iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** X **Month of publication:** October 2019

DOI: http://doi.org/10.22214/ijraset.2019.10115

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

Clustering EDP (Error Detection Program) Errors from Cloud Data Centres using Data Mining

K. Arunraj Bapuji[1], Prof. A. Vinayababu[2], B Anand Kumar[3], Mallaih Vorsu[4]
[1, 4]*Acharya Nagarjuna University, Guntur*
[3]*Rayalaseema University, Kurnool*

Abstract: ***Cloud computing emerging as a vital resource of every business firm in the world, as a result there is a tremendous usage and movement of data from the cloud data centres to client on 24x7. This phenomenon generates many types of errors or faults at various stages of cloud computing process. The errors or faults are identified and resolved using EDP (Error Detection Program) part 1[14] and stored in simple text file using many attributes for further processing [15]. In this paper author proposing a mechanism to import the errors from the simple text file using python script and an automated SQL query will now export the data into a table in the MySQL data base. All the errors are stored as per the priority generated by a SQL query. In the next paper author applies various data mining techniques such as clustering, classification, forecasting on MySQL database to alert the various cloud administrators to prevent the errors or faults.***
Keywords: Data Ming, K-means, clustering, classification, cloud computing, patching, kernel.

I. INTRODUCTION

The cloud error data has been stored in notepad file (eror.txt), the error data will be segregated into table format for further action on the data. In that segregated table of data will be tested by using MySQL query's in different formats. The error data used to forecast, analyse and to choose a command to resolve the error in a smooth way for accurate result and to take a decision to update a server on the cloud environment.

Cloud Computing is internet-based computing which is used to share resources software's and information will be supplied to user and to customers and also to other devices with on demand services. [3] And these are categorized into three basic service models and deployed models which are described below.

A. Service Models

1) *Infrastructure-as-a-Service [IaaS]:* IaaS provides access to fundamental resources such as physical machines, virtual machines, virtual storages, etc.
2) *Platform-as-a-Service [PaaS]:* PaaS provides the runtime environment for applications, development tools, etc.
3) *Software-as-a-Service [SaaS]:* SaaS model allow using software applications as a service to end users.

B. Deployed Models

Different kind of deployed models are provided by different service providers like IBM, Amazon, Microsoft, google.

1) *Public Cloud:* A public cloud allows using cloud environment to all the users to use their services public. Public cloud is good choice for start-up companies; because they can start their company by hiring the resources of Public cloud without any own IT infrastructure. In this model service provider will take care about IT infrastructure and support. Public cloud is less secure because resources accessed by the 'n" of users. [4]
2) *Private Cloud:* Private cloud is allowing all the services to the end users with in organization. Private cloud can be maintained or host by internal or service providers and service providers support in providing secured data commutation for 24/7. This cloud is more secure because it is accessible with in orgazation. Private cloud is good choice for who maintains secure date for their customers. And once private cloud is established it is not much difference for public cloud. [4]
3) *Hybrid Cloud:* Hybrid cloud is combination of both public and private cloud, whereas private cloud will perform all the secure and critical activities and public cloud will take care of all the non-critical activities. Most of the hybrid cloud service providers will ensures that always resources available, because it is combination of both public and private cloud. [4]
4) *Community Cloud:* Community cloud is like a public cloud, but it is only accessible to specific community. This cloud can be maintained by either service provider or own, but they have limited access to the public cloud. It is only accessible to the community members or outside members of the community also allowed to get access to the IT resources. [4]

©IJRASET: All Rights are Reserved

C. Data Mining

Data mining is process of extracting useful information from raw data. In cloud computing most of data will be unstructured. Data mining can help in cloud computing in extracting unstructured data to structured data. In could compute all the software's, storage servers and network maintained are centralized. Data mining will help to maintain these entire centralized infrastructures and secure reliable all services for users of the cloud.

Data mining is highly efficient to extract structured data from unstructured data for a service or task. And data can be from various sources or platforms. Data mining uses two kinds of models 1. Descriptive 2. Predictive [1].

Descriptive model generally used to characterize the general properties of the data in database [1]. Predictive model if performing inference on current data to make prediction [1][2]. Both models can archive different variety of data from raw data which show in in bellow figure [1].

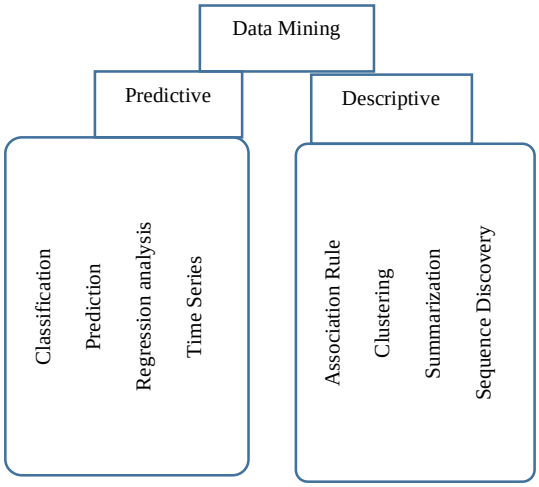


Figure: 1.0 Datamining

of analysing and extracting useful data in various fields where human interaction is available. This helps to all cloud customers to get their valuable information for just click on one button.

D. Data Mining Techniques in Cloud Computing

There are several virtual areas where data mining algorithms can be applied in cloud computing. And or other model can be applied depends up on model and architecture of cloud environment. Some of models useful in cloud environment: Classification, Prediction, Regression analysis, Summarization, Clustering [5].

Clustering: K-means is popular algorithm which is mostly used to analyses the real worked data. K-means algorithm tries to group the items in data set into desired number of clusters. To perform this task, it makes some interaction until it converges. After each interaction, calculated means are updated such that they become closer to final means [7]. Also, K-Means algorithm can efficiently extract data from large data bases and cloud environment for storage large data base without any cost [8].

E. Faults in Cloud Computing

There are various faults which can occur in cloud computing which are very critical in real time which may lead to SLA (Service level agreement). Some of them discussed below.

1) *Network Fault:* A Fault occur in a network due to network partition, Packet Loss, Packet corruption, destination failure, link failure etc.
2) *Physical Faults:* This Fault can occur in hardware like fault in CPUs, Fault in memory, Fault in storage, etc.
3) *Media Faults:* Fault occurs due to media head crashes.
4) *Processor Faults:* fault occurs in processor due to operating system crashes, etc.
5) *Process Faults:* A fault which occurs due to shortage of resource, software bugs, etc.
6) *Service Expiry Fault:* The service time of a resource may expire while application is using it.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

II. PREVIOUS WORK

Virendra Singh Kushwaha, Sandip Kumar Goyal & Priusha Narwariya, paper investigated a method for fault-tolerance in load balancing schemes in the cloud environment. This paper mainly focused on various fault tolerance implementations in cloud computing during the load balancing and they concluded one of the best and low-cost technique in the cloud computing. [6]

Mehdi Effatparva, Seyedeh Solmaz Madani, proposed a model in which, results are analysed by CPN tools and demonstrated the degree of reliability. This proposal concluded that, when there is an increase of requests, the FT is reduced, and same way reliability is also reduced and vice versa. The proposed method was implemented by using byzantine fault tolerance by using colour Petrinets simulant in interconnected network clouds, which is used to evaluate the consistency. This method is used because it has strong mathematical support. Results which are obtained from simulator they are compared with optional resources and they found same method is used in both-end. [7]

In recent years Mr Iqjot Singh, Prerna Dwivedi, Taru Gupta and P. G. Shynu team worked on, "Enhanced K-means clustering with encryption on cloud", they provided a solution to upload and download large data files with ensuring security in Hadoop and bigdata by using hashing in cloud environment.

They described that Hadoop is open source software which can store and manage large files in cloud environment. K-means clustering algorithm is an algorithm used to calculate distance between the centroid of the cluster and the data points. Hashing is algorithm in which we are storing and retrieving data with hash keys. The hashing algorithm is called as hash function which is used to portray the original data and later to fetch the data stored at the specific key**.** [9] After execution they concluded, by adding hashing them able to access files faster and with the help of encryption, the data stored in the HDFS is safe and secure. Degloved algorithm will not create load on the overall system and smooth retrieving is enabled to the user through which he can get the desired and applicable output. [9]

Mr. Sudhir M. Gorade1, Prof. Ankit Deo2, Prof. Preetesh Purohit, worked on “A Study of Some Data Mining Classification Techniques", they proposed a study of various data mining classification techniques like Decision Tree, K- Nearest Neighbour, Support Vector Machines, Naive Bayesian Classifiers, and Neural Networks. [10] They concluded that several classification techniques in datamining are available and they have their own advantages and disadvantages. Decision tree classifiers, Bayesian classifiers, classification by back propagation, support vector machines, and some of them are eager and lazy learner by tuples to construct a generalization model. nearest-neighbour classifiers and case-based reasoning. These store exercise tuples in design space and wait until presented with a test tuple before execution simplification. [10]

Renu Asnani worked "A distributed k-mean clustering algorithm for cloud data mining" she proposed a survey paper for cloud data storage and their utilization in various applications and databases, in addition of that a new model of cluster investigation of data is proposed which provides the clustering as service. [12] She concluded paper provides a complete overview of data mining, cluster analysis of data, recently developed different techniques available for clustering and their applications are learned and lastly the key objectives are established and a new model for sentiments-based text clustering data model is proposed. [12]

Krishan Rohilla, Shabnam Kumari, Reema, worked on "Data Mining based on Hashing Technique", they proposed an association rule based on the concept of Hardware support.

They first maintain the database and compare it with systolic array after this a pruning process is being performed to filter the database and to remove the rarely used items. Lastly, data is indexed according to hashing technique and the result is achieved in terms of support count. [13] They concluded that hash based pipelining technique products in market can be sold earlier because in HAPPI technique it removes bottleneck problematic, thereby providing earlier throughput and our sales process becomes faster because due to indexing hashing process converts faster. [13]

III.PROPOSED MODEL AND IMPLEMENTATION.

EDP (Error Detection Program) error.txt file has been inputted which is generated from the various cloud servers of the cloud centres worked on previous papers, [14,15] which contains the various columns such as customer name, IP address, host name, server connection status, OS version, date and time info, server uptime, kernel version, CPU architecture, cores, memory & swap, OMM messages, file system information, yum repo list, latest installed packages, duplicate packages, conflict packages, rpm db problems, last backup date, VMware tool status, network routes, network ethernets network, gateway access, NFS service status, group volume, physical volume, logical volume, filesystem mounted, yum summary, error status, comments.

The proposed methodology contains various components as described below.

©IJRASET: All Rights are Reserved

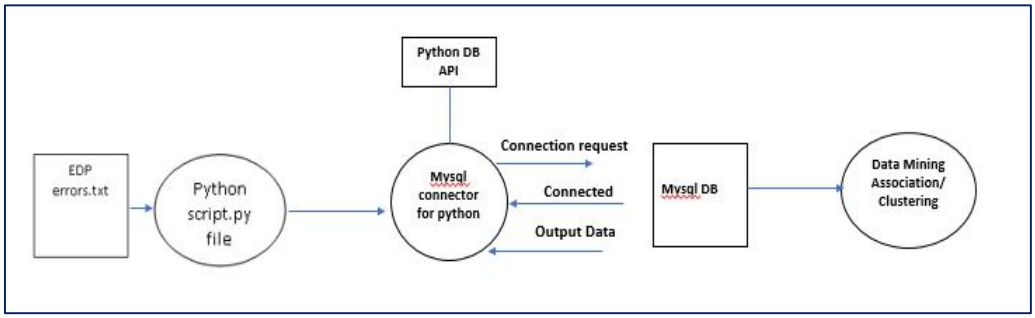


Figure: 3.0 Proposed model.

1) *Python DB API:* It consists of various API's such as Mysql, Oracle, Ms-Access, Ms-Sql Server etc., to be connected using the python scripts to the above-mentioned data bases.
2) *Python Script:* It is .py program which impose the specific API to connect to the target DB servers for example MySQL, MySql server, Oracle DB server, MS-Accesses etc. In this paper we are using python program to connect MySQL DB to establish connection with python DB API.
3) *MySQL – Python Connector:* It is the interface between the python script and MySQL DB, in the low level. It will convert EDPerrors.txt into the one or more table of MySQL DB based on the requirements of the system administrator.
4) *MySQL DB:* It is open source DB server which contains the various tables, table spaces to store and organize, manipulate, to summarize the various categories of data, in this paper we are using MySQL DB to store EDP errors in terms of one or more DB tables.
5) *Data Mining Process:* This process will be initiated once the EDP errors are successfully stored in MySQL DB this mechanism will be described in future work.
6) *Connectivity Models:* ODBC, JDBC, RDBMS, ADO

Sample Script

```
#!/usr/bin/python; import MySQLdb
***Database connection
        db=MySQLdb.connect("servername","user_name","password","error_db")
***prepare a cursor object using cursor() method
        cursor = db.cursor()
        executive sql query
        cursor.execute("SELECT VERSION()")
***Fetching row
        data = cursor.fetchone()
        print "Database version : %s " % data
***table creation
        sql = """CREATE TABLE errordb( CHAR(20) NOT NULL,CHAR(20), INT,  CHAR(1), FLOAT )"""
        cursor.execute(sql);
***disconnect from server;
        db.close()
```

A. Implementation

The proposed paper is extension of EDP module [14,15]. The data base connectivity module in the proposed paper is implemented in LAMP (Linux Apache MySQL Python) servers. The error data will be imported from EDP second module [15] to MySQL server data base using above connectivity model shown in the figure 3.0. In MySQL server a DB is created with a name as (test.sql) and the names of the tables are error_m_t, errorfile_m_t, errorfile_m_tmp, errormaster_m_t, errormaster_h_t, errortype_m_t, status_m_t. The Python script will connect to the DB and data will be imported from error.txt file. An automated SQL query will now export the data to the errorfiel_m_t table. All the errors will be stored as per the priority generated by a SQL query. Errors will be prioritized/indexed based on repetition of errors and leads to a issue. The most important table in this module is errorfile_m_t, where

©IJRASET: All Rights are Reserved

each error will be stored using an error number(errId) such as 1,2,3,…n. If error id and errName collectively is 1 then the error is in pending status otherwise if it is 2 then the error is resolved, and same thing will be updated in status_m_table.

If we have same errId with multiple errors, then all the errors will be updated in a notepad file by running a SQL query. This paper focuses only on data base connectivity using Python script and importing the errors from error.txt to MySQL data base (test). The errors DB are applied with different data mining techniques for further fore casting.

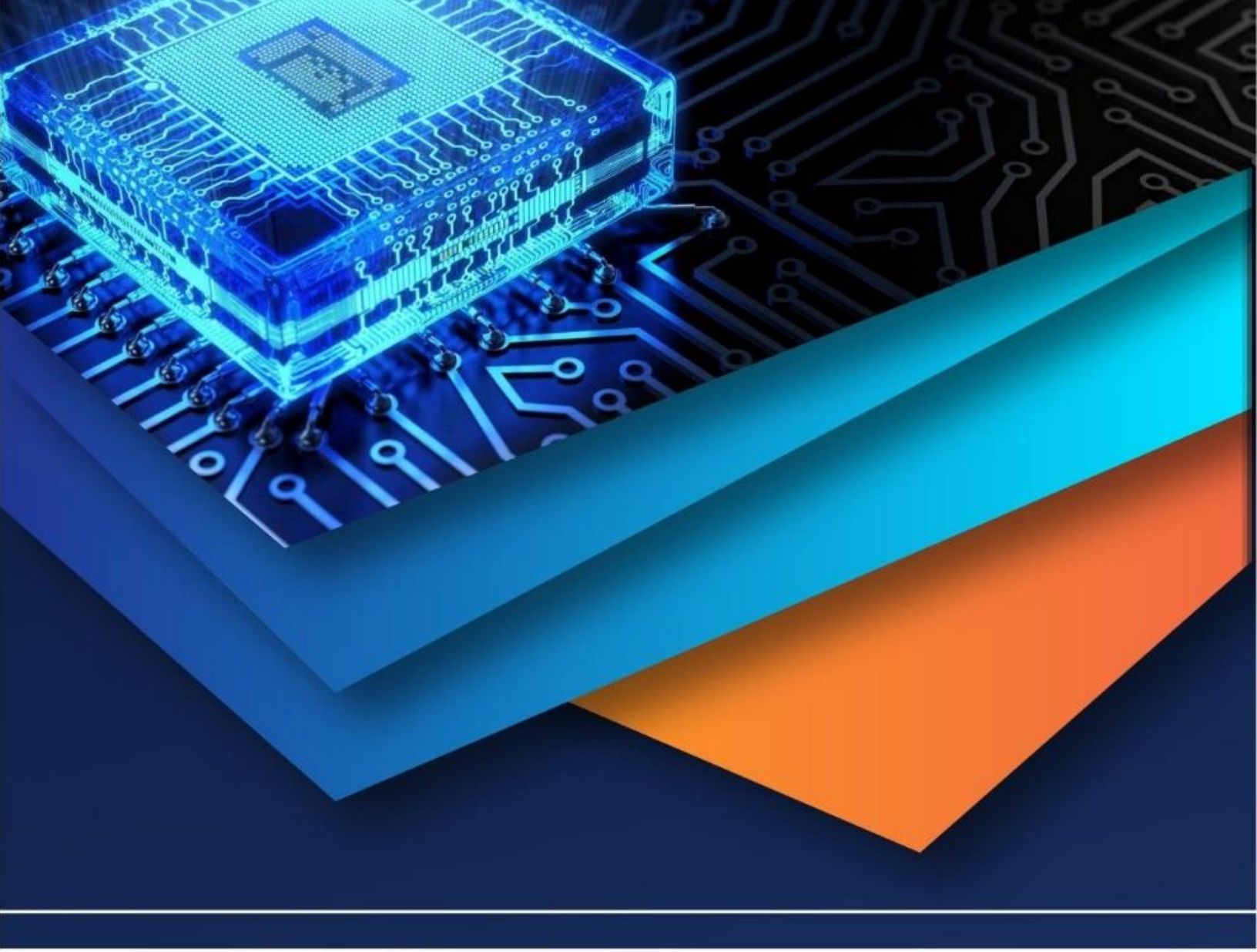
IV. CONCLUSIONS

The present paper focus on database connectivity between errors.txt file and MySQL database server. After implementation, the proposed method able to import data from error.txt file and export to database server in errorfile_m_t table. Further resolved and unresolved errors will be sent to datamining techniques for further fore casting.

REFERENCES

[1] Nikit Jain, Vishal srinivas “Data Mining techniques: a survey paper” ijret : international journal of research in engineering and technology, volume : 02 issue : 11 NOV – 2013
[2] Dr. M.H Dunham, “Data Mining, Introductory and Advanced Topics”, Prentic Hall, 2002.
[3] R. Kabilan, Dr. N. Jayaveeean, “Data Mining in Cloud Computing techniques: a survey paper” IJSEAS: International Journal of Scientific Engineering and Applied Science, Issue-8 NOV – 2015
[4] Youssef M. Essa, Bigdata Consultant, EMC, Cairo, Egypt, "A Survey of Cloud Computing Fault Tolerance: Techniques and Implementation" International Journal of Computer Applications (0975 – 8887) Volume 138 – No.13, March 2016
[5] Juna Li Pallvi Roy Samee u: Khan Lizhe Wang Yan Bhi” Data Mining Using Cloud; An Experimental Implementation of Apriori over mapreduce”
[6] Virendra Singh Kushwah1, Sandip Kumar Goyal2 & Priusha Narwariya3 "A Survey On Various Fault Tolerant Approaches For Cloud Environment During Load Balancing" (IJCNWMC) ISSN(P): 2250-1568; ISSN(E): 2278-9448 Vol. 4, Issue 6, Dec 2014, 25-34
[7] Mehdi Effatparva, Seyedeh Solmaz Madani, "Evaluation of Fault Tolerance in Cloud Computing using Colored Petri Nets " IJACSA, Vol. 7, No. 7, 2016.
[8] A.Mahendiran, N. Sarvanan N. Venkata Subramanaian, N Sriram Implementation of K-Means clustering in cloud computing environment research journals of applied science, engineering and technology 4(10) ISSN 2040-7467.
[9] Iqjot Singh, Prerna Dwivedi, Taru Gupta and P. G. Shynu,"Enhanced K-means clustering with encryption on cloud", IOP Conf. Series: Materials Science and Engineering 263 (2017) 042057 doi:10.1088/1757-899X/263/4/042057
[10] Mr. Sudhir M. Gorade1, Prof. Ankit Deo2, Prof. Preetesh Purohit," A Study of Some Data Mining Classification Techniques", IRJET e-ISSN: 2395 -0056 p-ISSN: 2395-0072 Volume: 04 Issue: 04 | value: 5.181 | ISO 9001:2008
[11] Renu Asnani "A distributed k-mean clustering algorithm for cloud data mining" International Journal of Engineering Trends and Technology (IJETT) – Volume 30 Number 7 - December 2015"
[12] Krishan Rohilla1, Shabnam Kumari2, Reema3,"Data Mining based on Hashing Technique", International Journal of Trend in Scientific Research and Development, Volume 1(4), ISSN: 2456-6470"
[13] B. Anand Kumar, A.Vinayababu, V. Malliah, K. Madhukar, " IJRASET, ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue XII, Dec 2018".
[14] B. Anand Kumar, K. Madhukar, V. Malliah, Vinay Babu, " (IJRASET)" ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887, Volume 7 Issue I, Jan 2019

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)