are collected and distributed to a set of servers that are running "reduce" functions, which then collects the results from the sub-parts and then re-combines them to get the single answer.

Many of the technologies within the Big Data environment are of an open source origin, due to participation, innovation,development and sharing between the commercial providers that are in open source development projects. The Hadoop framework based on Map Reduce, in conjunction with additional software components like R language and a wide range of open source Not Only Structured Query Language (NoSQL) tools which include Cassandra and Apache HBase, have become the basics of many Big Data discussions today. Vendors have now launched their own versions of such tools (e.g., Oracle's version of the NoSQL database) or tointegrate these tools with their own products (e.g., EMC's Greenplum Community edition which includes the open source Apache Hive, HBase and ZooKeeper) [15].

The HDFS (Hadoop data file system) is a considered to be fault-tolerant storage system that can store large volumes of information, scale up incrementally and survive on storage failure without losing data. Hadoop clusters are generally built with inexpensive computers. If one node (computer) fails, the cluster can still continue to operate without losing data or for that matter interrupting work by simply re-distributing the task to the remaining machines in the cluster. HDFS manages the storage on cluster actuallyby breaking files into small blocks and then storing duplicated copies of them over the pool of nodes.Comparing it with other redundancy techniques, including the strategies that are employed by Redundant Array of Independent Disks (RAID) machines found, HDFS has extra two key advantages. Firstly, HDFS does not require any special hardware as it can be built from hardware that is common. Secondly, it enables an efficient method of data processing in the form of MapReduce [16].

In addition to Map Reduce and HDFS, Hadoop also refers to a collection of other software projects that uses the MapReduce and HDFS framework. Some of the tools include: HBase, Hive, Pig, Mahout, Zookeeper, and Sqoop. The global market size of Hadoop projects in 2011 was US$77 million. The market is expected to grow almost ninefold to US$682.5 million by 2015 [1].

NoSQL database management systems (DBMSs) are available asopen source software and are designed for use in high data volume applications in the clustered environments. They usually do not have any fixed schema and are non-relational, unlike the traditional SQL database management system (also known as RDMS) present in many data warehouses today. Because they don't adhere to a fixed schema, NoSQL DBMS permit us with more flexible usage, allowing high-speed access to both semi-structured and unstructured data. However, SQL interfaces are also being used alongside the MapReduce programming paradigm. There are several types of NoSQL DBMS like:

Key-Value Stores: Key-value pair (KVP) tables are used to provide determination management for many of the other NoSQL technologies. The concept can be explained as: the table has two columns - one is for the key; the other is for the value. This value could be a single value or a data block containing more than one value, the format of value(s) is determined by program code. KVP tables may use one among these: indexing, hash tables or sparse arrays to provide rapid retrieval and insertion capability, depending on one's need for fast look-up, fast insertion or efficient storage. KVP tables are best to be applied for applied simple data structures and on the Hadoop MapReduce environment. Some Examples of key-value data stores include Amazon's Dynamo and Oracle's BerkeleyDB [17].

Document Oriented Database: A document-oriented database is a databasethat's designed for storing, retrieving and managing document-oriented or semi-structured data. The central concept of a document-oriented database is the idea of a "document" where the contents inside the document are encapsulated in some standard format such as JavaScript Object Notation (JSON), Binary JavaScript Object Notation (BSON) or XML. Examples of these databases include Apache's Couch DB and 10gen's Mongo DB.

BigTable Database: It is a distributed storage system based on the registered Google File System for managing structured data that is aimed to scale to a very large size – Petabyte's of data across thousands of servers. It is also known as Distributed Peer Data Store. This database is nearly similar to relational database except that the volume of data handled is very high and the schema does not order the same set of columns for all rows. Each cell has a time stamp and there can be multiple forms of a cell with different time stamps. In order to manage these huge tables, Bigtable breaches tables at row boundaries and saves them as tablets. Each tablet is nearly of 200MB in size, and each server can save about 100 tablets. This arrangement allows tablets from a single table to be spread among various machines, load balancing and fault tolerance is also managed by it. An example of a BigTable database is CassandraDB [18].

Graph Database: A Database that contains contains nodes, edges and properties to represent and store data is said to be Graph Database. In this database, every entity contains a direct pointer to its neighbouring element and no index look-ups are required. A graph database is valuable when large-scale multi-level relationship traversals are public and is best suited for processing complex many-to-many connections like social networks. A graph can be taken by a table store that supports recursive joins such as BigTable and Cassandra. Examples of graph databases include InfiniteGraph from Objectivity and the Neo4j an open source graph database [19].

All the above mentioned Information makes as realise even though that big data can be processed but still for small and medium sized industries to opt for such technologies is very expensive and out of dream for them." Besides this Big data possess various other problems in the field of Data Processing and resource Management, Data Integration, Data Storage, Data visualization and user Interaction, Model building and scoring. The organized approach toward data collection in order to enhance randomness in data sampling and reduce favouritism is not apparent in the collection of Big Data sets [20]". Big Data sets do not naturally reject data bias. The data collected can still be imperfect and inaccurate which, in turn, can lead to twisted conclusions. Twitter which is commonly inspected for insights about user feelings , that there is a natural problem with using Twitter because as a data source as only 40% of Twitter's active users are just listening and not contributing. This can suggest that the tweets are coming from a certain type of people (perhaps people who are more vocal and participative in social media) than from a true haphazard sample [21].Twitter makes a sample of its ingredients available to the public by its streaming Application Programming Interfaces (APIs).It is not however clear how actually sample of materials is derived [22]. In broader terms there are three areas of problems Associated with big Data which include Big Data Computation And Management, Big Data computation and Analysis and, BigData security. The Solution that is reliable and cheap in solving the case of BigData to benefit every class is Cloud Computing [43].

V. CLOUD COMPUTING AT RESCUE

Cloud Computing is a word used to describe a that new class of network based computing which takes place over the Internet or a model that depends on a large, centralized data center to store and process a wealth of information. Cloud computing has fundamentally been derived by the need to process a huge quantity of data [23].Data today is no longer measured in gigabytes but in Exabyte's as we are "Approaching the Zetta Byte Era". Cloud computing is all intended to access huge amounts of computing power by combining resources and offering a single system view [24]. Cloud computing has become a powerful architecture to perform extensive and complex computing, and has transformed the way that computing setup is abstracted and used. In addition to this, an important goal of these technologies is to deliver computing as a solution for tackling big data, such as large scale, multi-media and high dimensional data sets. Cloud computing is associated with new model for the provision of computing infrastructure and method for processing big data with all kinds of resources. Moreover, some innovative cloud-based technologies have to be implemented because dealing with big data for parallel processing is difficult. Cloud deployment solutions provide services that businesses would otherwise not be able to afford under the traditional hardware and software acquisition method. Cloud computing transforms the way information is handled, the typical organization models for cloud computing includes: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS) and hardware as a service (HaaS).

a) Infrastructure as a Service (IaaS):

In Infrastructure as a Service (IaaS) model, consumers are given full liberty to manage their data on the server. Here the service provider has to be responsible for raw storage, computing power, networks, firewalls, and load balancers and this is often demonstrated as a virtual machine [25]. A client business pays on a per-use basis whenever the equipment is used to support computing operations such as: storage, hardware, servers, and networking equipment. Infrastructure as a service is a cloud computing model that has received most attention from the market, with an expectation of 25% of enterprises planning to adopt a service provider for IaaS [26]. Some of the Services available to businesses through the IaaS model include disaster recovery, computing as a service, storage as a service, data center as a service, virtual desktop infrastructure, and cloud bursting, which is providing peak load capacity for variable processes. Profits of IaaS include increased financial elasticity, choice of services, business agility, cost-effective scalability, and increased security.

b) Platform as a Service (PaaS):

Platform as a Service is a level exceeding Infrastructure as a service (IaaS). In the PaaS model, customers are provided with an operating system, programming language execution environment, database, and web server. They are not concern with the either cost or management in the hardware and software layers. PaaS is the use of cloud computing to provide platforms for the development and use of routine applications [27]. Generally, PaaS solutions include application design and development tools, application testing, versioning, integration, deployment and hosting, state management, and other related development tools [28]. Businesses attain cost savings using PaaS through standardization and high utilization of the cloud-based platform across a number of applications [29].Some other advantages of using PaaS as: lowering risks by using pretested technologies, promoting shared services, improving software security, and lowering skill requirements needed for new systems development [30]. As related to big data, PaaS provides companies a platform for developing and using custom applications needed to analyse huge quantity of unstructured data at a low cost and low risk in an environment that is secure. Therefore maintaining the integrity of applications and with accurate authentication checks during the transfer of data across the entire networking stations is fundamental.

c) Software as a Service:

Software as a service (SaaS) is the level beyond Platform as a service (PaaS).Here; consumers are given access only to the application software, which can be run remotely from the data centres of the cloud service provider [31]. Here the provider is responsible for the maintenance and support of the infrastructure and operating platforms i.e. it provides businesses with applications that are stored and run on virtual servers in the cloud, as cloud service providers specialize in one area, they can provide consistent service at a fraction of the cost. The business is not charged for hardware, only for the bandwidth for the time and number of user's necessary. Advantage of SaaS is that this solution allows businesses to shift the risks associated with software acquisition while moving IT from being reactive to proactive [32]. Benefits of using SaaS as: easier software administration, automatic updates and patches management, software compatibility across the business, easier collaboration, and global accessibility [33].

d) Hardware as a Service (HaaS):-

HaaS is not yet widely used but it is a cloud service based upon time sharing model on minicomputers and mainframes from the 1960s and 1970s, contrarily to the SaaS and PaaS that provide applications and services to the customers, HaaS offers only the hardware [34]. Time sharing developed into the practice of managed services. In a managed service situation, the managed service provider (MSP) remotely monitors and administers hardware located at a client's site as contracted. A problem with managed services was the necessity for some MSPs to provide hardware on-site for clients, the cost of which needed to be built into the MSP's cost [35]. The HaaS model allows the customer to license the hardware directly from the service provider which alleviates the associated costs. Vendors in the HaaS arena include Google with its Chromebooks for Business, CharTec, and Equus. However, as cloud IT solutions are becoming more common and accepted the possible markets for cloud are also expanding rapidly [36].

BigData cloud alone is not a corporate solution but an IT tool. Traditionally companies have learned how to outsource certain elements to modernise their processes. Cloud computing is the next step that allows outsourcing of IT, instead of forming and maintaining their own IT department with physical servers and technical consultants companies hires a cloud service company to provide all its IT needs. In addition to this,cloud computing allows computing to be treated as a service. Earlier a company which was in need of computing power and storage was not only required to purchase its own processors and servers but also, maintain them. Lazy capacity and economic waste is recorded whenever those IT resources are not in use.

However, with the arrival of cloud computing, a company can acquire its exact computing needs. All these make it clear that if companies and governments need to use Big Data then Cloud computing is the best option they can opt for.It is clearly visible that cloud computing when used for BigData computation has certain benefits like Flexibility,Storage, Saving Time and reducing costs [41].

VI. CLOUD COMPUTING WITH RESPECT TO BIG DATA COMPUTATION: LATEST APPROACHES AND RESEARCH

It may look easy and comfortable to implement Cloud Computing for Big Data Processing but it a very tedious and complex task with various other Challenges associated with it. Various methods and approaches that are being used for carrying out analytics on cloud for Big Data applications. Some of the areas of work include a) Data Management b) Model Development and Scoring c) Visualisation and User interactions d) Business model, are under focus [49]. Present DBMS technologies has various issues like Scalability, reliability etc. They can't do justice to BigData as data here is heterogeneous i.e. both Structured and Un-Structured therefore what can be done is the use of NoSQL. Furthermore key Value stores for supporting riches set of applications is to be researched upon. This is due to the fact that cloud computing needs a DBMS that supports descriptive and deep analytics. Also it needs a support for updating of heavy Applications, Ad-hoc Analytics and decision support [37]. One of the main problem rather a challenge is how security can be managed while uploading and accessing and processing of data on cloud.Some of the benefits of using NoSQL which include:

a) Availability and Partition tolerance;
b) Consistency and availability; and;
c) Consistency and Partition tolerance.

With all these benefits of NoSQL there are certain things that need to be addressed like transitional Integrity, Authentication Mechanisms, and Insider Attack [42].Various categories of security challenges with possible Solutions. These challenges and solutions are as below:

a) Network level with File encryption and Network encryption as the solution.
b) Authentication level with logging as the solution
c) Data level with software format and Node maintenance, Node authentication as the solutions
d) Generic levelwith honey pot nodes,rigorous testing of MapReduce Jobs, layered framework of assuring cloud, Third party secure data publication to cloud, Access control as the solutions [38].There are several cryptographic techniques that can be used which include a) Homomorphic encryption b) Verifiable Computation c) Multiparty Computation .There is a huge scope for Multiparty Computation [40].

Since data here is heterogeneous in nature so Hadoop which works well for short jobs on Homogeneous data doesn't work well here. Therefore a solution in terms of LATE was developed them[39]. In order to increase the speed of query execution on Hadoop a project was carried out in Yale university by the name of Hadapt. In Hadapt data warehouse queries were carried out in split execution which gives the efficient results [44].To achieve the higher rates for data retrieval Webservers can embedded into cloud environment. This can be done by creating a multilevel index in web server with multilevel index key in data node [45]. One Major Problem that is associated with Cloud Computing on BigData computation is Network Intrusion. Solution to this can be integration of modern technologies, Hadoop File System and cloud Technologies with latest representation learning techniques and support to predict network intrusions through BigData classification strategies[46]. Something needs to be done on behalf of image processing in BigData i.e. on larger data sets this is due to the fact that Hadoop has one major drawback that it can't perform well on unstructured and heterogeneous data. To overcome a solution in terms of Hadoop Image processing Interface (HIPI) is provided. This system is developed to work on large data set of images' format is provided for storing images so that efficient access within the MapReduce Pipeline can be done. Here we have a culling stage is introduced before the Mapping stage, acting as a simpler way to filter the image sets and control the types of images in MapReduce tasks. Finally, image encoders and decoders are present that run behind the scenes and work to present the users with the float image types which are most useful for image processing and vision applications [48].Wealth of Information needs to be extracted out of BigData which if otherwise left untreated would make it a data monster ,but the problem here is that present mining Algorithms won't work well here, to overcome this we need to change our traditional mining algorithms and techniques [47]. A single Perfect Data Management solution is yet to be designed. Also effective techniques for dealing with the elasticity of cloud infrastructures, designing scalable, elastic and autonomic multitenant database systems with security and privacy needs to be designed [50].

VII. CONCLUSION – RESEARCH GAP

Now after the review of research papers, it is found that "Cloud Computing" is a ray of hope and one of the methods that can help researchers and industry people to achieve great heights when applied on BigData to unveil the wealth of knowledge. This Information can prove boon to almost every aspect of life i.e. from

agriculture to Space Missions. Since much of work have been done on this burning Topic, still we need something that is concrete and effective which can not only help us to maintain large amount of data but also provide the security and better access ,so that , mining becomes an easy task to do without changing much in its existing technologies and algorithms. A perfect solution to handle heterogeneous data with security and better access on clouds is yet to be formed. This means that a single perfect Data Management Solution needs to be formulated.

REFRENCES

[1] International Data Corporation, (2011), "*The 2011 Digital Universe Study: Extracting Value from Chaos*", Accessed at: http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

[2] W.H. Inmon, "Building the Data Warehouse", John Wiley. pp. 33, 1996

[3] M.S. Chen, J. Han, and P.S. Yu, " *Data Mining: An Overview from a Database Perspective*", IEEE Transactions on Knowledge and Data Engineering, 8(6), pp. 866-883, 1996

[4] R. Agrawal , R. Srikant, "*Mining Sequential Patterns*", In: Yu P, Chen A (Eds).Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan, pp.3–14,1995

[5] R.P Schumaker, et al., "*Sports Data Mining Methodology, Sports Data Mining, Integrated*", In: Information Systems 26, Springer Science+ Business Media, LLC, 2010

[6] S. Shabia, M.A.Peer, "*Expedition for the exploration of Apposite Knowledge*", International Journal of Computer Science and Information Technologies, 3 (5),pp.5164 – 5168,2012

[7] R. L. Villars, C. W. Olofson, and M. Eastwood, "*Big Data: What It Is and Why You Should Care*," White Paper, IDC, June 2011

[8] Coronel, S. Morris, & Rob, "*Database Systems: Design, Implementation, And Management*", (10th Ed.). Boston: Cengage Learning P.,2013

[9] White, "*Data Communications And Computer Networks: A Business User's Approach*", (6th Ed.). Boston: Cengage Learning, 2011

[10] J. Manyika, "*Big data: The next frontier for innovation, competition, and productivity*".Accessed:http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[11] IDC,(2012),"*Worldwide Big Data Technology and Services 2012-2015 Forecast*", Accessed: http://www.idc.com/getdoc.jsp?containerId=233485

[12] EddDumbill,(2012)"*WhatIsBigdata?*",Accessed:http://radar.oreilly.com/2012/01/what-is-big-data.html

[13] P.C Zikopoulos, C. Eaton, D. DeRoos, T. Deutsch,G. Lapis, "*Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data*", McGraw-Hill: Aspen Institute, 2012

[14] J. Dean and S. Ghemawat, "*MapReduce: Simplified Data Processing on Large Clusters*," in Proceedings of OSDI '04: 6th Symposium on Operating System Design and Implementation, December 2004 San Francisco, CA,USA

[15] H. Doug, (2012), "*Oracle Releases NoSQL Database, Advances Big Data Plans*", Accessed at:http://www.informationweek.com/software/informationmanagement/oracle-releases-nosql-database-advances/231901480

[16] W.O. Carl, V. Dan. (2012),"*Worldwide Hadoop – MapReduce Ecosystem Software 2012-2016*",Accessed at :http://www.idc.com/getdoc.jsp?containerId=234294

[17] W.O. Carl, "*The Big Deal about Big Data*". Accessed: http://www.idc.com/getdoc.jsp?containerId=226904\

[18] Google Inc., "*Bigtable: A Distributed Storage System for Structured Data*". Accessed:http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/bigtable-osdi06.pdf

[19] InfoGrid."*OperationsonalGraphDatabase*",Accessed:http://infogrid.org/blog/2010/03/operations-on-a-graph-database-part-4/

[20] "*BigData*".Accessed:http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

[21] M. Rosoff. "*Twitter Has 100 Million And 40% Are JustWatching*", Accessed: http://articles.businessinsider.com/2011-09-08/tech/30127585_1_ceo-dick-costolo-twitter-users

[22] Twiiter,Accessed:https://dev.twitter.com/docs/streaming-apis/streams/public

[23] Armbrust, Griffith, Joseph, Katz, Konwinski, Lee, Zaharia", *A View Of Cloud Computing*", Communications Of The ACM, 53(4),(2010)

[24] Cisco,(2009),http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11481374_ns827_Networking_Solutions_White_Paper.html

[25] Rouse, (2010), "*Infrastructure as a Service*", Accessed : http://searchcloudcomputing.techtarget.com/ definition /Infrastructure-as-a -Service-IaaS

[26] Cisco,(2009),"*Infrastructure as a Service: Accelerating time to profitable new revenue streams*".Accessed:http://www.cisco.com/en/US/solutions/collateral/ns341/ns991/ns995 /IaaS_BDM_WP.pdf

[27] Salesforce.com,(2012),"*The end of software: Building and running applications in the cloud*", Accessed: http://www.salesforce.com/paas/

[28] Géczy, Izumi, Hasida,"*Cloudsourcing: Managing Cloud Adoption*", Global Journal of Business Research, Vol.6, No.2, 2012

[29] Oracle,(2012),"*OraclePlatformAsAService*".Accessed:http://www.oracle.com/us/technologies/cloud/oracle-platform-as-aservice408171.html

[30] Jackson,(2012),"*Platformasaservice:Thegamechanger",*Accessed:http://www.forbes.com/sites/kevinjackson/2012/01/25/platform-as-a-service-the-gamechanger/

[31] Cole, (2012), "*Looking At Business Size, Budget When Choosing Between Saas And Hosted ERP. E-Guide: Evaluating Saas Vs. On Premise For ERP Systems"*. Accessed: http://docs.media.bitpipe.com/io_10x/io_104515/item_548729/SAP_sManERP_IO%23104515_EGuide_061212.pdf

[32] Carraro,Chong,(2006),"*Softwareasaservice:Anenterpriseperspective*",Accessed:http://msdn.microsoft.com/enus/library/aa905332.aspx #enterprisertw_topic3

[33] Rouse,(2010),"*Software As A Service*", Accessed: http://searchcloudcomputing.techtarget.com /definition/Software-as-a-Service

[34] ComputureWeekly.com,(2009),"*HardwareAsAService*",Accessed:http://www.computerweekly.com/feature/Hardware-as-a-Service

[35] Rouse,(2007)."*HardwareAsAService*".Accessed:http://searchitchannel.techtarget.com/definition/Hardware-as-a-Service-in-managed-services

[36] Panettieri, (2011),"*Can Google Take Hardware As A Service (Haas)*", Accessed:http://www.mspmentor.net/2011/06/13/can-google-takehardware- as-a-service-haas-mainstream/

[37] Agrawal, S. Das, and A. E. Abbadi." *Big Data And Cloud Computing: Current State And Future Opportunities*", Proceedings of the 14th International Conference on Extending Database Technology, EDBT/ICDT '11, p.530–533, 2011. ACM. II.2(b), New York, NY, USA

[38] V.N. Inukollu, S. Arsi ,and S. R. Ravuri, "*Security Issues Associated With Big Data In Cloud Computing*", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[39] M. Zaharia , A. Konwinski , A. D. Joseph , R. Katz , I. Stoica, "*Improving MapReduce Performance In Heterogeneous Environments*", Proceedings of the 8th USENIX conference on Operating systems design and implementation, p.29-42, December 08-10, 2008, San Diego, California ,USA

[40] S. Yakoubov, V.Gadepally, N. Schear, E. Shen, and A. Yerukhimovich, "*A Survey Of Cryptographic Approaches To Securing Big-Data Analytics In The Cloud*," Proceedings of High Performance Extreme Computing Conference (HPEC),p.1-7, September 9-11, 2014. Waltham, Massachusetts, USA

[41] O.A. Iyanda."*Big Data and Current Cloud Computing Issues and Challenges*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol.4, No.6, June 2014

[42] K. Zaki , "*NoSQL DATABASES: New Millennium Database For Big Data, Big Users, Cloud Computing And Its Security Challenges*", International Journal of Research in Engineering and Technology (IJRET),Vol.3,No.3,May 2014

[43] Ji, Y. Li, W. Qiu, U. Awada, K. Li , "*Big Data Processing in Cloud Computing Environments*", International Symposium on Pervasive Systems, Algorithms and Networks,p.17,2012

[44] K.Bajda-Pawlikowski , D. J. Abadi , A. Silberschatz , E.Paulson, "*Efficient Processing Of Data Warehousing Queries In A Split Execution Environment*", Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, June 12-16, 2011, Athens, Greece

[45] Shah, Annappa and K. C.Shet."*Design An Efficient Big Data Analytic Architecture For Retrieval Of Data Based On Web Server In Cloud Environment*", International Journal on Cloud Computing: Services and Architecture (IJCCSA), Vol.4,No.2,April 2014

[46] S. Suthaharan, "*Big Data Classification: Problems And Challenges In Network Intrusion Prediction With Machine Learning*," in Big Data Analytic workshop, in conjunction with ACM Sigmetrics, p.70-73 , Vol.41,No.4,2013

[47] B.R. Prakash ,Dr. M. Hanumanthappa,"*Issues and Challenges in the Era of Big Data Mining*", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Vol.3,S.No.4,July 2014

[48] S. Chris, L. Liu, A. Sean, and L. Jason, "*HIPI: A Hadoop Image Processing Interface For Image-Based Map Reduce Tasks*," B.S. Thesis. University of Virginia, Department of Computer Science, p. 2–3 ,2011

[49] M. D. Assunção , R. N. Calheiros , S. Bianchi , M. A.S. Netto , R. Buyyab, "*Big Data Computing And Clouds: Trends And Future Directions*," Journal of Parallel and Distributed Computing,p.156-175,Vol.75,No.13 August 2014

[50] Agrawal , S. Das , A. E. Abbadi, "*Big data and cloud computing: new wine or just new bottles?*", Proceedings of the VLDB Endowment, v.3 n.1-2 pp.1647 -1648, September 2010.

[51] http://www.bu.edu/datamanagement/background/importance/