ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

Analysis of Various Clustering Algorithm on Job Events Data of Google Cloud Tracelog

Mrs. Chethana C

Assistant Professor, Dept. of CSE, BMSIT&M

Abstract: The term Workload is defined as "the amount of work assigned to or done by a client, workgroup, server or system in a given time period" and consists of two components user and task. Analysis of cloud jobs and user benefits both providers and researchers as it enables a more in-depth understanding of the system. In this paper I am using clustering techniques available in WEKA tool to perform analysis. The data is taken from google cloud trace.

Keywords: Analysis, Cloud Computing, Cluster, Workload

I. INTRODUCTION

In computing, the workload is the amount of processing that the computer has been given to do at a given time. The workload consists of some amount of application programming running in the computer and usually some number of users connected to and interacting with the computer's applications [1].

The various types of jobs submitted by the users will arrive at cloud data centre. Every job includes certain self-defining attributes such as the submission time, user identity and resource requirements in terms of CPU, memory and disk space [6]. A tool is used to predict and plan future work and skills requirements based upon historical data. Once a workload baseline has been established using past performance adjustments are made for expected changes in demand or other factors which impact the project scope. Workloads by themselves may have properties or attributes that could dictate where workload can or can't run. This justifies existence of a workload as a separate entity - it is in theory possible to construct a workload for which no deployment can exist in any of the clouds available today.

There are many examples what kind of attributes a workload may possess. A workload may have a compliance attribute, which says that this workload must run in an environment with a certain certification. Another attribute may be a geo location requirement, whereas it must run within a certain geographic region for a legal reason [5] A workload may be time-bound ("runs for 5 hours") or time-unbound. A workload may have a specific start time or flexible start time, in which case it may have a hard stop time (for example, must finish by a certain time in the future). It can be interruptible or must run without interruptions.

Workload may have a budget associated with it, it may have redundancy requirements. It may require a certain OS or distribution. It may require certain. It may require a certain minimal access speed to some data source Each requirement is a restriction - the more requirements a workload has, the fewer clouds can potentially run it [4].

II. CLUSTERING

Clustering [2] is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. In this paper I have considered four types of clustering algorithm for the analysis purpose. They are Canopy, K-mean clustering, Farthest first clusters and Density Based clustering.

III. METHODOLOGY

The analysis is conducted using the data from the second version of the Google Cloud trace log. The data is downloaded using Google Cloud SDK Shell. I have considered only 70000 instances of the data. The tracelog consists of various tables as follows [3].

- Machine events are described by two tables which are machine events table and machine attributes table.
- Job and task events are described by the job events table, Task events table and task constraints table.
- task usage table

The present article focuses on job_events table which contains 500 CSV files.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

The job events table contains the following fields:

1. timestamp
2. missing info
3. job ID
4. event type
5. user name
6. scheduling class
7. job name
8. logical job name

The clustering algorithms in weka tools is applied on the fields 1 ,3, 4, and 6

IV. PERFORMING CLUSTERING IN WEKA

For performing cluster analysis the data is loaded into WEKA in .ARFF format.

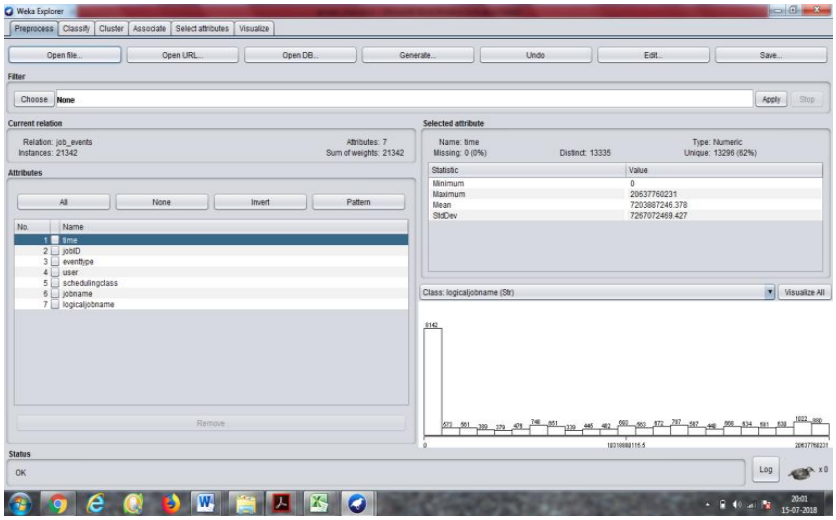

Figure1: Load dataset into WEKA Tool.

The various clustering [2] algorithms available in WEKA is displayed by clicking on the cluster button on menu bar and analysis is done by choosing algorithm one at a time.

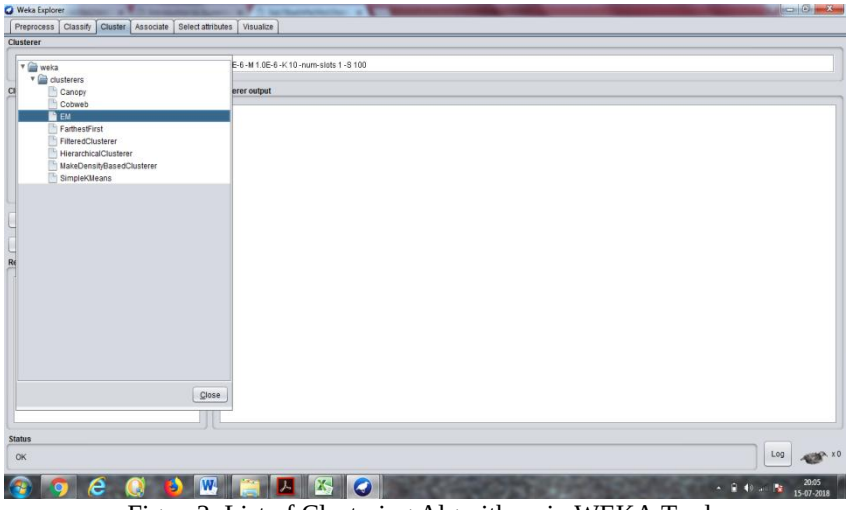

Figure2: List of Clustering Algorithms in WEKA Tool.

Canopy: Canopy based clustering algorithm uses two steps, initially the datasets are partitioned into overlapping subsets called canopies and then clustering process has been performed on subsets. It consumes less time to provide result.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

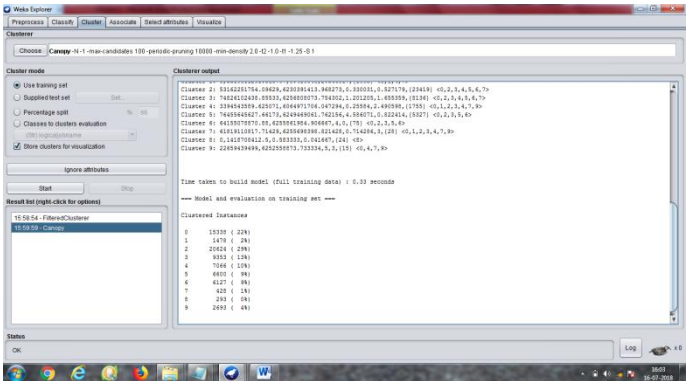

Figure3: Canopy Clustering Algorithm

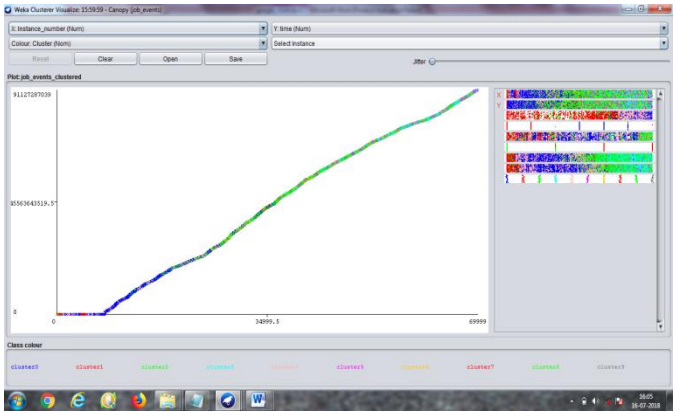

Figure4: Result of Canopy Clustering in form of graph

Farthest First: Farthest first algorithm has same procedure as kmeans, this also chooses centroids and assigns the objects in cluster but with max distance and initial seeds is value which is at largest distance to the mean of values. The cluster assignment is different, at initial cluster.

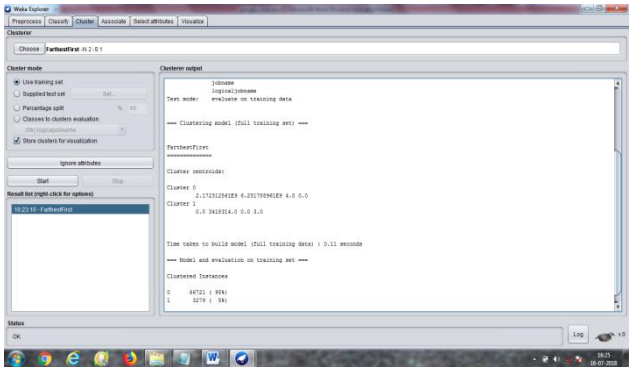

Figure5: Farthest First Clustering

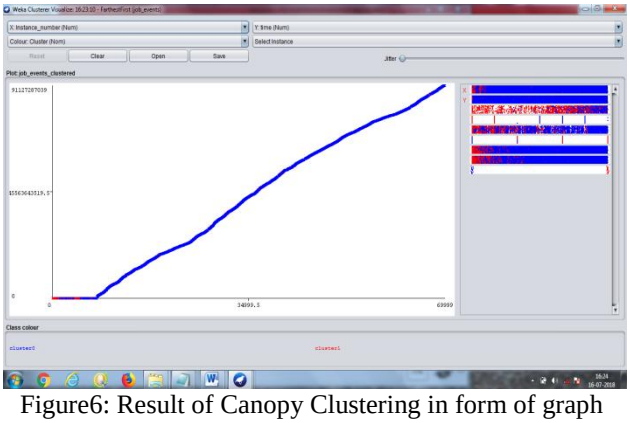

Figure6: Result of Canopy Clustering in form of graph

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

SIMPLE K-Means Clustering: K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

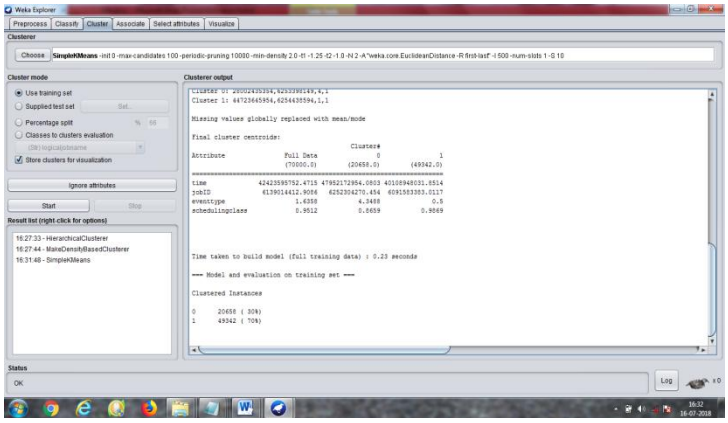

Figure7: Simple K-means Clustering

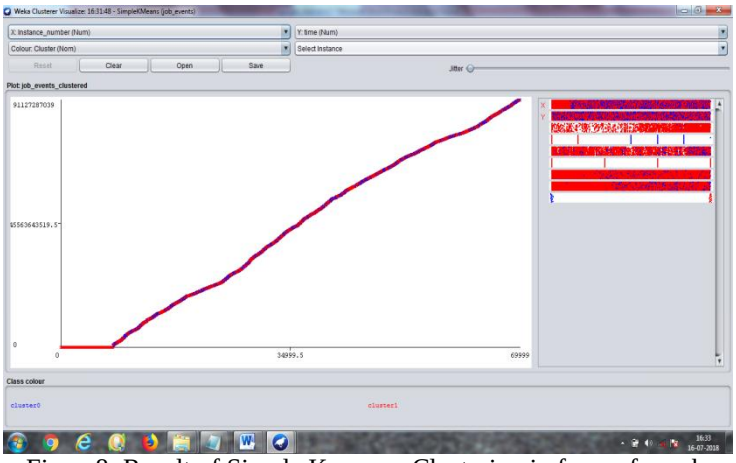

Figure8: Result of Simple K-means Clustering in form of graph

Density Based Cluster: DBSCAN (for density-based spatial clustering of applications with noise) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. OPTICS can be seen as a generalization of DBSCAN to multiple ranges, effectively replacing the parameter with a maximum search radius.

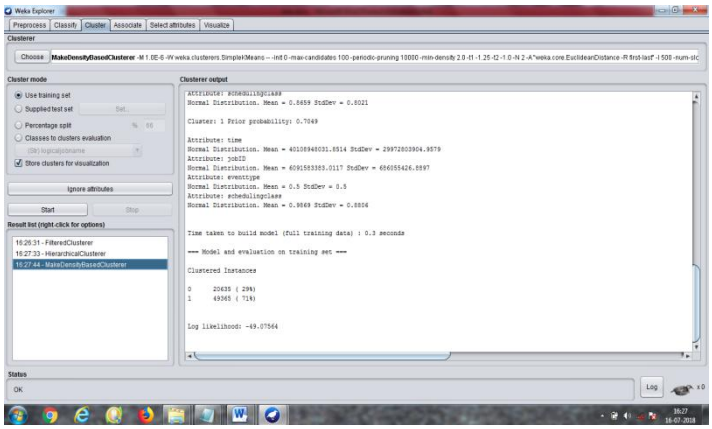

Figure9: Density Based Clustering

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

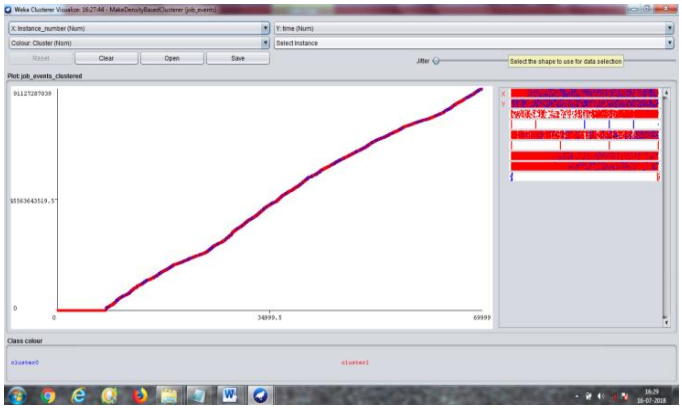

Figure10: Result of Density Based Clustering in form of graph

V. COMPARISON OF RESULT OF ALGORITHMS

SL. NO	Name of Cluster	Instances	No.of cluster selected by cross Validation	Log Likely hood	Clustered Instances in terms of percentage										Time to build Model In Seconds
					0	1	2	3	4	5	6	7	8	9	
1	Canopy	70000	10	--	22	2	29	13	10	9	9	1	0	4	0.2
2	Farthest First	70000	2	--	95	5	-	-	-	-	-	-	-	-	0.06
3	Density basedclusterer	70000	2	-49.07564	29	71	-	-	-	-	-	-	-	-	0.3
4	SimpleKmeans	70000	2	--	30	70	-	-	-	-	-	-	-	-	0.18

VI. CONCLUSION

This paper gives the analysis few clustering algorithm of WEKA tool. The data is taken from the job events table of the second version Google Cloud tracelog. 70000 instances of the data were taken for analysis. Among four clustering algorithm simple k-means clustering algorithm is taking less time for clustering.

REFERENCES

[1]. M.Alam, K.A.Shakil, S.Sethi, "Analysis and Clustering of Workload in Google Cluster Trace based on Resource Usage", 2016 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), DOI: 10.1109/CSE-EUC-DCABES.2016.271,August 2016.
[2]. A.Verma,L.Pedrosa, M.Korupolu,,"Large-scale cluster management at Google with Borg", Google Inc,ACM,2015.
[3]. C. Reiss, J. Wilkes, and J. Hellerstein, "Google Cluster-Usage Traces: Format þ Schema," Google Inc., Mountain View, CA, USA, White Paper, 2011.
[4]. N Sharma, A Bhajpai, R Litoriya, "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2,2012.
[5]. G. D. Costa,L. Grange, I. de Courchelle, "Modeling, Classifying and Generating large-scale Google-like Workload", preprint submitted to Sustainable Computing: Informatics and Systems (SUSCOM) January 8, 2018.
[6]. J. Panneerselvam, J. Hardy, N. Antonopoulos, Analysis, Modelling and Characterisation of Zombie Servers in Large-Scale Cloud Data centres