International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

Technological Prospects of Cloud Computing in Web Mining: Recent Trends and Opportunities

Santosh Kumar Jha [1]
[1] *Assistant Professor –MCA, L. N. Mishra Institute of Economic development and Social Change, Patna, Bihar*
Orcid ID: 0000-0003-2329-0920

ABSTRACT
Web has immense potentials to grow with new technologies and flourish with new opportunities in almost every sphere of human life. The internet grew by WWW and explored with e-business and social network. The Web is known and participating in mining of various kinds of data ranging from users views and patterns to Bitcoin applications.
Web mining includes different set and types of data and extract useful information and from various sources of web and gain knowledge using applicable data mining technique on dataset. Web mining types broadly categorised in three main areas, web uses mining, web content mining, and web structured mining.
Each category of web mining is involved to handle issues of heterogeneous behaviour of web data. All three technique and applications are highly required of high end architectures which can facilitate infrastructures and support for all required criteria. Cloud computing is an emerging technology and blowing with intensive support for varieties of applications. The techniques and applications of Web usage mining are extremely demanded in cloud computing. Cloud computing allow these technique to retrieve relevant and useful data through virtually integrated mode of data warehouse. It helps the users to reduce cost and infrastructure for implementation. This paper presents methodologies of web mining using Cloud Computing technology and its prospects.
Keywords: Web Uses Mining, Web Content Mining, Web Structure Mining, MapReduce Model, Cloud Computing.

1. Introduction
The Internet and WWW have changed the computing world from parallel computing to distributed computing, and now from grid computing to cloud computing, The Web has immense potential and web mining has a major role to discover knowledge from valuable information from the web.
For comparing Web data mining with traditional data mining, the three main differences are main parameters to consider-
(i) Scalability- In traditional data mining, data processing are around 1 million records from a database, whereas data records are large enough exceeding 10 million pages in web mining.
(ii) Accessibility- the data is privately made available and often requires access rights to read in the processing of data mining tasks. However, the data is public in web mining process, and hardly requires access rights for processing.
(iii) Structure – In traditional data mining task, information are accessed from a database, which provides some level of explicit structuring of data. However, in a typical web mining task, unstructured or semi-structured data are processed from web pages. In case, the underlying information for web pages comes from a database, this, too, is masked by HTML markup [1, 2].

1.1 WEB DATA MINING
Web data mining techniques are the outcomes of a longer process of research and product advancement. Web data mining is based on knowledge extracted from the Web. It targets to discover useful information or knowledge from Web hyperlinks structure, page contents, and usage data [4]. Many new mining tasks and algorithms have been designed in the past decade. On the basis of the

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

types of data used in the mining, Web data mining activities can be framed into three types as shown in Fig. 1.

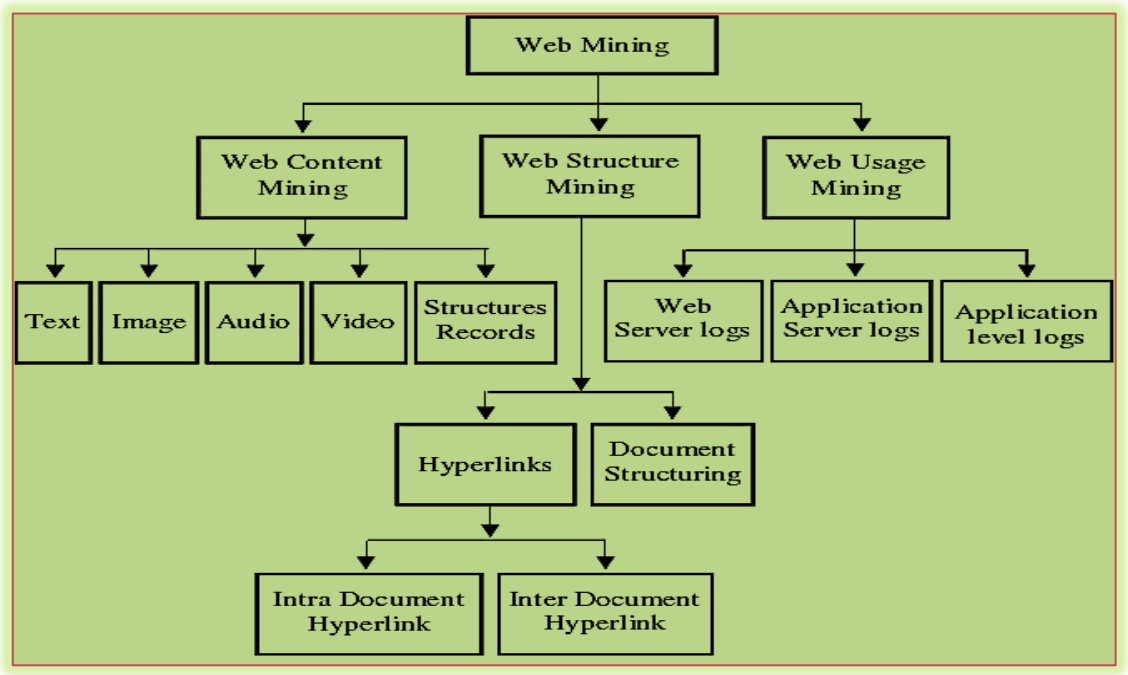

Fig. 1. Web mining categories

1.2 CLOUD COMPUTING

Cloud computing is a type of Web-based computing platform which provides on-demand delivery of resources that include computing applications, storage, database, server, software and infrastructure as services facilitated to customers as per their requests over the Internet. Major types of cloud computing are broadly categorised in two models deployment model and service model as follows:

Types of Deployment Model

Deployment model is categorised mainly in the following three types:

I. **Private cloud** – This model especially offered over a private network for an organization having more control and security. It is designed for the use of single organisation and provides highest level of security.

II. **Public cloud** – This model is suitable for anyone to the universal public and offered by third-party service provider over the Internet. It is free or pay-per-usage and known as reasonably priced model globally.

Hybrid cloud – This model is a combined approach of two or more clouds comprising private, community or public. It is known as "best of both worlds" model. It can be joined by private and public, achieved by real-time demand, and/or devoted services with cloud resources.

Types of Service Model

The service models of Cloud computing are mainly categorised in three major types:

I. Infrastructure-as-a-service (IaaS)

It relates to sharing infrastructures and provides resources in terms of server, network, and virtualization. The user accesses all these through apps and responsible for managing apps OS and middleware. All the functionalities such as data storage and other resources are controlled and managed by the service provider, through the Internet. Users pay only for the demanded services and no need to purchase licence, hardware and software. Figure 2 is shown as an example.

II. Platform-as-a-Service (PaaS)

PaaS facilitates the users a shared cloud platform. It helps for application development and management. Besides, IaaS application layer, the user needs for other requirements:

- Runtime Environment for Applications (an important DevOps component)

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

- Development and Data Processing Platforms using PaaS

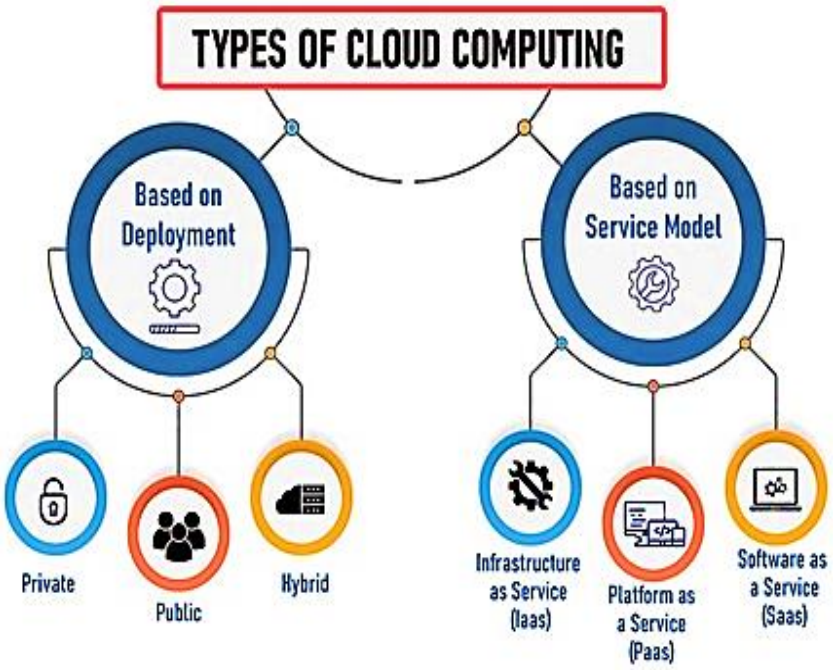

Fig. 2. Major types of Cloud computing

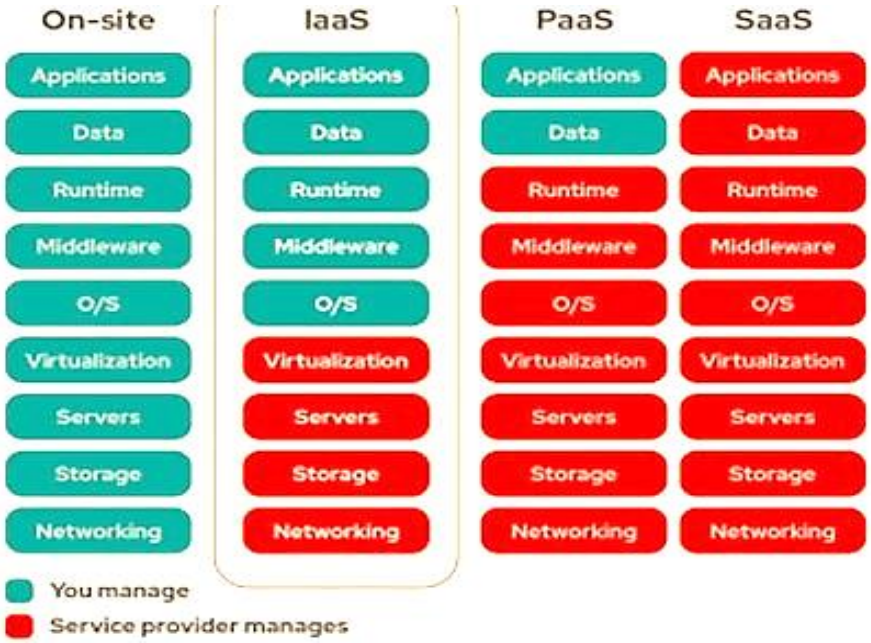

Fig. 3. Cloud Services at different models

III. Software-as-a-Service (SaaS)

SaaS is delivering software applications to the users and make them free to install on PC or mobile. These services are remotely offered as a Web based service nearly at no costs.

Moreover, one more model is gaining popularity, known as **Function-as-a-Service (FaaS)**. This model is used to execute code in response to events. Infrastructural starvation is managed in this model.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

2. Literature Review

I have studied various articles of web usage mining. Some of these are: **Data Mining, Web Data Mining Techniques, Tools and Algorithms: An Overview:**
This paper highlights to mine Web usage data on client side, whereas majority of papers focus on the server side Web usage mining. In the mining of Web usage data at client side, the more complete knowledge about Web usage can be obtained [5][19].
Cloud Computing Use for Web Mining: This paper mainly presents the technology of cloud computing used for implementing web mining. Web mining is the process of extracting useful information using data mining techniques from the web and gaining knowledge from the mined data. There are numerous online resources made available on web content mining, web structure and other types and access through the web servers [9].

3. Research Gap & Research Problem

3.1 Drawbacks in the Existing Methodologies

1. It is observed by the users that the response time is too long.
2. A heavy demand on networking is figured by the continuous explosive growth of the Web.
3. Multiple resources and Web servers are highly demanded.
4. One solution to improve the quality is to increase the bandwidth, but it is hampered by increasing economic cost of Web services.
5. While working with catching system, Web caching system has three major drawbacks:
(i) proxy updating is not properly managed,
(ii) stale data might be received by a user, and,
(iii) origin servers typically become hold-ups in the case as the number of users grows.
6. The ideal effectiveness of Web catching is weakened by various factors. The bottleneck is its limited system resources of cache servers. These resources are memory space, processing capacity, disk storage, various network resources and I/O bandwidth. In addition, some other factors are momentous problems even if the cache space is provided unlimited. Basically the problem of updating a voluminous Web objects are unmanageable that a large cache cannot solve it.
7. Main drawbacks of systems are related with prefetching scheme which increase the network traffic and the load of Web server. The reason is its enhanced prefetching policies that some pre-fetched objects may be accessed which eventually not requested by the users. [10]

3.2 Cloud computing challenges

Cloud computing moved away from client node and an enterprise application server to cloud services as an ultimate. The superficial appearance of cloud computing has made a remarkable impact on the industry. At the moment, industry desires Cloud services for best opportunities to the society.

3.3 Cloud computing issues

We get any new technology by sacrificing, and, the implementation of cloud computing is not free from such issues. There are some important prevalent factors given as follows:

- Security and Privacy concerns
- Service quality promised
- Interoperability balances,
- Portability, Performance and Bandwidth Cost to the requirements
- Improper Cloud Service Exit policy
- Multi-jurisdictional issues
- Multi-tenancy, Virtualization

4. The Objectives of the study

The primary objective of this paper is coined to discuss the major issues in web data mining applications. First is heterogeneous behaviour of web data. That, increasing the uses of web and mobile applications generates huge amount of data and information. These are found as

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

interconnected web pages. Web pages provide data in three different forms- structured, semi-structured and unstructured. Second is dynamic nature of web data and web applications. Third is searching and finding patterns form web in the form of web uses mining, web content mining, and web structure mining [3]. At the end, the research scope of web mining in the arena of cloud computing is also discussed.

5. Research Methodology/Design/Approach/Hypothesis

Key Reasons behind the Adoption

These days, the Cloud mining is made an alternative approach and adopted for the following main reasons:

a. Increased reliability and dependability on web and mobile applications.
b. Support of increased security which is well trusted by the customers.
c. Increased and improved bandwidth and Low cost of ownership.

5.1 Web Mining and Cloud Computing

Web Usage Mining is known as one of the popularly used technologies in Web Mining [1]. Now a days, the reduced cost efficiency and flexibility of the Cloud computing has enforced to adapt it for the Web Usage mining [11]. However, Cloud computing is termed as an emerging new IT procurement model. It has made an improved movement and attention to accept the cloud computing as continual concerns, considerably in the Web mining. Profoundly, the Cloud Mining is novel approach to express search interface for our data. The security issue is a major challenge when SaaS (Software-as-a Service) is offered for web mining. The SaaS is used for slashing the cost and designated as a cloud mining technique. There is a need of an influential framework of web mining that Hadoop and MapReduce communities solve for projecting analytics [12].

Now, I discuss in the next section the use of MapReduce Model in Cloud Computing and various benefits of using this model for the Web mining.

5.2 Cloud Computing and MapReduce Model

Cloud computing has gained its popularity as the most captivating areas which allows lots of services to be utilized. The Cloud computing facilitates us to fully utilize the various resources as needed and disseminate at various places [13]. A programming model, proposed by Google, is known as the MapReduce model. The MapReduce model is mainly used for processing voluminous data sets. These days, MapReduce model is gaining more popularity in cloud computing [14][15]. In a single day, this model processes around 20 petabytes of data. The MapReduce model is used for parallel and distributed processing of massive data sets on clusters [16]. Various popular applications of MapReduce frameworks include the following:

At Google:

- Used for the Index building for Google Search
- Applied in the Article clustering for Google News
- Assisting in the Statistical machine translation

At Yahoo!:

- Used for the Index building for Yahoo! Search
- Applied in the Spam detection for Yahoo! Mail

At Facebook:

- Used for the Ad optimization
- Applied in the Spam detection

5.3 System Framework of MapReduce Model

The basic architecture of MapReduce is mentioned in Figure 2. The MapReduce involves mainly two basic phases [17]:

Map: It is used to perform filtering and sorting of datasets and
Reduce: It is used for performing a summary operation.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

In the MapReduce framework, the input and output are provided in the way of key-value pairs. First, the input data is partitioned into segments of applicable size. After that, the Map process takes a series of key-value pairs and generates processed key-value pairs. These processed key-value pairs are supplied to a particular reducer by using a certain partition function. Next, the data are sorted and the shuffling process is done. After shuffling, the Reduce procedure integrates the results. The MapReduce model achieves scalability using a large volume of CPU with low implementation costs. The selection of a single server or multiple servers is a smart proposition of this model.

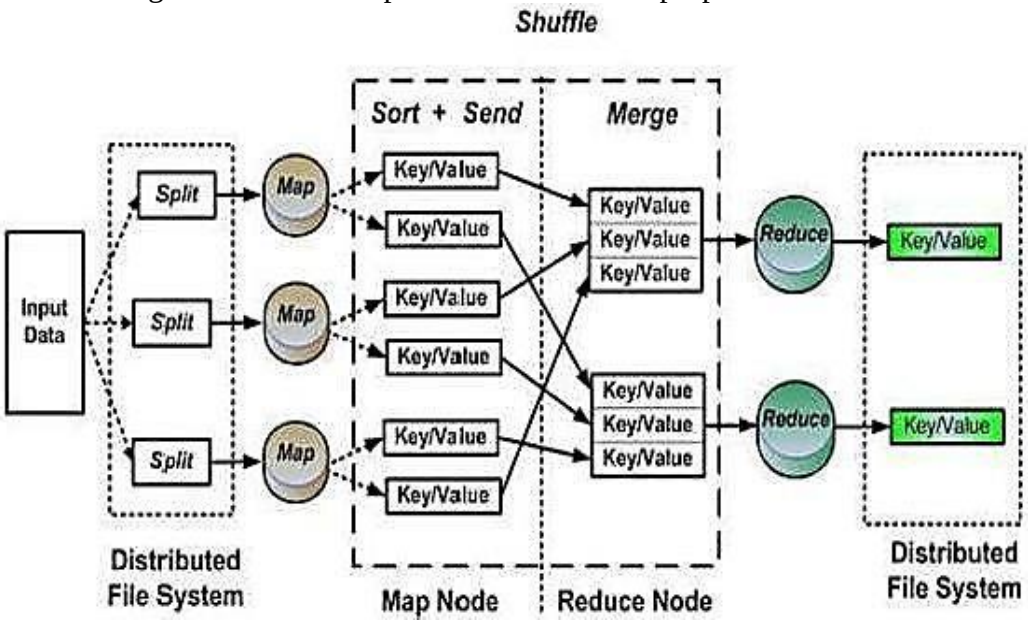

Fig. 4. MapReduce system framework

5.4 Advantages of MapReduce Framework:

The main advantage of the MapReduce framework is its fault-tolerant capability that reports from each node in the cluster, periodically. It helps the master node to notice if any node is delayed or silent for a longer interval as expected, then it helps the main node to process the reassignment of the delayed tasks. Some other main advantages of this model are given as follows:

i. **Scalability and Distributed Processing:** The scalability of the MapReduce model is achieved by using the Hadoop platform. The Hadoop platform has the capability to accumulate and distribute large data sets across ample of servers that operate in parallel which leads to reduced costs.

ii. **Flexibility:** This model is suitable for any type of data, i.e., Structured and Unstructured data from a variety of sources mainly gathered from email, e-commerce, social media, and other web-enabled data.

iii. **Fast:** This framework works on a Distributed architecture which makes it capable to process huge amounts of data ranging from Terabytes to petabytes. It is too fast in processing that it takes minutes to process terabytes of data, and petabytes of data can be processed in hours only.

iv. **Security and Authentication:** Security is a major area of concern in almost every field. The MapReduce framework works with HDFS and HBase security which allows only authenticated users to access the data.

CONCLUSION

Mining of web data using the Cloud permits companies to easily manage their resources with effective, reliable, and secure services for their users. Cloud Computing is the union of services that is definitely one of the widely used technologies and provides cost-efficient and flexible services to users. Reliably, Web Usage Mining uses Cloud Computing Service mainly SaaS (Software as a Service) to improve security as well as it reduces overall cost. I have discussed the basic MapReduce model and its advantages in this paper, however, future work will have attention to a different fresh

DOI:10.46647/ijetms.2023.v07i01.017 ISSN: 2581-4621

approach to explore the current model. There is a high need for a more precise and faster approach to discovering the Web Usage mining relies on the Cloud Computing approach.

Reference

1. Web Data Mining Exploring Hyperlinks, Contents, and Usages Data By Bing Liu Published by Springer
2. Georges Dupret, Mounia Lalmas, 2013, Absence Time and User Engagement: Evaluating Ranking Functions,WSDM'13, February 4–8, 2013, Rome, Italy
3. Hicham Snoussi, Laurent Magnin, and Jian-Yun Nie, Heterogeneous Web Data Extraction using Ontology
4. R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, 1997, pp. 558-567.
5. "A Framework for Personal Web Usage Mining".
6. Kaikala Anjani Sravanthi1, "Web Mining Using Cloud Computing" , ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 4, April 2013.
7. Web mining, "www.wikipedia.com", 10-02-2015.
8. Rajni Pamnani, Pramila Chawan Department of computer technology, VJTIUniversity, Mumbai "Web Usage Mining: A Research Area in Web Mining".
9. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for we personalization. Data Mining and Knowledge Discovery, 6(1):61{82, January 2002.
10. Faten Khalil, Jiuyong Li and Hua Wang ―A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses‖ ,Proc. Fifth Australasian Data Mining Conference (AusDM2006), CRPIT Volume 61,177-184.
11. Robert Grossman , Yunhong Gu, "Data mining using high performance data clouds: experimental studies using sector and sphere", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2008
12. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using modelbased clustering. In In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 280{284, Boston, Massachusetts, 2000.
13. Zhu, W., & Lee, C. (2014). A new approach to web data mining based on cloud computing. Journal of Computing Science and Engineering, 8(4), 181–186. doi:10.5626/jcse.2014.8.4.181
14. "MapReduce." Wikipedia. N.p.: Wikimedia Foundation, 11 Jan. 2017. Web. 2 Jan. 2017.
15. Divestopedia, and Securities Institute. What is MapReduce? - definition from Techopedia. Techopedia.com, 2017. Web. 2 Jan. 2017.
16. Posted, and Margaret Rouse. What is MapReduce? - definition from WhatIs.com. SearchCloud Computing, 25 June 2014. Web. 2 Jan. 2017.
17. Hornung, T., Przyjaciel-Zablocki, M., & Schätzle, A. (2017). Giant data: MapReduce and Hadoop » ADMIN magazine. Retrieved January 10, 2017, from http://www.admin-magazine.com/HPC/Articles/MapReduceand-Hadoop
18. Lee, K.-H., Lee, Y.-J., Choi, H., Chung, Y. D., & Moon, B. (2012). Parallel data processing with MapReduce, ACM SIGMOD Record, 40(4), 11, doi:10.1145/2094114.2094118.
19. Muhammd Jawad Hamid Mughal. Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview- (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018.