iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** III **Month of publication:** March 2018

DOI: http://doi.org/10.22214/ijraset.2018.3361

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: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

A Proposal for Duplicate Data Detection in Big Data

Nancy Jasmine Goldena[1]
[1]*Department of Computer Applications, Sarah Tucker College, Tirunelveli, India*

***Abstract:** Big Data is now the most talked about research subject. Over the years with the internet and storage space expansions vast swaths of data are available for would be searcher. But the problem that plagues the internet storage space is that multiple copies of the same data exits. This not only degrades the search results but also concedes time. Also it prevents accurate data analysis. In order to solve these problems a novel proposal has been proposed here. Traditional data mining approaches work well with dataset of small sizes. When the size of the dataset grows newer techniques are needed as it would consume more time to implement an operation on the large Big dataset. Hence a simpler approach is being proposed here, which does need a creation of new technique to process Big Data but proposes a unique strategy so that we can make use of existing data mining techniques efficiently.*
***Keywords:** Big Data, Data Cleaning, Fuzzy rule based approach.*

I. INTRODUCTION[1]

Data has now become the buzz word in technology as it can do a whole array of things which we previously didn't know about. Data has the potential, not only to predict but also to fine tune existing practices. Data about an application or firm can help make it better, by enabling better services and thereby enriched profits. The amount of data also plays a role. The more data you have, the better fine tuned your predications can be. Take for example the problem of detecting weather. Data stored from previous years form patterns and help in prediction. If the data points are more, then it is easier to connect the points rather than trying to connect too far of data points. More data eventually gives us a clustered center of data points which in turn helps prediction more accurate.

This concept of more data has eventually created a field around it called Big Data. Big Data is generally an extension of data mining with the catch being that Data size is exceedingly huge. Data here are measured in petabytes[5]. The proper definition of Big Data is described best by the 5V's (Figure 1) which interesting evolved from three. They are Volume, Velocity, Variety, Veracity and Value. The last two have being added by IBM and Oracle lately. Volume describes the huge amounts of data in the field of Big Data. The amount of data being dumped into the internet is relatively huge beyond comprehension. Variety pertains to the different types of data contributed by various media like text, images, video and audio. Velocity is the speed at which data is normally generated. Variability is the variation in the type of data. For eg:- an email might have data which would be diverse whereas a photo website like tumblr would have only image data. Value is the outcome of processing or analyzing these huge volumes of data.

These best describe what Big Data really is[1]. The term Big Data was used post 2008[2] and is now relatively used in many fields containing data like government, business enterprises, medical, law enforcement agencies, cyber security etc [3].

During the 90's the only common device that could access the internet was the computer. Nowadays a number of devices can access the internet. They range from mobile phones to household items like refrigerators, wash machines etc, to transport vehicles like train, cars etc, to name a few. These abounding numbers of devices that can be hooked up to the internet are called the internet of things (IOT) and they are responsible for populating the internet databases with huge amounts of data. These include items embedded with electronic and hooked to the internet like refrigerators, television etc to embedded sensors in automobiles, actuators etc., An estimation of things that have joined the internet space over the years is shown in figure 2[11]. Figure 3 shows a potential roadmap of what the internet of things might be able to do in the distant future.[12] The x axis of both the graphs shown in figure 2 and 3 shows the year which is time and you can see the graph raising up slowing showing the extent of internet usage over time through varied devices.

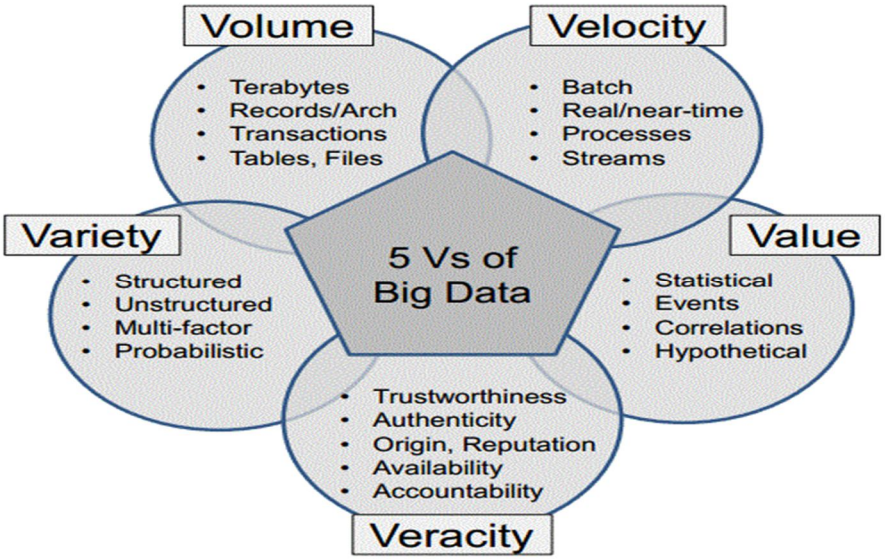


Figure 1 Big Data

Since there are many sources contributing data, chances are that sometimes multiple copies of the data may be present in multiple or same data storage points across the internet[4]. These are called duplicate data. Some duplicate data are intentional like result publishing websites. They incorporate the strategy to balance the load of the server. But some are not intentional. Like for example take a encyclopedia site like Wikipedia. Duplicate data can wreak havoc here. Apart from the obvious problems posted by duplicate data in the internet, it can alter predictions and patterns when analyzed upon. Many companies now mine information from the internet. So to prevent the above mentioned problems it is crucial to eradicate these. Apart from this the other problems posed by duplicate data are increase space accumulation, increased search times, unwanted or repeated duplicate search results. It also affects the efficiency of the remote host.

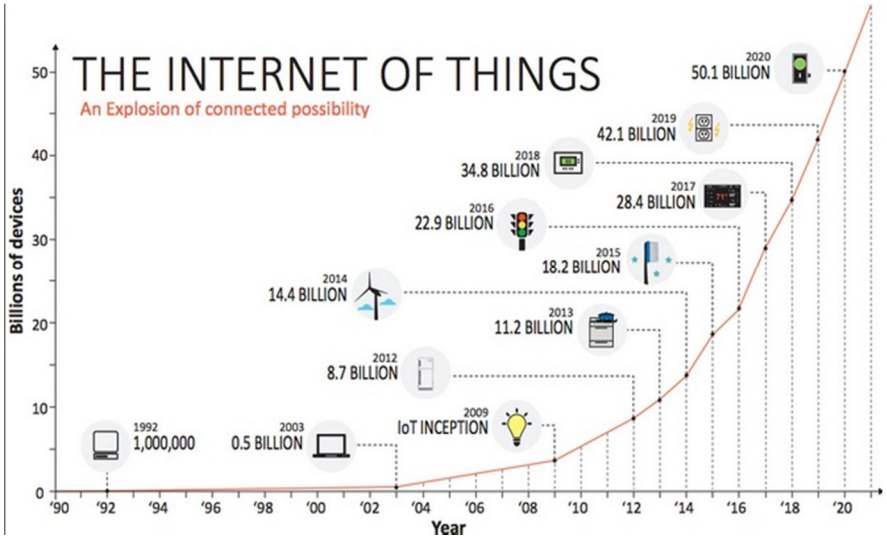


Figure 2 Internet of Things

To solve these problems posed by Big Data a two prone solution is devised here. In my previous work research was conducted to detect the so called duplicate data using various techniques like Outlier Detection, Fuzzy Clustering, and also by combining Fuzzy Clustering with Radon Transformation. All three works had fantastic results. In this work we try to make them to work on Big Data. These approaches on their own won't work with Big Data hence here we propose to segment the data into more manageable segments so that the ultimate goal of detecting duplicate data can be achieved.

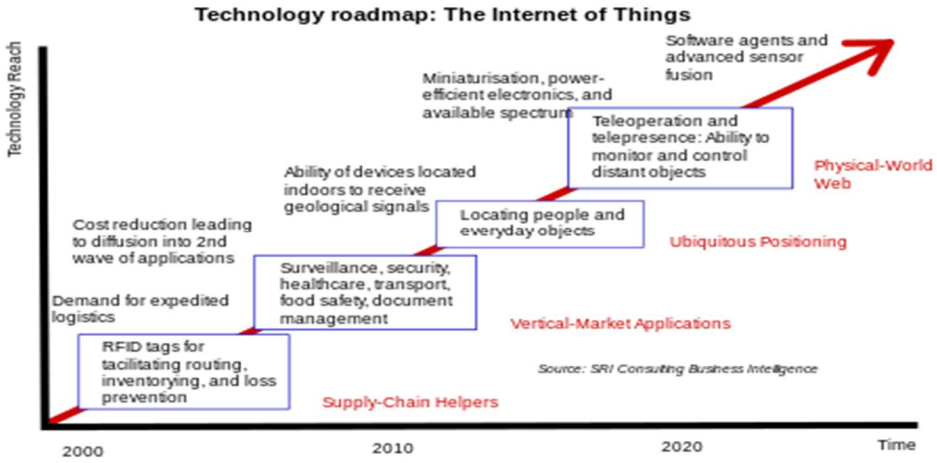


Figure 3 Roadmap of IoT

II. SURVEY OF LITERATURE

As a fact of matter it's no surprise that Big Data means really huge quantum's of data. Although the computer is capable of computing vast amount of data it all boils done to time and the recourses provided for by the system. Here the primary problem for Big Data is of course space. There aren't petabytes capacity hard disks available in the market. So processing such huge volumes of data is impossible to start with. Also the second problem that might arise, suppose that you manage to load your data to process is that traditional known methods can't handle such large data. The computer although it is a multiprocessing and multitasking environment it works in a linear fashion only. Hence it takes more time to compute. To counter this linearity problem there are various methods available in the literature which we shall see one by one.

A. Cluster Analysis

Clustering is a method by which data points present in the database try to be in close proximity to similar data points. When these data points are displayed in a scatter plot the above explanation can be visually seen as shown in figure 4. This figure shows the purchasing power of individuals and you can easily see patterns forming here. Using this we can take in the needed data and eliminate the rest.

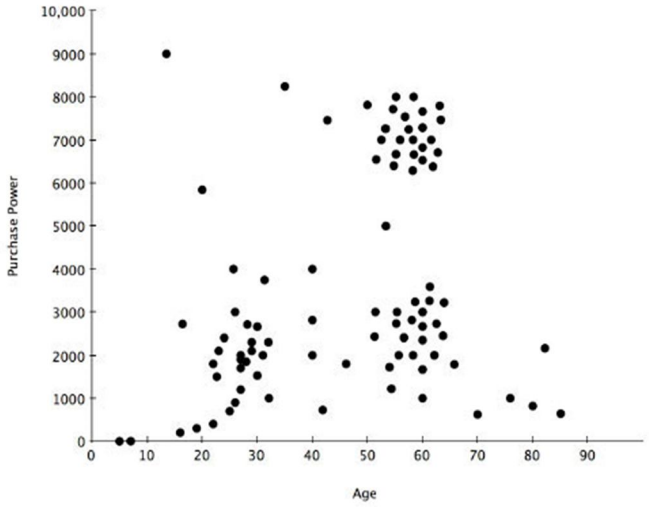


Figure 4 Sample cluster plot

B. Correlation Analysis

It determines the strength between the various groups of data present in the dataset. The definition of relationship function is quite important. Because that is what, will help us segregate the data. [13]

C. Statistical Analysis

This technique is based on a statistical model which enables to group data. Mostly in statistics uncertainty and randomness are modeled using the probability theory. This can be used to summarize datasets [13]

D. Regression Analysis

It is also a mathematical technique that reveals correlations between a few variables. It is based on observations and experiments. [13]

So if we see from the above class of data analysis techniques all reply on splitting the data into manageable bits. Here in this work we intend to do the same but with a slight twist, the details of which are explained in the next section.

III. PROPOSED APPROACH

Here the chosen problem is duplicate data detection. Although we have successfully done this in the previous work here we are dealing with Big Data or in other words huge amounts of data. The plan here is a simple two pronged approach as illustrated in the algorithm below.

Step 1. Get in the dataset set to be weeded of duplicate data
Step 2. Segment the data into small manageable chunks of pieces
Step 3. Implement fuzzy technique to find duplicate data and thereby remove it.
Step 4. Then unify the dataset by joining the segments and thereby remove any duplicate data that may be found in two segments by using step 2
Step 5. Continue step 4 until total segment is 1

A. Segmentation

Segmentation in general is splitting a large entity into smaller pieces. Segmentation is used in various fields where in the primary objective is to convert the given data to be processed into small manageable parts. Doing so yields benefits like better computational efficiency to better zero in on the needed data. To divide the data into small chunks a constraint like a threshold value is needed. This value is purely dependent of the data set taken into consideration. Here to divide the dataset into smaller pieces of manageable data we have used the size as the constraint. The reason for deciding on size at the constraint is that we are focusing on eradicating duplicate data. In such a set we cannot eliminate certain segments. Hence the whole dataset has to be processed. So therefore the size is taken as the constraint. The following equation is used for segmentation.

$$D\_S(i) = B\_D / s$$

Where D_S is the data segment size of one segment where i is one of the n segments, B_D is the Big Data and s is the segment size. The reason for doing this is the traditional approaches of data mining can be used on the large dataset. It is more or like divide and conquer strategy. The segment size s is chosen in consideration with how many segments you need. It is determined by the processing power of the system being used. Suppose for example the dataset that you have is of size B_D and processing time of $D\_S(i)$ is estimated to be t. Then size of D_S should be decided by t. If t registers a higher value then the value of $D\_S(i)$ should be less and vice versa.

B. Fuzzy C-means Approach

Fuzzy C-means approach works by way wherein you put data into one of two clusters based on a criteria[8]. Here the fuzzy partition is performed using the following algorithm.

1) First choose the primary prototypes(V_i)

2) Compute the degree of membership for the data present in the clusters using the following equation $\mu_{x,i} = \frac{\left(\frac{1}{d^2(Z_x,V_i)}\right)^{(1/m-1)}}{\sum_{i=1}^{C}\left(\frac{1}{d^2(Z_x,V_i)}\right)^{(1/m-1)}}$

3) Compute the new centroids $V_i^t = \frac{\sum_{x=1}^{N}\mu_{x,i}^m Zx}{\sum_{x=1}^{N}\mu_{x,i}^m}$

4) Update the degree of membership $\mu_{x,i}$

5) If $\mu_{x,i} < \forall$ stop, else go to step 3

Where $\forall$ is a termination criterion between 0 and 1, $d^2(Z_x,V_i) = \|Z_x - V_i\|^2$

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

First the entire dataset is taken and the ones that need to be in a said set should be determined. This is done by computing the membership. What data should be present in a particular cluster is given in the equation given in step 2 which is the one that calculates the membership. Then the centroid is computed. Based on the centroid the values that need to be in a particular set are determined. These steps are repeated until all the data are put into various segments. In doing this the duplicate data can be easy detected. As the membership output would be same for two similar data. Therefore by doing this we are also able to weed out the data [6][7]. This step is essentially an extension of our previous work. It has been literally fine tuned to work with Big Data by just segmenting the large dataset into small manageable pieces.

C. Data set Unification

The data set segments are now weeded out of duplicate data. But however there is a possibility that two segments might contain the same duplicate data. To weed this duplicate data out the same fuzzy approach is implemented after a union of two nearby data segments. The formula of which is given belo $\cup D\_S = D\_S(i) \cup D\_S(i+1)$
where $D\_S(i)$ and $D\_S(i+1)$ are neighboring segments. We join two segments at a time and the result of each unified data segment is subjected to the fuzzy approach discussed in the previous step and the unification process continues until the numbers of segments become one. The whole process of the work is shown diagrammatically in figure 4. It started with segmenting the dataset into segments, which then are weeded out of duplicate data using the fuzzy c-means and are then unified through various levels until the number of segment becomes 1. And while unifying each pair of segments the fuzzy c-means is again implemented on the newly joint segment to weed out any duplicate data.

IV. EXPERIMENTAL RESULTS

To test the hypothesis a Wikipedia data set was used to test it[9]. A screen shot of it is shown in figure 5. All these files contains link to data. The reason for choosing Wikipedia data is that it too faces the same problem as being dealt in this paper. The link given in [10] is Wikipedia way of determining duplicate documents. Hence we have sourced data from that site. The running of the problem dataset was disastrous as the system crashed but it was successful when executed with the segmentation approach. The execution rate is graphically depicted in Figure 6 Hence the technique with which we have devised was able to run and achieve results successfully. The segmentation part was quite influential to do the last part of the task.

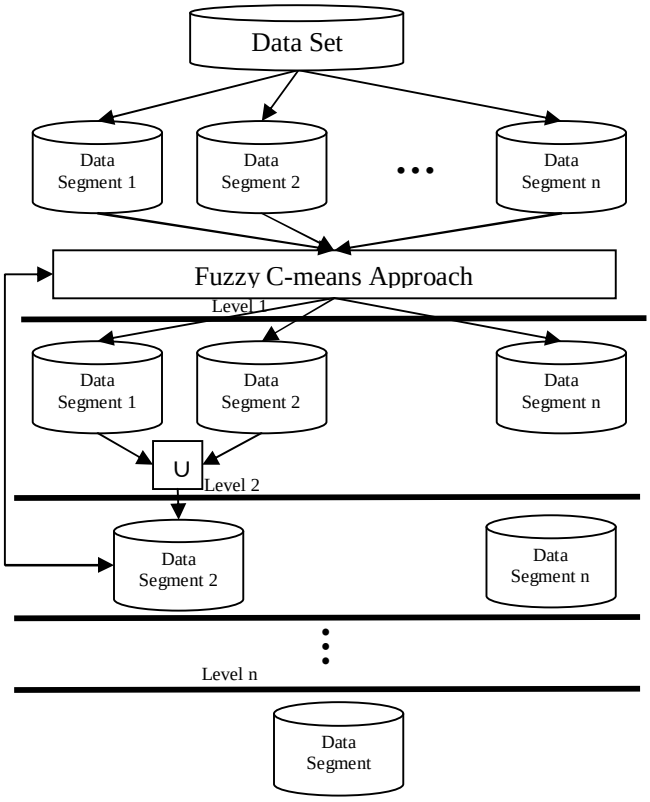


Figure 4 Block Diagram of the proposed approac

V. CONCLUSION

The main problem with Big Data is that large swaths of data are nearly impossible to search and with the addition of duplicate data it only makes matters worse. Hence segmenting the data into manageable pieces can help balance the load and lead to a more robust of way of effective analysis. Further the technique of using fuzzy c-means in each of the data segments helps to ensure the weeding out of unwanted data that are copies hence thereby refining the data analysis performed. As shown in the experimental analysis the execution rate of the proposed technique is always high no matter what hardware environment you possess. If you have a lower configuration then just increase the number of segments otherwise decrease the number of segments. The execution rate is inversely proportional to the number of segments. In future, methods could be incorporated to determine the correct number of segments in proportion to the hardware configuration.

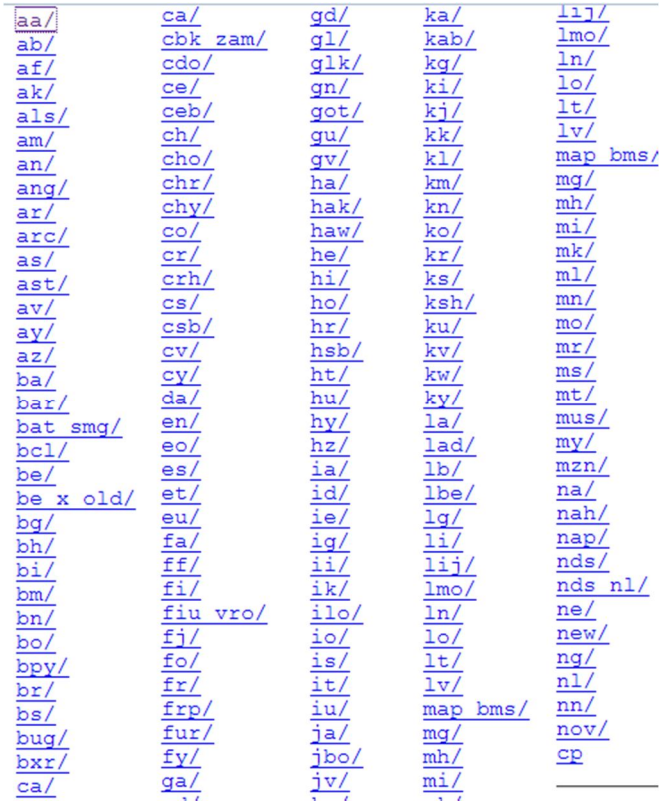


Figure 5 Wiki dataset

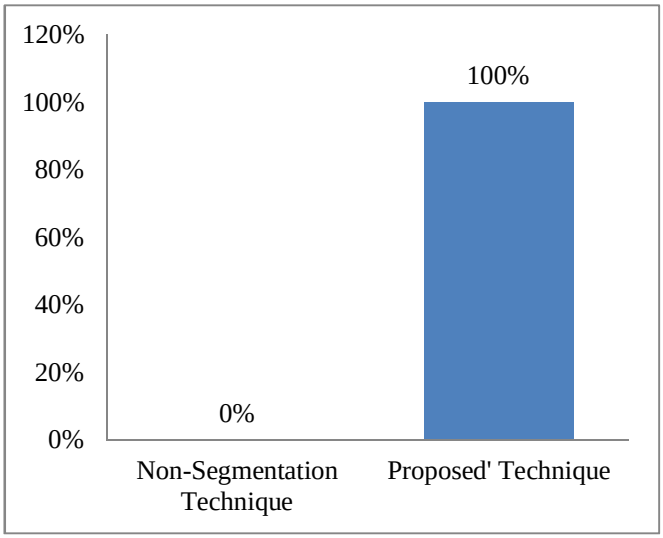
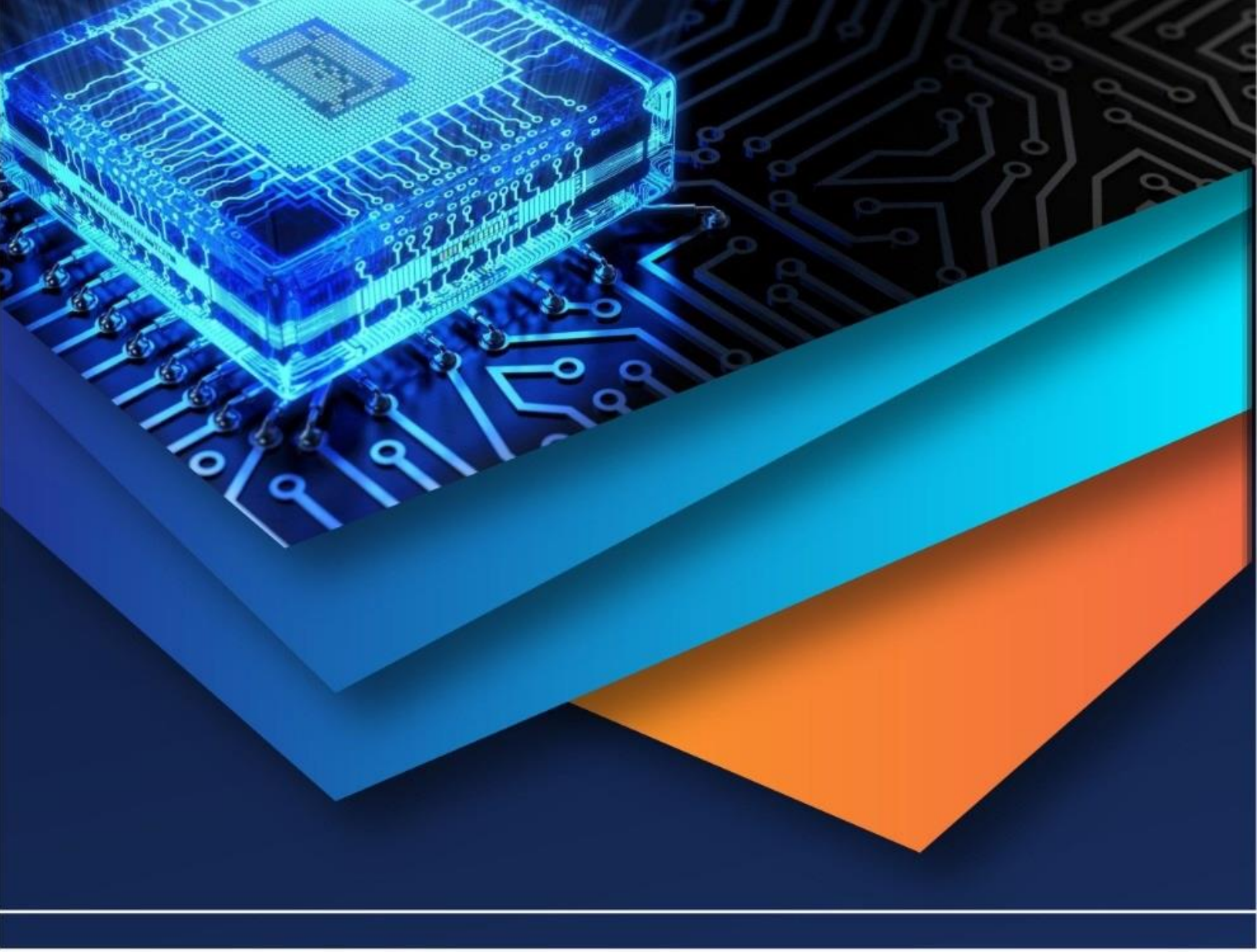


Figure 6 Execution rate

REFERENCES

[1] Baaziz, and L. Quoniam, "How to use Big Data technologies to optimize operations in Upstream Petroleum Industry,"International Journal of Innovation, vol. 1, no. 1, pp. 19-29, 2013.
[2] Editorial: "Community cleverness required," Nature, Vol. 455, No. 7209, pp. 1-1, 4 September 2008. http://www.nature.com/news/specials/bigdata/index.html
[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.
[4] Vermesan, Ovidiu; Friess, Peter (2013). Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems (PDF). Aalborg, Denmark: River Publishers. ISBN 978-87-92982-96-4
[5] Everts, Sarah (2016). "Information Overload". Distillations. **2** (2): 26–33. Retrieved 17 February 2017.
[6] Nancy Jasmine Goldena, and Dr. S.P. Victor, "Entropy-based Fuzzy Clustering Approach with Radon Transformation on Duplicate Data Detection and Identification." International Journal of Advanced Research in Computer Science and Software Engineering 4.8 (2014): 1021-028.
[7] Dr. S.P. Victor and Nancy Jasmine Goldena, "Sophisticated Fuzzy Clustering Algorithm for Duplicate content Detection based on Outlier Detection." International Journal of Advanced Research in Computer Science 5.6 (2014).
[8] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57
[9] https://dumps.wikimedia.org/ accessed on May 2017
[10] http://tools.wmflabs.org/dupdet/ accessed on May 2017
[11] http://plug.hani.co.kr/futures/2772091 accessed on May 2017
[12] https://en.wikipedia.org/wiki/File:Internet_of_Things.svg accessed on May 2017
[13] Khan, Nawsher, et al. "Big data: survey, technologies, opportunities, and challenges." The Scientific World Journal 2014 (2014).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY