Efficient Storage Management over Cloud using Data Compression without Losing Searching Capacity

Amitkumar P Gohil
Parul Institute of Technology,Vadodara
Gujarat, India-391760

Amish Desai
Assi.Prof. Parul Institute of Technology,Vadodara
Gujarat, India-391760

ABSTRACT
Nowadays due to social media, people may communicate with each other, share their thoughts and moments of life in form of texts, images or videos. We are uploading our private data in terms of photos, videos, and documents on internet websites like Facebook, Whatsapp, Google+ and Youtube etc. In short today world is surrounded with large volume of data in different form. This put a requirement for effective management of these billions of terabytes of electronic data generally called BIG DATA. Handling large data sets is a major challenge for data centers. The only solution for this problem is to add as many hard disk as required. But if the data is kept in unformatted the requirement of hard disk will be very high. Cloud technology in today is becoming popular but efficient storage management for large volume of data on cloud still there is a big question. Many frameworks are available to address this problem. Hadoop is one of them. Hadoop provides an efficient way to store and retrieve large volume of data. But Hadoop is efficient only if the file containing data is large enough. Basically Hadoop uses a big hard disk block to store data. And this makes it inefficient in the area where volume to data is large but individual file is small. To satisfy both challenges to store large volume of data in less space. And to store small unit of file without wasting the space. We require to store data not is usual form but in compressed form so that we can keep the block size small. But if we do so it added one more dimension of problem. Searching the content in a compressed file is very in-efficient. Therefore we require an efficient algorithm which compress the file without disturbing the search capacity of the data center. Here we will provide the way how we can solve these challenges.

Keywords
Cloud, Big DATA, Hadoop, Data Compression, MapReduce

1. INTRODUCTION

1.1 What is Cloud Computing?
Today, most popular work of people is to surf the internet download song, movies or watch online videos or playing internet online games. And for these websites like Google, Yahoo! and Facebook mostly used and get clicked millions in some minutes. With this they are earning lot of money. But due to heavy use of internet, generates invaluable data which stores lots of terabytes of disk spaces. So nowadays we are getting lack of disk space for storing the internet data. For solving this problem, we are now able to store the data on cloud storage instead of local hard disk. This will make data availability is very high. We can get our data everywhere any time without need of any extra storage device. For this, a number of cloud computing technologies have been introduced in last few years. Cloud computing is a technique which provide services over the internet. The cloud, in terms of technical terms, refers to the datacenter which can be virtual hardware and software that will provide supports a client's needs, often in the form of data storage and online hosted applications. These type of infrastructures enable organizations to reduce costs by eliminating the use of physical hardware, allowing organizations to outsource data and computations on demand. Cloud infrastructure developers with innovative ideas for Internet services now no need for large capital investment in hardware to deploy their services. Everyone is depending upon the datacenters for their data but datacenters are also facing the problems of storing the data because day by day number of increasing internet users very much they are using their data online. Social media websites are being the most important role for this. You can see in Fig.1 graph which is showing the usage of internet in the year of 2007. And this may be crucial in next 4-5 years.

Cloud Computing is a new technology which describe a new class of network based computing that is taking place over the Internet. Using the Internet for communication and transport this technology provides hardware, software and networking services to clients. It is a one kind of platform provided to the user over the internet. These platforms hide the complexity and details of the infrastructure from users and applications by providing very simple and easy graphical interface or API (Applications Programming Interface).

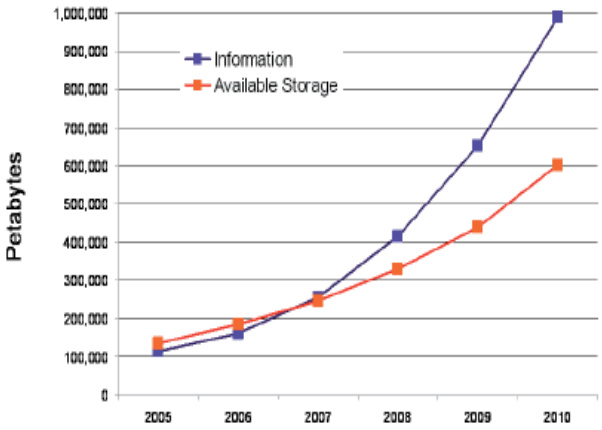

Fig.1 Information VS Available Storage [11]

1.1.1 Cloud Deployment Models
The most common types of cloud computing deployment models as below:

Private cloud – This type of cloud infrastructure is being operated in a solely organization or provided by the third party within a single organization.

Community cloud – here cloud infrastructure is shared by several organizations and supports for a specific community that has common interests (e.g., forum, blogs, mission, games, industry collaboration, or compliance requirements). This may be managed by the creators of this community or by the third party.

Public cloud – here cloud infrastructure is available to all general public or a large industry group and is owned by an organization who is selling cloud services.

Hybrid cloud – this cloud infrastructure is a combination of all above part means two or more cloud infrastructures, may be they are private, public or community. Means if any organization wants to share their some private cloud data to the public user, it share the data to public cloud. So this combination of cloud infrastructure is hybrid.

1.1.2 Cloud Service Delivery Models

Cloud service delivery models are those in which form cloud service providers are providing cloud services and the most common are:

Software as a Service (SaaS) – Applications use to perform specific functions or processes (e.g., email, customer management systems, enterprise resource planning systems, and spreadsheets). In a simple way, applications which cannot be installed in some computer due to low system configuration or some another problems, we can get service of that application over the internet via cloud. E.g. Email Servers

Platform as a Service (PaaS) – In older system or computer Java was not supported. So it was difficult to run java Script over net unless user install it in his system manually. So PaaS is providing platform to the user so user can run java apps into their old systems over the internet via cloud. E.g. Java Virtual Machine.

Infrastructure as a Service (IaaS) Sometimes we need to test our project or like other task in which we need of networks of the systems. Or if we need an extra storage capacity but we have not enough money for buying new hard disk. So here IaaS is providing infrastructure to the user over internet via cloud. E.g. Google Drive

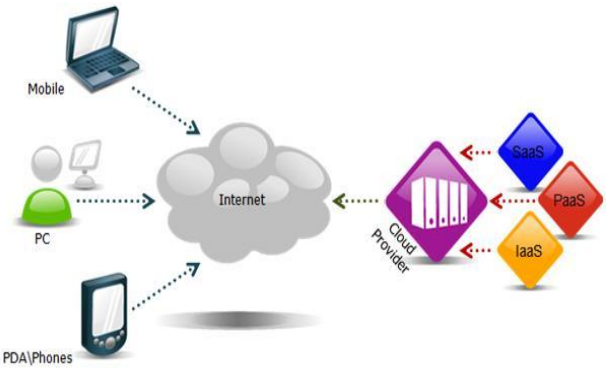

Fig.2 Cloud Computing [6]

1.2 What is BIG DATA?

Big data is a data intensive technology which is used for a collection of large and complex data sets that are difficult to process by using traditional database management tools and applications. It represents a new technology drift for applications in different fields of study such as science and technology, design, production, data storage, research, networking systems etc. Big-Data technologies comprise of a new set of technologies and architectures designed to derive large set of data which is analyzed and determine quality of research in science and technology, determine real-time on road traffic conditions, prevent diseases, combat crime etc.

Big DATA is the technology which supports the several aspect of business economic model. "Big Data technologies" in a collective term that includes the platform, tools and techniques which can be used for capture, process, investigate and visualize large datasets in a very easy reasonable timeframe not accessible for traditional IT technologies. Big Data differs itself from standard traditional technologies in three aspects: the volume of data, velocity i.e. the rate of generation and transmission of data, and type of data under consideration. In order to achieve benefits of big data technologies it is required to adopt analysis methods and follow with the new storage technologies. This forces business persons to determine the new methodologies to serve their information needs because ignoring this technology or methods of big data will eventually make the business organization non-competitive. As big data infrastructure composes of huge volumes of data. Organizations using big-data generally come across significant risks and threats to these storehouses. Organizations are generating more and more data now and due to this understanding its importance, context, usability and protection mechanisms are very crucial factors for us.

1.2.1 Opportunities and challenges for BIG DATA

Big Data environment has started to impact almost all types of organizations, since it has the potential power to extract useful knowledge from huge volumes of data and operate upon it as per the requirements on real time basis. The opportunities provided by Big Data systems can be explained by analyzing its applicability in different areas as below:

1. **Healthcare:** The healthcare industry is quickly moving to electronic medical records and images that it may be useful for public health monitoring and in epidemiological research programs. In healthcare industry, Big Data is also associated with the massive volume of patient-specific data. A valid example is in medical imaging where small pathological features measuring only a few millimeters can be detected in magnetic resonance imaging and in CT scans.

2. **Mobile Networks:** The amount of mobile data traffic is expected to grow to 10.8 Exabyte per month by 2016 due to increased usage of smart phones and tablets [11]. Big Data is needed for managing and operating mobile networks and it also aim of improving network quality, considering issues such as isolation and correlation of network faults, security breach detection, traffic planning, hardware maintenance predictions etc.

3. **Video surveillance:** Video surveillance is in a transition phase from CCTV to IPTV cameras and recording systems that organizations want to analyze for behavioral patterns. Big data can be used to analyze huge volumes of data so as to generate security and service enhancement.

4. **Media and Entertainment:** The media and entertainment industry has shifted to digital recording, production, and delivery in recent times and Big Data approach could be used for collecting the huge volumes of rich content to find and analyze user viewing behaviors.

5. **Life sciences:** In field Life Sciences the low-cost gene sequencing can produce tens of terabytes of information required to be analyzed to find genetic variations, DNA sequencing, treatment success rate etc.

6. **Transportation:** Sensor data is being generated at an accelerating rate from fleet GPS transceivers, RFID tag readers, smart meters, and cell phones and this data is being used to optimize business operations to realize incoming business opportunities.

7. **Environment Study:** Efficient environment study requires collecting and analyzing data from thousands of sensors that monitor air and water quality and meteorological conditions. Careful analysis of collected data can then be used to direct simulations of climate and groundwater models for predicting long term trends and changes in environment such as increased CO2 emissions, ground water table level etc.

1.3 What is Mapreduce?

MapReduce is a programming model and a connected implementation for processing and generating huge data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is combination of a Map() function that performs filtering and sorting (such as sorting batsmen by first name into queues, one queue for each name) and a Reduce() function that performs a summary operation (such as counting the number of batsmen in each queue, yielding name frequencies). The "MapReduce System" (also called "MapReduce infrastructure" or "MapReduce framework") composes the processing by assembling the distributed servers in order, running different tasks in parallel, managing all communications and data transfers between the different parts of the system, and also redundancy and fault tolerance is provided. This model is well known for map and reduce functions commonly used in functional programming, even though their purpose in the MapReduce framework is not the same as in their original methods. Single threaded implementation of MapReduce will usually not be faster than a traditional database implementation, it usually works with multi-threaded implementations. Here shuffle operation reduces cost of network communication, and fault tolerance features of MapReduce framework is the use of this model advantageous. MapReduce libraries have been written in various programming languages, with different levels of optimization. A popular open-source implementation is Apache Hadoop. The name MapReduce originally referred to the proprietorship of Google Inc., but it has been generalized in Apache-Hadoop.

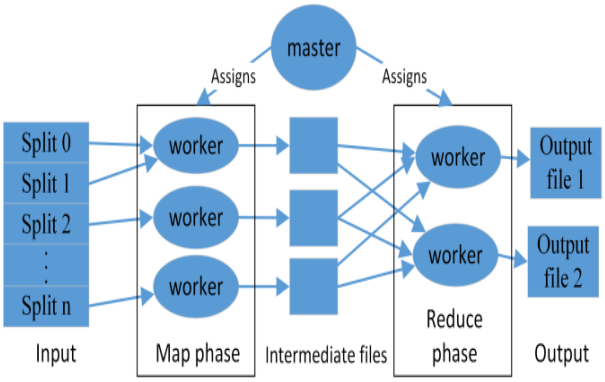

Fig.3 MapReduce Flow [2]

MapReduce is a programming model for processing large data sets in distributed environments [7]. Map Reduce is actually an analytical tool which analyse the data, compare the data which helps in terms of reducing the duplication of data thus we can save the disk space. In the MapReduce model, the Map function performs filtering and sorting, while the Reduce function carries out grouping and aggregation operations. The 'An apple in a day keeps doctor away' of MapReduce is the word counting example: it counts the occurrence of each word in a set of documents. The Map function splits the document into words and for each word in a document it produces a (key, value) pair.

Function [2] map(name, document)

for each word **in** document

emit (word, 1)

The *Reduce* function is responsible for aggregating information received from *Map* functions. For each key, word, the *Reduce* function works on the list of values, partialCounts. To calculate the frequency of each word, the *Reduce* function groups by word and sums the values received in the partial Counts list.

Function [2] reduce (word, List partialCounts) sum = 0

for each pc **in** partialCounts

sum += pc

emit (word, sum)

The final output is the list of words with the count of occurrences of each word. This algorithm is known as MapReduce method.

1.4 KMP Algorithm for Searching

KMP algorithm is based on three author name Knuth, Morris and Prat. It is used for detecting the mismatch. It means if any mismatch occur by comparing two strings, it starts false. This algorithm will helpful for searching.

By comparing two string, KMP algorithm will start matching the pattern of string by letters one by one. If any letter is detected different, it will generate interrupt and tells that word is different from each other.

E.g. First string is 'ABCDE' and second string is 'ABDCE'

Here 'A' and 'B' will pass the string comparison but third letter is different from two strings so with this detection we can say that word is different.

Components of KMP algorithm:

• The prefix function, Π

The prefix function,Π for a pattern encapsulates knowledge about how the pattern matches against shifts of itself. This information can be used to avoid useless shifts of the pattern _p'. In other words, this enables avoiding backtracking on the string _S'.

• The KMP Matcher

With string _S', pattern _p' and prefix function _Π' as inputs, finds the occurrence of _p' in _S' and returns the number of shifts of _p' after which occurrence is found.

The prefix function, Π

Following pseudocode computes the prefix fucnction, Π:

Compute-Prefix-Function (p)

1. m <- length[p] //'p' pattern to be matched
2. Π[1] <- 0
3. K<- 0
4. for q <- 2 to m
5. do while k > 0 and p[k+1] != p[q]

6. do k <- Π[k]
7. If p[k+1] = p[q]
8. then k <- k +1
9. Π[q] <- k
10. return Π

The KMP Matcher:

The KMP Matcher, with pattern ‗p', string ‗S' and prefix function ‗Π' as input, finds a match of p in S.Following pseudocode computes the matching component of KMP algorithm:

KMP-Matcher(S,p)

1. n <- length[S]
2. m <- length[p]
3. Π <- Compute-Prefix-Function(p)
4. q <- 0 //number of characters matched
5. for i <- 1 to n //scan S from left to right
6. do while q > 0 and p[q+1] != S[i]
7. do q<- Π[q] //next character does not match
8. if p[q+1] = S[i]
9. then q <- q + 1 //next character matches
10. if q = m //is all of p matched?
11. then print ―Pattern occurs with shift‖ i – m
12. q <- Π[q] // look for the next match

Algorithm 1: KMP algorithm. [13]

2. RELATED WORK

Sasiniveda.G, Revathi.N [9] author show that with MapReduce, efficiency will result in lowered system costs, energy usage and management complexity it increase the performance of the system and enhanced MapReduce programming model will be evaluated for better performance over cloud.

Samira Daneshyar and Majid Razmjoo [10], shows that MapReduce is an easy, effective and flexible tool for large scale fault tolerant data analysis. Author experiments the MapReduce technique using Hadoop into Amazon EC2 cloud and showing that performance can be better with MapReduce.

Nidhi Grover [1] is giving overview to BIG DATA technology. This paper gives information regarding the architecture for Big Data, issues future work and challenges. Architecture of Big Data shows that how it is working in real time environment. It also defined some issues those will help me to solve those problems in my dissertation. Author conclude that Big Data technologies represent a new generation of infrastructures and technologies those developed in order to mine value from a very large volumes of a wide variety of data by newly enabling high-velocity data capture, discovery, and analysis. It is a term used for large and complicated data sets that are difficult to be processed by standard traditional data processing applications and tools. Big data has established the ability to improve performance, save cost, efficient data processing and better decision-making in diverse fields of application such as traffic control, healthcare weather forecasting, fraud control, media and entertainment, disaster prevention, education etc. Big data poses opportunities and challenges in its application areas that need further significant research efforts. This paper presented a review on the recent efforts dedicated to big data.

Katarina Grolinger[2], Michael Hayes[2], Wilson A. Higashino[2], Alexandra L'Heureux [2], David S. Allison, Miriam A.M. Capretz [2] authors in this paper mapreduce methodology is introducing the solution for the problems arises in BIG DATA technology. Problems in the big data are Big Data tasks types: data storage, Big Data analytics, online processing, and security and privacy. They conclude that traditional data processing and storage approaches are facing many challenges for meeting the constantly increasing computing demands of Big Data. This work engrossed on MapReduce, one of the key enabling approaches for meeting Big Data demands by means of highly parallel processing on a large number of commodity nodes. Issues and challenges of MapReduce faces when dealing with Big Data are identified and categorized according to four main Big Data task types: (1) data storage, (2) analytics, (3) online processing, and (4) security and privacy. Furthermore, efforts aimed at improving and extending MapReduce to address identified challenges are presented. By identifying MapReduce challenges in Big Data, this paper provides an overview of the area, facilitates better planning of Big Data schemes and identifies opportunities for future research.

Jens Dittrich [3], JorgeArnulfo Quian´eRuiz [3] author show in this paper how Hadoop useful for MapReduce methodology. This will help for efficient big data processing. They conclude by providing an all-inclusive view on how to control state-of-the-art approaches (presented in the first three parts) to significantly improve the performance of Hadoop MapReduce jobs. So we will identify open challenges for MapReduce. In next step of future author show us to see how MapReduce will work for Hadoop.

Dr. Siddaraju [4], Sowmya C L [4], Rashmi K [4], Rahul M [4], this authors suggest various methods for catering to the problems in hand through MapReduce framework over HDFS. MapReduce technique has been studied at in this paper which is needed for implementing Big Data analysis using HDFS. They conclude that this paper exploits the MapReduce framework for resourceful analysis of big data and for resolving challenging data processing problems on large scale datasets in different domains. MapReduce delivers a simple way to scale your application. It effortlessly scale from a single machine to thousands, providing Fault tolerant & high performance.

Praveen Kumar [5], Dr Vijay Singh Rathore [5] authors defines in this paper several solutions to the big data problem have developed which includes the Map Reduce environment defended by Google which is now available open-source in Hadoop. Hadoop's distributed processing, Map Reduce algorithms and overall architecture are a major step headed for achieving the promised benefits of Big Data. Authors conclude that Hadoop with its effective DFS & programming framework based on concept of mapped reduction, is a commanding tool to manage large data sets. With its map-reduce programming patterns, overall architecture, ecosystem, fault- tolerance techniques and distributed processing, Hadoop deals a complete infrastructure to handle Big Data.

3. PROBLEM DEFINITION

BIG Data is nothing but a technique which stores the large volume of the data. And to store large amount of data certainly requires more number of hard disk. And there is no

other solution available and ever be possible. What we can do at most is to design an algorithm which will store the file and takes a less amount of space so that we can minimize the number of hard disk required. ZIP or RAR solution is available but it losses the searching capability. Cloud as a whole provides two ways:

- Data Storage
- Data retrieval

User try to retrieve the data on various parameters. If we lose searching capability we cannot provide a user much more convenience even though we are saving space. So we require a solution which saves the space as well as does not harm searching capabilities.

4. EXISTING WORK

Traditional data management with SQL is not enough for Big Data. This data will be in forms of thousands of TB disk spaces. So there are some systems which can handle these data effectively. Here we have got MapReduce technology but this technology is also facing so many problems. MapReduce itself is schema-free and index-free. This provides great flexibility and enables MapReduce to work with semi-structured and unstructured data. Moreover, MapReduce can run as soon as data is loaded. However, the lack of indexes on standard MapReduce may result in poor performance in comparison to relational databases [2]. This may be outweighed by MapReduce scalability and parallelization. So we are proposing an algorithm which will work on idea of MapReduce technology with better management on Big Data over Cloud.

5. PROPOSED WORK

1. To store large volume of data in less space.
2. To store small unit of file without wasting the space. We require to store data not in usual form but in compressed form so that we can keep the block size small. But if we do so it added one more dimension of problem.
3. Searching the content in a compressed file is very in-efficient.

6. PROPOSED FLOW DIAGRAM

Phase 1

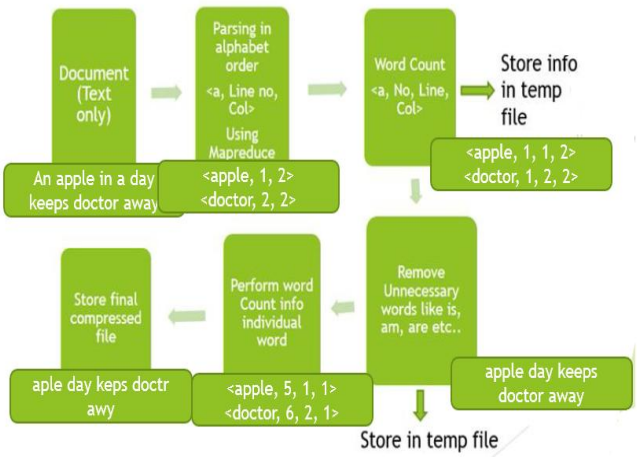

Fig.4 Compression of text file

This is my proposed flow where I will upload text file. This text file is processed by extended MapReduce algorithm and finally it will create the compression form of uploaded text file. E.g. I will upload text file containing text "An apple in a day keeps doctor away." Here 'an', 'in', 'a' etc. will be removed and text file will store as " aple day keps doctr awy". In fig.4 among upper boxes, <apple, 1, 2> shows 'apple' keyword is in 1st row and 2nd column. And then <apple, 1, 1, 2> shows 'apple' keyword occur 1 time in 1st row and 2nd column. In down box <apple, 5, 1, 1> show that 'apple' keyword is having 5 letters and is in the 1st row and 1st column. Final result of uploaded text will be "aple day keps doct awy". Here we are getting more compressed file the original MapReduce work.

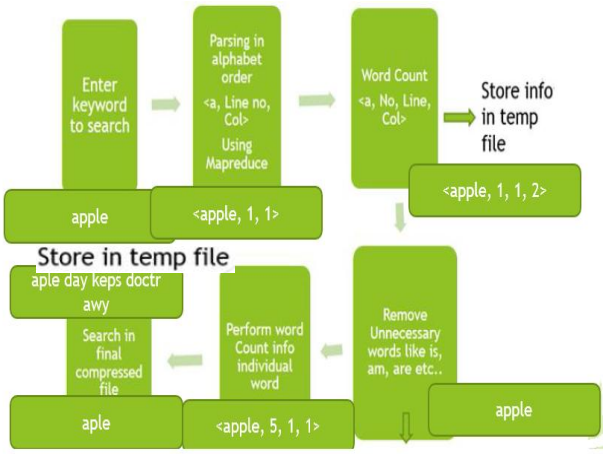

Fig. 5 searching in compressed file

Phase 2

As per shown in fig. 5, I will upload a keyword for searching. It will take same approach for compression method and finally it will compressed and search into the compressed file. E.g. I will upload keyword "apple" for searching. It will compress the keyword into "aple". This compressed keyword will be searched by KMP algorithm in compressed file as I have shown you into Figure 4.

7. CONCLUSION

Big Data is used for large and complicated data sets that are difficult to be processed by standard traditional data processing applications and tools. But large and complicated data sets are difficult to manage so we require MapReduce technology. MapReduce is working with semi-structured and unstructured data but due to lack of indexes on standard MapReduce may result in poor performance. This algorithm will compress data more than the original method of MapReduce function. So my proposed work is to use such algorithm that will solve the problem of storing large data sets in such a way there will not be loss of searching capacity in them. This work will be beneficial to data-centers for storing more large data sets. So we can provide better service for storing data via cloud technology without losing its searching capacity.

8. REFERENCES

[1] Nidhi Grover, "'Big Data'- Architecture, Issues, Opportunities and Challenges", International Journal of Computer and Electronics Research [Volume 3, Issue 1, February 2014]

[2] Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'Heureux David S. Allison, Miriam A.M. Capretz, "Challenges for MapReduce in Big Data", IEEE 2014

[3] Jens Dittrich, JorgeArnulfo Quian´eRuiz, "Efficient Big Data Processing in Hadoop MapReduce", IEEE 2014

[4] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M, "Efficient Analysis of Big Data Using Map Reduce Framework", International Journal of Recent Development in Engineering and Technology, ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014

[5] Praveen Kumar, Dr Vijay Singh Rathore, "Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014

[6] http://www.gethackingsecurity.com/wp-content/uploads/2014/11/cloud_computing.jpg

[7] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), pp. 107-113, 2008.

[8] John F. Gantz, "A Forecast of Worldwide Information Growth through 2010", IDC 2007.

[9] Sasiniveda.G, Revathi.N, "Data Analysis using Mapper and Reducer with Optimal Configuration in Hadoop", International Journal of Computer Trends and Technology- volume4Issue3- 2013

[10] Samira Daneshyar and Majid Razmjoo, "Large-scale data processing using Mapreduce in cloud computing Environment", International Journal on Web Service computing (IJWSC), Vol.3, No.4, December 2012

[11] John F Gantz, "A Forecast of Worldwide Information Growth Through 2010", March 2007

[12] Mr. Chandrapal U. Chauhan, **"**Signature Based Rule Matching Technique in Network Intrusion Detection System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012

[13] Ashish Prosad Gope, Rabi Narayan Behera, "A Novel Pattern Matching Algorithm in Genome Sequence Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014