ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

# A Study on Big Data with Cloud Computing and their related Issues

G.Rekha, D.Bhanu Sravanthi

Academic Consultant, School of Engineering & Technology, SPMVV, Tirupathi, Andhra Pradesh, India.

Academic Consultant, School of Engineering & Technology, SPMVV, Tirupathi, Andhra Pradesh, India.

**ABSTRACT:** The term big data arose under the explosive increase of global data as a technology that is able to store and process big and varied volumes of data, providing both enterprises and science with deep insights over its clients/experiments. Cloud computing provides a reliable, fault-tolerant, available and scalable environment to harbor big data distributed management systems. Within the context of this paper we present an overview of both technologies and cases of success when integrating big data and cloud frameworks. Although big data solves much of our current problems it still presents some gaps and issues that raise concern and need improvement. Security, privacy, scalability, data governance policies, data heterogeneity, disaster recovery mechanisms, and other challenges are yet to be addressed. Other concerns are related to cloud computing and its ability to deal with Exabyte's of information or address exaflop computing efficiently. Cloud and Big Data: A Compelling Combination This paper presents an overview of both cloud and big data technologies describing the current issues with these technologies.

**KEYWORDS:** big data, cloud computing, big data issues.

## I. Introduction

In recent years, there has been an increasing demand to store and process more and more data, in domains such as finance, science, and government. Systems that support big data, and host them using cloud computing, have been developed and used successfully (Hashem et al., 2014) . Whereas big data is responsible for storing and processing data, cloud provides a reliable, fault tolerant, available and scalable environment so that big data systems can perform (Hashem et al., 2014). Big data, and in particular big data analytics, are viewed by both business and scientific areas as a way to correlate data, find patterns and predict new trends. Therefore there is a huge interest in leveraging these two technologies, as they can provide businesses with a competitive advantage, and science with ways to aggregate and summarize data from experiments such as those performed at the Large Hadron Collider (LHC). To be able to fulfil the current requirements, big data systems must be available, fault tolerant, scalable and elastic. In this paper we describe both cloud computing and big data systems, focusing on the issues yet to be addressed. We particularly discuss security concerns when hiring a big data vendor: data privacy, data governance, and data heterogeneity; disaster recovery techniques; cloud data uploading methods; and how cloud computing speed and scalability poses a problem regarding exaflop computing. Despite some issues yet to be improved, we present two examples that show how cloud computing and big data can work well together.

Our contributions to the current state-of-the-art is done by providing an overview over the issues to improve or have yet to be addressed in both technologies. Cloud delivery models offer exceptional flexibility, enabling IT to evaluate the best approach to each business user's request. For example, organizations that already support an internal private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of valuable external data sources and applications provided in public clouds. Using cloud infrastructure to analyse big data makes sense because: Investments in big data analysis can be significant and drive a need for efficient, cost-effective infrastructure. The resources to support distributed computing models in-house typically reside in large and midsize data centers. Private clouds can offer a more efficient, cost-effective model to implement analysis of big data in-house, while augmenting internal resources with public cloud services. This hybrid cloud option enables companies to use on-demand storage

DOI: 10.15680/IJIRCCE.2017. 0504181

space and computing power via public cloud services for certain analytics initiatives (for example, short-term projects), and provide added capacity and scale as needed.

Big data may mix internal and external sources. While enterprises often keep their most sensitive data in-house, huge volumes of big data (owned by the organization or generated by third-party and public providers) may be located externally—some of it already in a cloud environment. Moving relevant data sources behind your firewall can be a significant commitment of resources. Analyzing the data where it resides—either in internal or public cloud data centers or in edge systems and client devices—often makes more sense. Data services are needed to extract value from big data. Depending on requirements and the usage scenario, the best use of your IT budget may be to focus on analytics as a service (AaaS)—supported by your internal private cloud, a public cloud, or a hybrid model.

## II. RELATED WORK

**Big Data Trends :**

What makes cloud computing such a cost-effective delivery model for big data analytics? How are big data and cloud technologies converging to make big data analytics in clouds a reasonable option? For big data analytics:

Data is becoming more valuable:Today the conversation is shifting from "What data should we store?" to "What can we do with the data?" Enterprises are looking to unlock data's hidden potential and deliver competitive advantage. Gartner predicts that enterprise data will grow by 800 percent from 2011 to 2015, with 80 percent unstructured (for example, e-mails, documents, video, images, and social media content) and 20 percent structured (for example, credit card transactions and contact information).1

**What Is Big Data Analytics?**

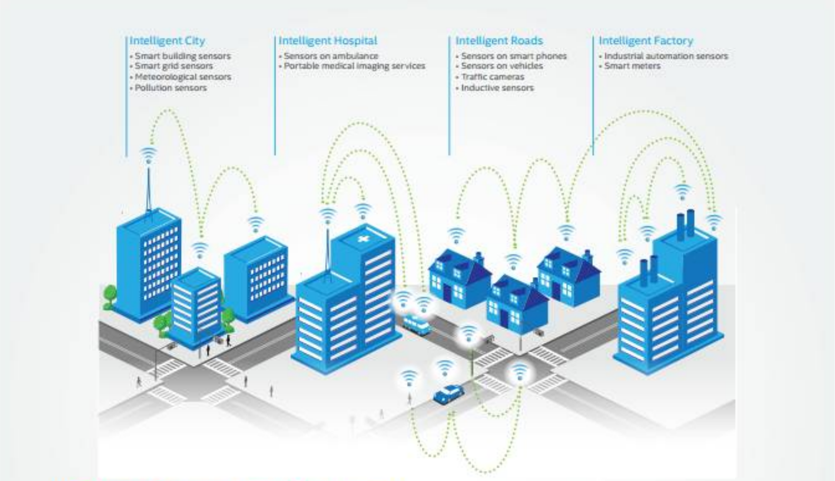
Big data refers to huge data sets that are orders of magnitude larger (volume); more diverse, including structured, semistructured, and unstructured data (variety); and arriving faster (velocity) than you or your organization has had to deal with before. This flood of data is generated by connected devices—from PCs and smart phones to sensors such as RFID readers and traffic cams. Plus, it's heterogeneous and comes in many formats, including text, document, image, video, and more. The real value of big data is in the insights it produces when analyzed—discovered patterns, derived meaning, indicators for decisions, and ultimately the ability to respond to the world with greater intelligence. Big data analytics is a set of advanced technologies designed to work with large volumes of heterogeneous data. It uses sophisticated quantitative methods such as machine learning, neural networks, robotics, computational mathematics, and artificial intelligence to explore the data and to discover interrelationships and patterns.

With the potential for so much data to reveal insights that can boost competitiveness, companies must find new approaches to processing, managing, and analyzing their data—whether it's structured data typically found in traditional relational database management systems (RDBMSs) or more varied, unstructured formats. Plus, combining diverse data sources and types has the potential to uncover some of the most interesting unexplored patterns and relationships.

**Data analytics is moving from batch to real time**. Intel's 2012 survey of 200 IT managers in large enterprises found that while the amount of batch versus real-time processing is split evenly today, the trend is toward increasing real time to two-thirds of total data management by 2015.2 At the same time, the technology for processing real-time or near-real-time information is moving past hype to early stages of maturity.

**Real time supports predictive analytics** :Predictive analytics enables organizations to move to a future-oriented view of what's ahead and offers organizations some of the most exciting opportunities for driving value from big data. Real-time data provides the prospect for fast, accurate, and flexible predictive analytics that quickly adapt to changing business conditions. The faster you analyze your data, the more timely the results, and the greater its predictive value.

**The scope of big data analytics continues to expand**: Early interest in big data analytics focused primarily on business and social data sources, such as e-mail, videos, tweets, Facebook* posts, reviews, and Web behavior. The scope of interest in big data analytics is growing to include data from intelligent systems, such as in-vehicle infotainment, kiosks, smart meters, and many others, and device sensors at the edge of networks—some of the largest-volume, fastest-streaming, and most complex big data. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Interest in applying big data analytics to data from sensors and intelligent systems continues to increase as businesses seek to gain faster, richer insight more costeffectively than in the past, enhance machine-based decision making, and personalize customer experiences.



**Big Data in Context: Smart City Example**

In addition to the transactional, social, and location data generated by people, device sensors generate in real time some of the fastest-growing big data. Processing and analytics can be applied to these valuable data sources via provisioned embedded, cloud, or dedicated IT infrastructure and storage and high-performance computing solutions.

## III. BIG DATA & CLOUD COMPUTING

The concept of big data became a major force of innovation across both academics and corporations. The paradigm is viewed as an effort to understand and get proper insights from big datasets (big data analytics), providing summarized information over huge data loads. As such, this paradigm is regarded by corporations as a tool to understand their clients, to get closer to them, find patterns and predict trends. Furthermore, big data is viewed by scientists as a mean to store and process huge scientific datasets.

This concept is a hot topic and is expected to continue to grow in popularity in the coming years. Although big data is mostly associated with the storage of huge loads of data it also concerns ways to process and extract knowledge from it (Hashem et al., 2014). The five different aspects used to describe big data (commonly referred to as the five "V"s) are Volume, Variety, Velocity, Value and Veracity (Sakr & Gaber, 2014): Volume describes the size of datasets that a big data system deals with. Processing and storing big volumes of data is rather difficult, since it concerns: scalability so that the system can grow; availability, which guarantees access to data and ways to perform operations over it; and bandwidth and performance. Variety concerns the different types of data from various sources that big data frameworks have to deal with.

## IV.BIG DATA ISSUES

Storing and processing big volumes of data requires scalability, fault tolerance and availability. Cloud computing delivers all these through hardware virtualization. Thus, big data and cloud computing are two compatible concepts as cloud enables big data to be available, scalable and fault tolerant. Business regard big data as a valuable business opportunity. As such, several new companies such as Cloudera, Hortonworks, Teradata and many others, have started to focus on delivering Big Data as a Service (BDaaS) or DataBase as a Service (DBaaS). Companies such as Google, IBM, Amazon and Microsoft also provide ways for consumers to consume big data on demand. Next, we present two examples, Nokia and RedBus, which discuss the successful use of big data within cloud environments.

**Security :**Cloud computing and big data security is a current and critical research topic (Popović & Hocenski, 2015). This problem becomes an issue to corporations when considering uploading data onto the cloud. Questions such as who is the real owner of the data, where is the data, who has access to it and what kind of permissions they have are hard to describe. Corporations that are planning to do business with a cloud provider should be aware and ask the following questions: a) Who is the real owner of the data and who has access to it? The cloud provider's clients pay for a service and upload their data onto the cloud. However, to which one of the two stakeholders does data really belong? Moreover, can the provider use the client's data? What level of access has to it and with what purposes can use it? Can the cloud provider benefit from that data? In fact, IT teams responsible for maintaining the client's data must have access to data clusters. Therefore, it is in the client's best interest to grant restricted access to data to minimize data access and guarantee that only authorized personal access its data for a valid reason. These questions seem easy to respond to, although they should be well clarified before hiring a service. Most security issues usually come from inside of the organizations, so it is reasonable that companies analyse all data access policies before closing a contract with a cloud provider. b) Where is the data? Sensitive data that is considered legal in one country may be illegal in another country, therefore, for the sake of the client, there should be an agreement upon the location of data, as its data may be considered illegal in some countries and lead to prosecution.

**Privacy :**The harvesting of data and the use of analytical tools to mine information raises several privacy concerns. Ensuring data security and protecting privacy has become extremely difficult as information is spread and replicated around the globe. Analytics often mine users' sensitive information such as their medical records, energy consumption, online activity, supermarket records etc. Although, doing so might discourage organizations from using de-identification methods and, therefore, increase privacy and security risks of accessing data. Privacy and data protection laws are premised on individual control over information and on principles such as data and purpose minimization and limitation. Nevertheless, it is not clear that minimizing information collection is always a practical approach to privacy. Nowadays, the privacy approaches when processing activities seem to be based on user consent and on the data that individuals deliberately provide. Privacy is undoubtedly an issue that needs further improvement as systems store huge quantities of personal information every day.

**Heterogeneity:**Big data concerns big volumes of data but also different velocities (i.e., data comes at different rates depending on its source output rate and network latency) and great variety. The latter comprehends very large and heterogeneous volumes of data coming from several autonomous sources. Variety is one of the "major aspects of big data characterization" (Majhi & Shial, 2015) which is triggered by the belief that storing all kinds of data may be beneficial to both science and business. Data comes to big data DBMS at different velocities and formats from various sources. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also result in diverse data representations (Wu et al., 2014). Dealing with such a wide variety of data and different velocity rates is a hard task that Big Data systems must handle. This task is aggravated by the fact that new types of file are constantly being created without any kind of standardization. Though, providing a consistent and general way to represent and explore complex and evolving relationships from this data still poses a challenge.

**Data Governance** The belief that storage is cheap, and its cost is likely to decline further, is true regarding hardware prices. However, a big data DBMS does also concern other expenses such as infrastructure maintenance, energy, and software licenses (Tallon, 2013). All these expenses combined comprise the total cost of ownership (TCO), which is estimated to be seven times higher than the hardware acquisition costs. Regarding that the TCO increases in direct proportion to the growth of big data, this growth must be strictly controlled. Recall that the "Value" (one of big data Vs) stands to ensure that only valuable data is stored, since huge amounts of data are useless if they comprise no value. Data Governance came to address this problem by creating policies that define for how long data is viable. The concept consists of practices and organizational polices that describe how data should be managed through its useful economic life cycle. These practices comprise three different categories:

1. Structural practices identify key IT and non-IT decision makers and their respective roles and responsibilities regarding data ownership, value analysis and cost management (Morgan Kaufmann, 2013).

2. Operational practices consist of the way data governance policies are applied. Typically, these policies span a variety of actions such as data migration, data retention, access rights, cost allocation and backup and recovery (Tallon, 2013).

3. Relational practices formally describe the links of the CIO, business managers and data users in terms of knowledge sharing, value analysis, education, training and strategic IT planning. Data Governance is a general term that applies to organizations with huge datasets, which defines policies to retain valuable data as well as to manage data accesses throughout its life cycle. It is an issue to address carefully. If governance policies are not enforced, it is most likely that they are not followed. Although, there are limits to how much value data governance can bring, as beyond a certain point stricter data governance can have counterproductive effects.

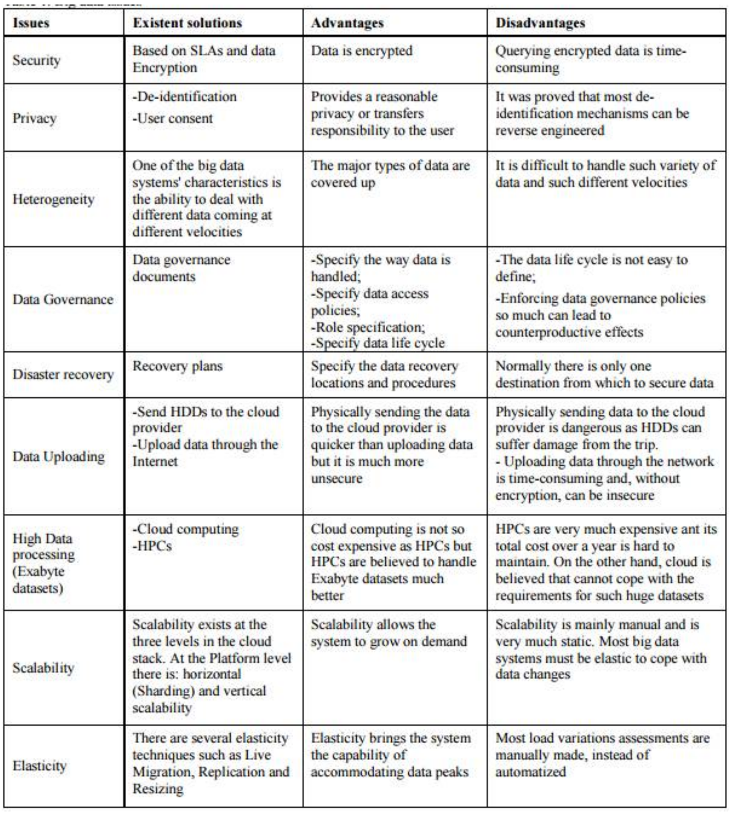
**Disaster Recovery:** Data is a very valuable business and losing data will certainly result in losing value. In case of emergency or hazardous accidents such as earthquakes, floods and fires, data losses need to be minimal. To fulfil this requirement, in case of any incident, data must be quickly available with minimal downtime and loss. However, although this is a very important issue, the research in this particular area is relatively low (Subashini & Kavitha, 2011), (Wood et al., 2010), (Chang, 2015). For big corporations it is imperative to define a disaster recovery plan – as part of the data governance plan – that not only relies on backups to reset data but also in a set of procedures that allow quick replacement of the lost servers (Chang, 2015). From a technical perspective, the work described in (Chang, 2015) presents a good methodology, proposing a “multi-purpose approach, which allows data to be restored to multiple sites with multiple methods”, ensuring a recovery percentage of almost 100%. The study also states that usually, data recovery methods use what they call a “single-basket approach”, which means there is only one destination from which to secure the restored data. As the loss of data will potentially result in the loss of money, it is important to be able to respond efficiently to hazardous incidents. Successfully deploying big data DBMSs in the cloud and keeping it always available and fault-tolerant may strongly depend on disaster recovery mechanisms.

**Other problems** :The current state of the art of cloud computing, big data, and big data platforms in particular, prompts some other concerns. Within this section we discuss data transference onto the cloud; Exaflop computing, which presents a major concern nowadays; and scalability and elasticity issues in cloud computing and big data:

a) Transferring data onto a cloud is a very slow process and corporations often choose to physically send hard drives to the data centres so that data can be uploaded. However, this is neither the most practical nor the safest solution to upload data onto the cloud. Through the years there has been an effort to improve and create efficient data uploading algorithms to minimize upload times and provide a secure way to transfer data onto the cloud (Zhang, et al. 2013), however, this process still remains a major bottleneck.

b) Exaflop computing (Geller, 2011), (Schilling, 2014) is one of today’s problems that is subject of many discussions. Today’s supercomputers and clouds can deal with petabyte data sets, however, dealing with exabyte size datasets still raises lots of concerns, since high performance and high bandwidth is required to transfer and process such huge volumes of data over the network. Cloud computing may not be the answer, as it is believed to be slower than supercomputers since it is restrained by the existent bandwidth and latency. High performance computers (HPC) are the most promising solutions, however the annual cost of such a computer is tremendous. Furthermore, there are several problems in designing exaflop HPCs, especially regarding efficient power consumption. Here, solutions tend to be more GPU based instead of CPU based. There are also problems related to the high degree of parallelism needed among hundred thousands of CPUs. Analysing Exabyte datasets requires the improvement of big data and analytics which poses another problem yet to resolve.

c) Scalability and elasticity in cloud computing and in particular regarding big data management systems is a theme that needs further research as the current systems hardly handle data peaks automatically. Most of the time, scalability is triggered manually rather than automatically and the state-of-the-art of automatic scalable systems shows that most algorithms are reactive or proactive and frequently explore scalability from the perspective of better performance. However, a proper scalable system would allow both manual and automatic reactive and proactive scalability based on several dimensions such as security, workload rebalance (i.e.: the need to rebalance workload) and redundancy (which would enable fault tolerance and availability). Moreover, current data rebalance algorithms are based on histogram building and load equalization (Mahesh et al., 2014). The latter ensures an even load distribution to each server. However, building histograms from each server’s load is time and resource expensive and further research is being conducted on this field to improve these algorithms.



| Issues | Existent solutions | Advantages | Disadvantages |
|---|---|---|---|
| Security | Based on SLAs and data Encryption | Data is encrypted | Querying encrypted data is time-consuming |
| Privacy | -De-identification<br>-User consent | Provides a reasonable privacy or transfers responsibility to the user | It was proved that most de-identification mechanisms can be reverse engineered |
| Heterogeneity | One of the big data systems' characteristics is the ability to deal with different data coming at different velocities | The major types of data are covered up | It is difficult to handle such variety of data and such different velocities |
| Data Governance | Data governance documents | -Specify the way data is handled;<br>-Specify data access policies;<br>-Role specification;<br>-Specify data life cycle | -The data life cycle is not easy to define;<br>-Enforcing data governance policies so much can lead to counterproductive effects |
| Disaster recovery | Recovery plans | Specify the data recovery locations and procedures | Normally there is only one destination from which to secure data |
| Data Uploading | -Send HDDs to the cloud provider<br>-Upload data through the Internet | Physically sending the data to the cloud provider is quicker than uploading data but it is much more unsecure | Physically sending data to the cloud provider is dangerous as HDDs can suffer damage from the trip.<br>- Uploading data through the network is time-consuming and, without encryption, can be insecure |
| High Data processing (Exabyte datasets) | -Cloud computing<br>-HPCs | Cloud computing is not so cost expensive as HPCs but HPCs are believed to handle Exabyte datasets much better | HPCs are very much expensive ant its total cost over a year is hard to maintain. On the other hand, cloud is believed that cannot cope with the requirements for such huge datasets |
| Scalability | Scalability exists at the three levels in the cloud stack. At the Platform level there is: horizontal (Sharding) and vertical scalability | Scalability allows the system to grow on demand | Scalability is mainly manual and is very much static. Most big data systems must be elastic to cope with data changes |
| Elasticity | There are several elasticity techniques such as Live Migration, Replication and Resizing | Elasticity brings the system the capability of accommodating data peaks | Most load variations assessments are manually made, instead of automatized |

Table-1: Big data issues

## V. Conclusion and Future Work

With data increasing on a daily base, big data systems and in particular, analytic tools, have become a major force of innovation that provides a way to store, process and get information over petabyte datasets. Cloud environments strongly leverage big data solutions by providing fault-tolerant, scalable and available environments to big data systems. Although big data systems are powerful systems that enable both enterprises and science to get insights over data, there are some concerns that need further investigation. Additional effort must be employed in developing security mechanisms and standardizing data types. Another crucial element of Big Data is scalability, which in commercial techniques are mostly manual, instead of automatic. Further research must be employed to tackle this problem. Regarding this particular area, we are planning to use adaptable mechanisms in order to develop a solution for implementing elasticity at several dimensions of big data systems running on cloud environments. The goal is to investigate the mechanisms that adaptable software can use to trigger scalability at different levels in the cloud stack. Thus, accommodating data peaks in an automatic and reactive way. Within this paper we provide an overview of big data in cloud environments, highlighting its advantages and showing that both technologies work very well together but also presenting the challenges faced by the two technologies.

DOI: 10.15680/IJIRCCE.2017. 0504181

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

**International Journal of Innovative Research in Computer and Communication Engineering**

*(An ISO 3297: 2007 Certified Organization)*

*Website: www.ijircce.com*

Vol. 5, Issue 4, April 2017

## REFERENCES

1. Popovic. K. & Hocenski,Z. "cloud computing security issues and challenges.",(January),vol.3,Issue 1 pp.344–349, 2015.
2. Zhang.L,"Moving big data to the cloud.",*INFOCOM, 2013 Proceedings IEEE*, vol.4,Issue 2, pp.405–409, 2013.
3. Oracle, "Database as a Service ( DBaaS )" using Enterprise Manager 12c, vol.3,Issue 1,2012.
4. Schomaker "Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons" , Proc. of 17th Int. Conf. on Pattern Recognition (ICPR), IEEE Computer Society, pp. 683-686, vol. 2, 2004
5. Sakr, S. & Gaber, M.M., "Large Scale and big data: Processing and Managemen"t Auerbach, ed., 2014.
6. Majhi, S.K. & Shial, G., "Challenges in big data cloud Computing And Future Research Prospects: A Review". *The Smart Computing Review*, 5(4), vol.3,Issue 1,pp.340–345, 2015.
7. Subashini, S. & Kavitha, V., "A survey on security issues in service delivery models of cloud computing". *Journal of Network and Computer Applications*, 34(1), pp.1–11.

## BIOGRAPHY

**G.Rekha** is an Academic consultant in the Computer Science & Engineering Department, School of Engineering and Technology, Sri Padmavathi Women's University. She received her Master of Technology (M.TECH) degree in 2016 from JNTUA, Anantapur, India.

**D.Bhanu Sravanthi** is an Academic consultant in the Computer Science & Engineering Department, School of Engineering and Technology, Sri Padmavathi Women's University.She received her Master of Technology(M.TECH) degree in 2014 from JNTUA, Anantapur, India.