CLOUD COMPUTING WITH BIG DATA AS A SERVICE

G.Somasekhar
Dept. of Computer Science Engineering,
JNTUA College of Engineering Pulivendula,
AndhraPradesh,India.
somasekhar_giddaluri@rediffmail.com

Abstract – Big data management is becoming crucial now a days because of the evolution of large number of social networking websites, powerful mobile devices, sensors, and cloud computing. According to IDC analysis, the global data volume is going to grow 44 times between 2009 and 2020.It may also go beyond limits such that we cannot control. The existing technology and infrastructure may not support to maintain these large chunks of data. This paper deals with possibilities, vulnerabilities, problems, solutions and advantages of maintaining the bigdata in the area of cloud computing.

Keywords-cloud computing, bigdata, data mining, bigdata processing, bigdata management.

I. INTRODUCTION

Cloud computing is nothing but Information technology as a service. The name cloud computing was inspired by the cloud symbol that's often used to represent the Internet in flowcharts and diagrams. It develops nothing new. It provides IT services via internet to number of clients at low costs .Now a days all the IT companies and their clients are working with enormous amounts of data and the datasize is growing rapidly from time to time as new data is created. This is technically termed as bigdata, maintenance of which became inevitable now. For this, cloud computing is a viable solution. Datamining is the concept connected with both bigdata and cloud computing. In the following sections these three topics and their inter-relation is thoroughly examined.

II. CONCEPTS OF CLOUDCOMPUTING

Through cloudcomputing, the fundamental services offered are the following.

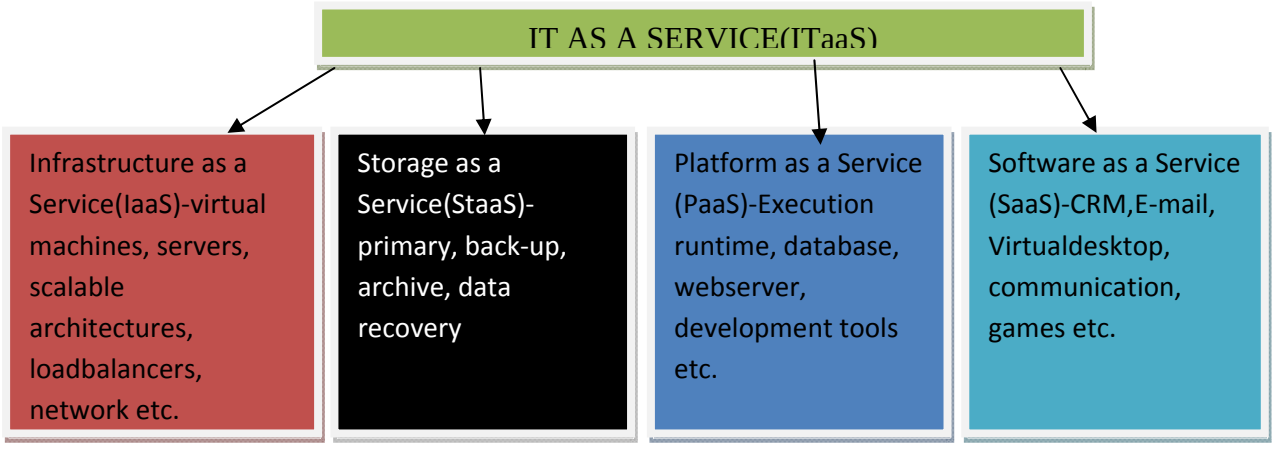

Figure 1.Major services offered by Cloudcomputing in order

- Software as a Service(SaaS)-It is the king of all the services which delivers the application via internet or intranet.

It contains single instance of software running in cloud environment and provides multitenancy.

e.g.,GoogleApps including Gmail,Google Calendar,Docs and sites.

- Platform as a Service(PaaS)-delivers a platform or solution stack on a cloud infrastructure integrating an OS,middleware,application software and even a development environment. May be encapsulated and and served by an Application Program Interface(API).

e.g.,GoogleApp Engine.

- Infrastructure as a Service(IaaS)-delivers computer infrastructure in a virtualized environment, basic storage and computing capabilities for handling workloads. Services include servers, switches, routers and so on.

e.g., Joyent which supports Facebook on it's applications developer program and produce a line of virtualized servers to provide on-demand infrastructure.

- Storage as a Service(StaaS)-provides services like back-up and data recovery.

e.g., Amazon provides DynamoDB service which uses Solid State Drives(SSD) for bigdata storage.

Cloud services may be provided at any of the traditional layers from hardware to applications.

By taking visibility into account, clouds are classified as four types.

- public clouds-can provide large number of benefits to their customers with the ability to scale up and down on demand run by third parties, often hosted away from customer premises.
- private clouds-Services are provided to only trusted users and enterprises and intended for the exclusive use of one client.
- Hybrid clouds-It is the combination of public and private clouds and distributes the applications across both a public and private clouds.
- community clouds-The information is shared among organizations on the cloud managed by themselves or a third party and hosted by service provider.

III. MERITS AND DEMERITS OF CLOUDCOMPUTING

Merits:

- For individuals Cloudcomputing saves money on hardware.
- For enterprises,less expenditure for onsite hardware, capital amount, administration, and maintenance is needed and it is scalable to handle variable business needs.
- For the environment, electricity costs and heat release of local servers is saved. Virtual cloud servers can save on electricity by sharing.

Demerits:

- Monthly fees
- Business data is stored off-site.
- Data is unsafe if our service provider goes out of business. Encoding of data transmission and storage needs are taken into consideration. Then the services will be limited.
- Training is needed for programmers with cloud standards.

IV. CHALLENGES AND VULNERABILITIES WITH BIGDATA

A) The existent cloud infrastructure may not support bigdata:

The traditional clouds must be flavoured with additional features in order to solve the complex problems associated with bigdata. The research is still in early stage to support bigdata in the clouds.

Example Applications:

Table I. Application areas,the bigdata used in,algorithms applied, and computing concepts are summarized.

Application	**Bigdata**	**Algorithms**	**Compute Style**
Scientific study(e.g.,EarthQuake Study)	GroundModel	EarthQuakeSimulation,Thermal Conduction,…	HPC
Internet Library Search	Historic Web Snapshots	DataMining	MapReduce
Virtual World Analysis	Virtual World Database	DataMining	TBD
Language Recognition	Text Corpuses	Speech Recognition	MapReduce & HPC

B) The storage may not be sufficient for handling bigdata:

The conventional structures may not support bigdata. The scalable architectures are one solution to uncover this problem.

C) We need separate analytic software for bigdata analysis:

As the data working with is grown enormously, the maintenance became a bigproblem. The analytic softwares and datamining tools are developed for taking prompt decisions during risktimes. Cloudcomputing is another solution for costreduction to store bigdata and management.

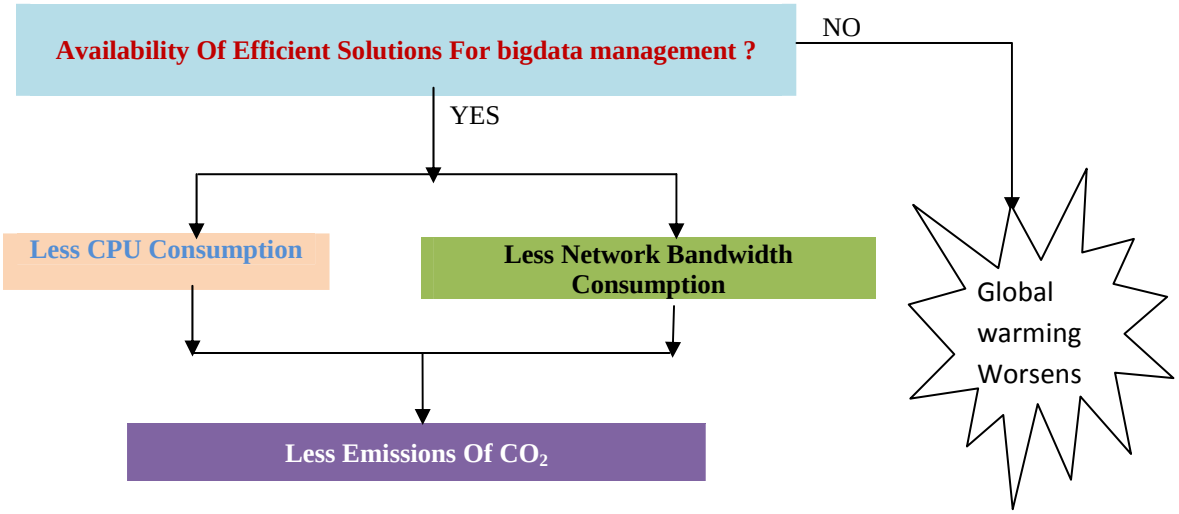

Figure 2. Natural Hazards caused by bigdata

D) Though we efficiently maintain bigdata, the provision of security is difficult in cloudcomputing environment.

Security now a days is a big problem as the existing internet already had been suffering from cracking.China and India are the leading vulnerable countries in the world due to cracking.

E) Maintaining the balance between ethical values and bigdata management is quite hard to achieve.

The very idea of handling bigdata itself invites problems of ethics and it is even hard to accomplish the balance between ethics and bigdata maintenance.

F) Highscale data may lead to global warming.

Heavy usage of CPU and network bandwidth cause global warming which is a big natural hazard for human life

As depicted in Figure 2.

As the globally generated data per year increases by 40%,many companies are looking for a better solution.

To some extent,cloudcomputing is a solution to this problem.

G)bigdata recovery is also a problem.

Handling bigdata is itself a big problem.It needs high storage capacity.If we fail to save this bigdata in the current location where there is the infrastructure to store bigdata,the recovery can become a big problem without sufficient back-up mechanisms.

H)managing large datafiles,managing many files,simultaneous access to files,longterm data management are the problems with bigdata processing.

I)industries rely on simulated data-driven environments rather than physical experiments.

J) Intersect360 research report in setember 2011 states that bandwidth and security are two major hurdles for bigdata and high performance computing.

K) Existing wire and wireless network and frequency management also need diverse techniques for handling complicated and various Big Data.

Amazon is a leader in bigdata and analytics. Clustercompute is Amazon's supercomputer to handle large amounts of data.

V. ADVANTAGES OF CLOUDCOMPUTING WITH BIGDATA

- Heavy cost reduction

"Big Data: The next frontier for innovation, competition, and productivity" of McKinsey [3] describes monthly 30 billion contents are shared through Facebook, and IT expenses have increased 5%. It also includes that Big Data can give 330 billion dollars value production possibility in the area of USA medicine (more than twice of medical expenses of Spain yearly), and 250 billion euro reduction in European public service area (equals the amount of GDP of Greece), and in USA until 2018, 140~190 thousand analysis specialists and 1.5 million data administers can be needed.

- Multitenancy

The cloudservice providers can give benefit to number of clients with lowcost and the clients feel as if they have their own infrastructure, database, middleware and other facilities.

- Business market creation

A big businessmarket can be created through cloudcomputing and it's service management

- speed in bigdata processing

massive volumes of data can be processed at higher speed in cloud environment. Analytics can help in achieving this task.

VI. LAYERED STRUCTURE OF BIGDATA AS A SERVICE

In the cloudcomputing environment which provides bigdata as a service, the structure of the network can be depicted in the Figure 3. At the bottom layer,it is the cloud infrastructure at the lowest level of abstraction providing computing and storage as services.In the next layer upwards it is data fabric service provided by

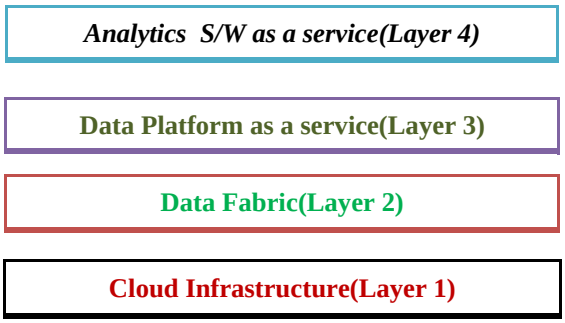

Figure 3. Big Data as a service(Layered approach)

service providers. Certain database management services like data aggregation may be considered as data fabric services. At next higher level, a service provider, in addition to data management, provides execution environment. At the higher level abstraction it is the predictive analytics S/W in the topmost layer as a service. This layer executes scripts, queries and generates reports, visualizations.

VII. PROBLEMS WITH BIGDATA AND HOWFAR CLOUDCOMPUTING CAN BE A BETTER SOLUTION?

Gartner, estimates that, by 2016, 50% of data will be stored on the cloud. In order to maintain such big unstructured, complex, variable data, scalable architectures are designed. The advantage with scalable architectures is, the storage can be added as data grows from time to time. But complex pipelining and parallel processing mechanisms are difficult to manage. Another problem is, due to poor data locality, graph structured data is difficult to handle. Joins and multiple access paths in high level queries, increase concurrency control requirements.

In addition to this, In bigdata processing, manytimes data is moved from storage array to computation cluster and viceversa which is the convention in traditional systems.Now,instead of moving data from storage to work,the reverse process is considered.i.e.moving work to stored data which simplifies the task.The web servers and service oriented applications are difficult to deploy across many machines, which led to cloudcomputing. By the combination of strong processing engine with cloudcomputing architecture, the execution is made effortless without threading and distribution of applications. According to Oracle CEO Larry Ellison, the interesting thing about cloudcomputing is that it is redefined to include everything what we had already done. The existent Data as a Service in cloudcomputing(DaaS) can be extended by appending bigdata processing mechanisms. By scalability and elasticity, clouds are ideal for bigdata analytics.

According to Intersect360 research study, goals of bigdata storage include scalability, throughput, data access, and data sharing. Data fragmentation is a challenge to bigdata storage. It is obstacle to scale storage, cause to performance degrade, requires additional administrative overhead.

Challenges for Clouds to handle bigdata:

- However, supporting bigdata analysis is an architectural hurdle to cloud providers.
- The cloud's distributed nature can be problematic for big data analysis.
- The big problem with clouds is that making the storage perform well, which is basic reason for discarding clouds for bigdata processing.
- Lack of knowledge to use bigdata analytics and lack of training to perform data transactions securely in a cloud environment are the major obstacles to cloudcomputing with bigdata as a service.

VIII. APPLICATION OF DATAMINING ALGORITHMS

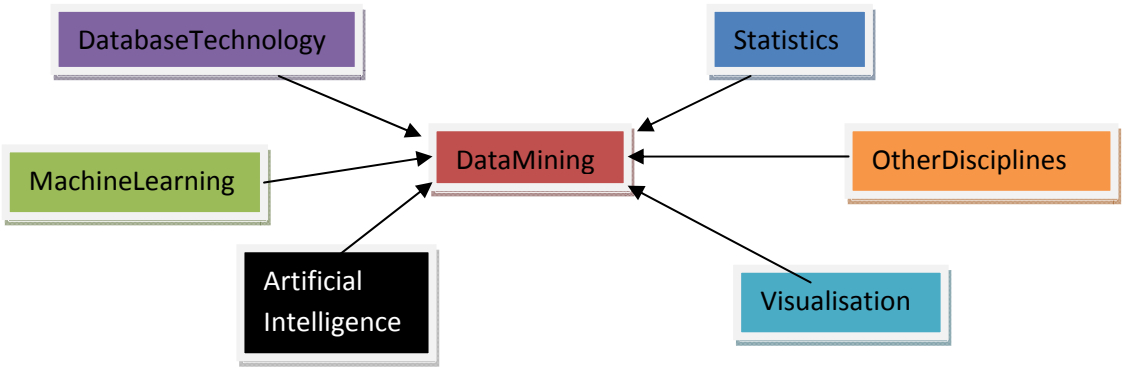

Figure 4.Applications of datamining

The below algorithms in datamining shown in Table II are conventional and also applicable to big datamining. But bigdata takes large volumes of data, there is the need for parallelisation of above traditional datamining algorithms. The strategies for parallelizing datamining algorithms can be in three ways.

A)Independent Search:

In this method, each processor can access the whole dataset. But eachone starts searching from different location in the search space. The starting point may be chosen randomly.

B)Parallelize a sequential datamining algorithm:

In this method, the intermediate results are partitioned across processors. The local concepts are checked for global correctness by taking the entire dataset.

Table II. Illustration of Fundamental Algorithms in datamining

DataMining Algorithm	Example
Classification & Prediction	1)Classify Countries based on climate 2)Classify Cars based on mileage
Outlier Analysis	Detecting Financial Crimes like CreditCard Fraud
Cluster Analysis	Clustering Social networks like Google, Facebook, Twitter etc.
Mining Association Patterns	Finding frequent patterns for improving profitability in any business.

C)Replicate a sequential datamining algorithm:

In this method each processor is allocated to a partition of the entire dataset. All processors work on their local partitions, execute sequential algorithm, and get result which is only locally correct. Finally, the processors exchange their results to check whether they are globally correct or not.

Deployment of some parallel strategies for the mining tasks is the greatest challenge in this area. Protecting privacy of data owners and conducting datamining tasks over federated clouds is another big problem.

IX. LATEST REALTIME SOLUTIONS TO BIGDATA PROBLEMS

In order to reduce delays, data is not moved to and fro for computation and processing, instead analytics and process are moved to data.

Here is one example cloud infrastructure which supports mobile access to file servers, on-demand access, and centralized control. For small business Windows File Server has served a lot in local area networks for using information in certain files by file sharing and file locking facilities for accessing and security purposes respectively. But this is only for users who are locally connected. When remote user needs to use the local file, instead of connecting to Virtual Private Networks(VPN) to access the local file for just modifying and viewing which is not so easy, we can use a hybrid local storage and cloud storage solution. It is of two variations.

- Gladinet cloud[10] with the File Server Agent Component
- Gladinet cloud Enterprise[10] without the File Server Agent Component

After installing File Server Agent Component, our local file can be exposed to online access.Distributed locking is provided with group policy through which local word files and excel files lock automatically on editing.

In 2012,

IBM provided –

- data security solutions for Hadoop and bigdata environments, a special masking scheme for sensitive data, support for latest Key Management Interoperability(KMIP) protocol.
- prevention of unauthorized access and protection against the latest threats with advanced analytics.
- QRadar security intelligence platform to collect, store, analyze and retrieve information on a log, threat, ulnerability and security from distributed locations.

The Cisco Global Infrastructure Services team provided –

- Offering a cloud storage service to any employee who needs to store large unstructured data, such as video.
- A storage cloud service called S-Cloud to store, manage, and protect globally distributed, unstructured content.

CloudSigma built a tiered architecture with Solid State Drives(SSDs) and magnetic storage, for storing large volumes of data.

Some datacenters like Cloud datacenter for IBM,is analyzing the logfiles before dropping them to improve problem anticipation and overall business efficiency of the datacenter.

According to Intersect360 research study, High Performance Computing(HPC)users and bigdata applications prefer private clouds rather than public clouds as private cloud deployment is well suited to bigdata computation.

Aspera on Demand delivers bigdata scale-out transfer capacity.

X. OTHER SOLUTIONS FOR BIGDATA PROCESSING

A)Key-Value stores:

As the data increases queries also increase and key-value stores is a NoSQL strategy to handle many number of queries at a time. This replaces the need for fixed data model and allows to store schema less data.It is always a Process of fast magnetic seek.

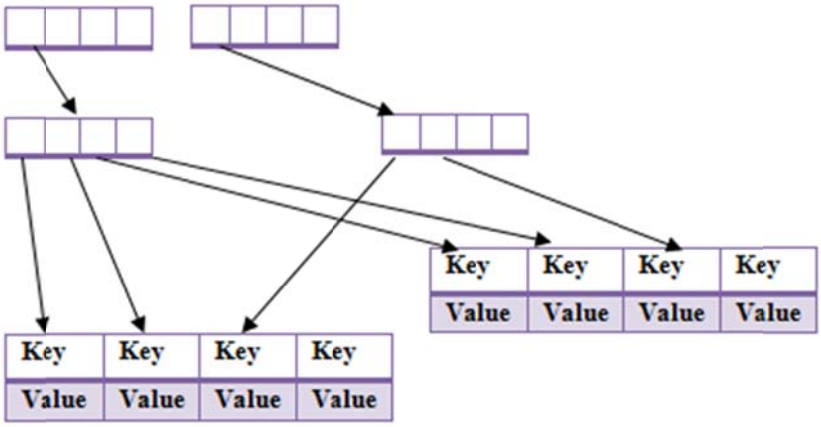

Figure 5: Key-Value stores with B Tree index

B)MapReduce paradigm:

MapReduce is a programming model for processing large datasets with a parallel,distributed algorithm on a cluster as depicted in Figure 6.

Solutions like Hadoop MapReduce can solve complexity problems of bigdata.

C)Tracking the path of access links the strategy to capture the intruder while processing the bigdata in active social networking for security reasons.

D)Scale-up technique is used to store large volumes of data at one end. But it is not a gain when performance is degraded because of huge storage capacity. Then we have to select a reverse strategy called scale-out which sees performance requirements added with storage capacity.

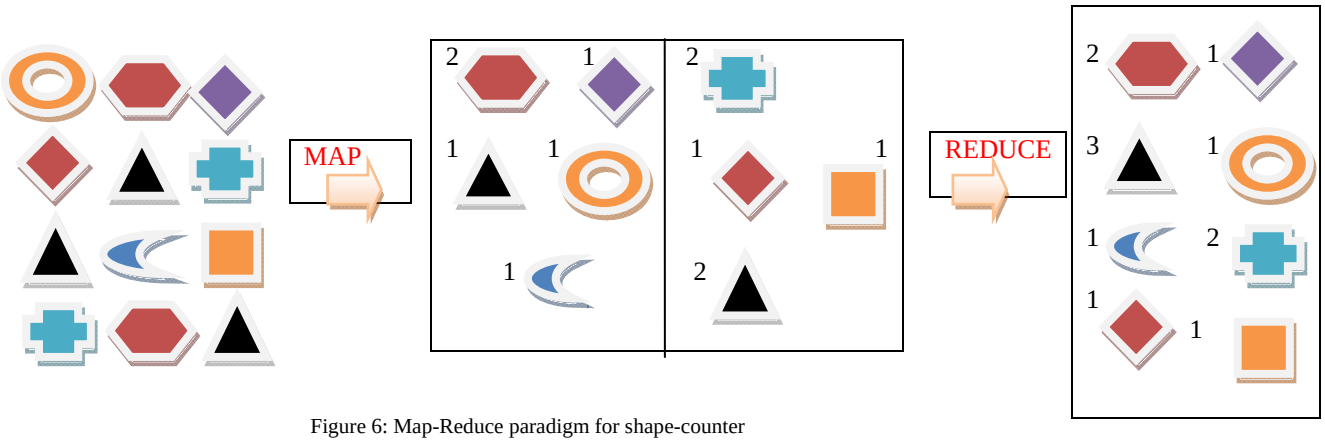

Figure 6: Map-Reduce paradigm for shape-counter

E) The Searchable Encryption System(SES)[1] is another technique when users want a file or document to be confidential. Here, the file or document "D" to be hidden can be accessed via a key provided with certain retrieval helping words as follows.

D = {w1,w2,w3,……..}.

F) GStreamer [2] is a pipeline-based multimedia framework written in C language used for conversion of stream media like video files and audio files.

These solutions for bigdata processing can be supplied on demand by following the decisions of bigdata analysts in cloud environments to provide bigdata as a service.

XI. RECENT TRENDS

Volume, variety, velocity, veracity(The term introduced by IBM means reality and trustworthy) are four characteristics of bigdata. The first two are discussed so far to some extent. In this section we focus on later two.

Now a days there is a vast change in human life. From 3G,4G communications, cell phones and wallets the trend is towards smart phones by which we can do the following tasks.

- making phone calls
- sending texts and e-mails
- sending photos
- tracking locations
- checking bank account balances
- Sending secure payments for consumer transactions and so on.

i.e., we can do almost all the business limited to individuals. For organizations and bigdata centers, handling large volumes of data, to facilitate this bigdata at high velocities,there should be network throughput. Several application areas of big velocity data are, financial services, stock brokerage, weather tracking, movies/entertainment, and on-line retail. Unfortunately, most bigdata analytics are available in static data environment. But, pertaining to realtime events, so far there are no sufficient analytics to support bigdata at high velocities or moving bigdata. Still research is going on towards moving bigdata strategies and companies depend on these studies to gain competitive advantage for their existence as bigdata is given that much importance today.

XII. CONCLUSION

There is a huge movement towards gaining advantage from cloudcomputing which can provide efficiency, speed and flexibility to process bigdata. Another technical term called bigdata analytics as a service is created in this competitive real world, giving importance to analytics. Even small and medium scale business organizations which have not leveraged their data to a certain volume, are also using the bigdata analytics.

Organisations realized that the users are interested in insights of data rather than only data. As bigdata analytics provides insights, it can be future topic for enhancement. Certainly cloudcomputing is going to be a big deal in handling massive amounts of data and data analytics.

ACKNOWLEDGMENT

I would like to express my gratitude to Prof. G.V.N.PRASAD, working as Head Of the department in CSE branch in Sri Indu College of Engineering and Technology,Ibrahim patnam,RR dist.,A.P.,India, for his encouragement to work on this paper.

REFERENCES

[1] N. S. Jho and D. W. Hong, "Technical Trend of the Searchable Encryption system", Electronic and Telecommunications Trends, vol. 23, no. 4, (2008).
[2] GStreamer, http://www.gstreamer.net.
[3] "Big Data: The next frontier for innovation, competition, and productivity" - McKinsey
[4] S. K. Eun, "Cloud Computing Security Technology Trends", Review Security and Cryptology, vol. 20, no.2
[5] Building Data Mining Applications for CRM [Paperback] by Alex Berson (Author), Stephen J.Smith (Author), Berson (Author), Kurt Thearling (Author).
[6] Data-Driven Marketing: The 15 Metrics Everyone in... by Mark Jeffery.
[7] Cloud Computing with the Windows Azure Platform By Roger Jennings
[8] Moving To The Cloud: Developing Apps in the New World of Cloud Computing, By Dinkar Sitaram, Geetha Manjunath
[9] The Cloud Computing Handbook - Everything You Need to Know about Cloud Computing, By Todd Arias Gladinet, www.gladinet.com
[10] Eaton, Deroos, Deutsch, Lapis, & Zikopoulos. (2012). *Understanding big data: Analytics for enterprise class Hadoop and streaming data.* New York: McGraw-Hill

AUTHORS PROFILE

Mr.G.Somasekhar is working as ad-hoc lecturer in Jawaharlal Nehru Technological University Ananthapur College of Engineering, Pulivendla, A.P ,India. He is doing his research on cloud computing. His interesting areas are cloud computing, bigdata management, network security, and data mining. He received his M.Tech. from Sri Indu College of Engineering and Technology, Ibrahim patnam, RR district, A.P., India.