

# Compare Process of some clustering algorithms of data Mining.

[Jamal Mbarki] [Sidi Yasser El Jasouli ]

**Abstract**— The aim of this paper is to create a process that compare time response of some clustering algorithms in data Mining , the subset of this experimentation is a data set stored in Teradata . In this paper, we present a novel hierarchical clustering algorithm called CHAMELEON that measures the similarity of two clusters based on a dynamic model. We present also k Means as representative of Partitioned ones, and DBSCAN as part of Density Based.

**Keywords**—component; data mining; Clustering ; response time ;

## I. Introduction

Data mining is a multi-disciplinary field which uses the three main scientific components: statistics, machine learning, artificial intelligence and database technology. It's commonly known by its acronym KDD: Knowledge Discovery in Database, and refers to all methods and algorithms used for data exploration or prediction in large data bases volumes, Data mining is very important in various fields such as sciences, business and other areas deal with a large data set (J. Mbarki. 2014 ).

## II. Data warehouse

It's a large reservoir of detailed and summary data that describes the organization and its activities, repartitioned into a various business dimensions. The customer dimension remains the most challenging dimension within DWH (J.Mbarki at all, 2014), DWH is a collection of what is commonly known data mart "(Inmon, 1996), data mart is subject oriented , each topic is stored separately.

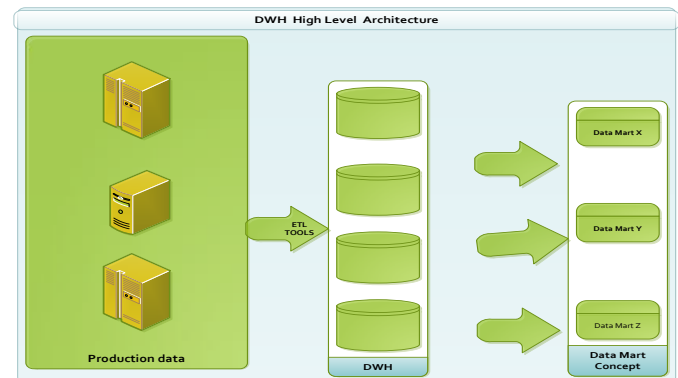
### Jamal Mbarki

Laboratory of Computer Science Research (LARI), Faculty of Sciences, University Mohamed Ier, Oujda, Morocco

### Sidi Yasser El Jasouli

Integrated & Efficient solutions IT.sprl  
Belgium

Example: CDB: Customer Data Base contains typically information on customers, while Billing contains information on invoices. The Data mart enables the business user to report and provides decision support in organised manner. data mart is built to allow a quick and easy use of data.



## III. Clustering

The learning problems may be classified into 2 main components ; either supervised or unsupervised. In supervised learning, the objective is to predict the value of a variable called predictor based on a number of input variable called most of the time descriptors; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures.

Clustering is one the most important task of unsupervised learning; a cluster is a subset of data which are similar. Clustering (also called also segmentation ) is the process of dividing a dataset into groups, where the members of each group are as much similar (close) as possible to each other. However, groups are as dissimilar (far) as possible from the each other. Clustering can uncover previously undetected relationships in a dataset. There are many applications for cluster analysis. For example, in business, cluster analysis can be used to discover and characterize customer segments for marketing purposes.

The Launch of new product/service uses the customer segmentation. There are other fields where these technics are used such scoring in Finance, or in biology, for classification of plants and animals.



## B. Data choice

We create a table in Teradata representing

The data set for our experimentation

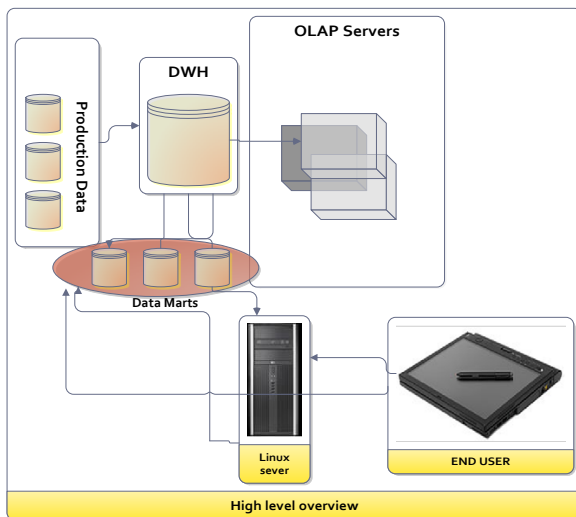
Where the attribute A5 is taken as descriptor .

The cluster\_number is set to default value =0(before run , after the run of the process it's taking the value of the cluster ....

Below described the table creation statement:

```
CREATE SET TABLE Data_mart01.DM_fact_cluster01
,NO FALLBACK , NO BEFORE JOURNAL,
NO AFTER JOURNAL, CHECKSUM = DEFAULT,
DEFAULT MERGEBLOCKRATIO
(A1 VARCHAR(3) CHARACTER SET LATIN NOT CASESPECIFIC,
A2 VARCHAR(30) CHARACTER SET LATIN NOT CASESPECIFIC,
A3 VARCHAR(15) CHARACTER SET LATIN NOT CASESPECIFIC,
A4 VARCHAR(60) CHARACTER SET LATIN NOT CASESPECIFIC,
A5 integer,
Cluster_number integer)
PRIMARY INDEX (A1 ,A2 ,A3 ,A4 , Cluster_number);
```

## C. Process choice.



## D. Steps for Creating and Execution Clustering Algorithms:

**Step-1** Computational step: Create the shell script in the LINUX server *cluster.sh* according to clustering algorithm principle,

Where principle is according respectively: to

## CHAMELEON, k Means, DBSCAN

**Step-2** Create -a set of data representing respectively the input of each process.

**Step-3** Execute Fast load tool in order to import data for treatment.

**Step-4** Execute the *sh* script and load data into tables.

**Step-5** Go to Teradata playground and select table to view its content.

## E. Deployment & evaluation.

### E. 1 CHAMELEON

Two-phase approach will be used:

#### Phase-I

Use a graph partitioning algorithm to divide the data set into a set of individual clusters.

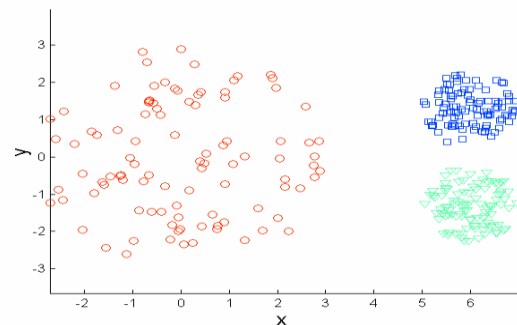
#### Phase-II

Use an agglomerative hierarchical mining algorithm to merge the clusters

### E. 2 Kmeans

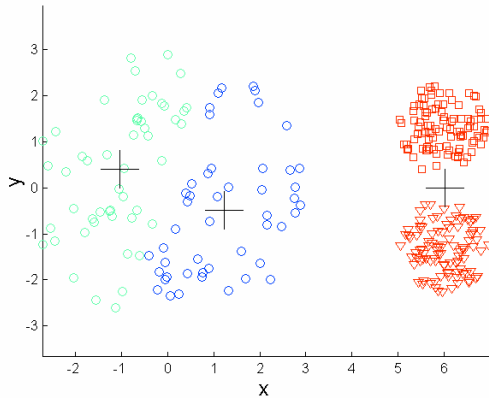
1. Divide set of objects into k not null;
2. Calculate centroid for each partition/cluster;
3. Assign each object to the nearest centroid /cluster;
4. Loop and go to 2, stop when all clusters are stables.

Original points



K-means (3 clusters)

✓ K means : k=3 → k is specified in advance



```
select distinct Cluster_number from
```

```
Data_mart01.DM_fact_cluster01
```

```
Cluster_number
```

```
1 1
```

```
2 2
```

```
3 3
```

✓ DBSCAN

⇒ Number of partition returned = 4

```
select max( Cluster_number) as partition from
```

```
Data_mart01.DM_fact_cluster01
```

```
partition
```

```
1 4
```

✓ CHAMELEON

Number of returned partition = 3:

```
select max( Cluster_number) as partition from
```

```
Data_mart01.DM_fact_cluster01
```

```
partition
```

```
1 3
```

E. 3 DBSCAN

DBSCAN is a density based clustering algorithm

(Jain, A.K.)

- Density is number of points within a specified radius (*Eps*).
- A point is a *core point* if it has more than the Specified number of points (*MinPts*) within *Eps*
- Core point is in the interior of a cluster.
- A *border point* has fewer than *MinPts* within *Eps* but is in neighbourhood of a core point.
- A *noise point* is any point that is neither a core point nor a border point

E. 4 deployment in production

Before process run ;

```
select distinct Cluster_number from
```

```
Data_mart01.DM_fact_cluster01
```

⇒ Only one record is returned

```
Cluster_number
```

```
1 0
```

After script execution;

E. 5 Time reponse of each process:

We will process many runs, by changing the record count of the input data set; and will calculate each time the time response value of the process. The following Unix command will be used:

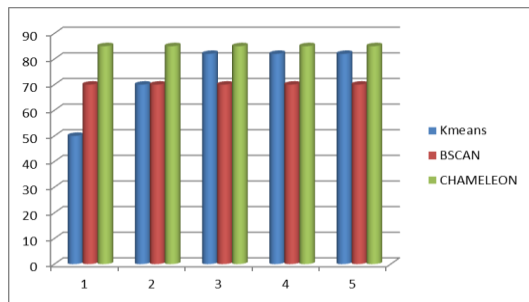
Example:

```
univ_med1@aa799dev]$time.  
/cluster.ksh
```

```
real 0m25.491s  
user 0m20.236s  
sys 0m20.140s
```

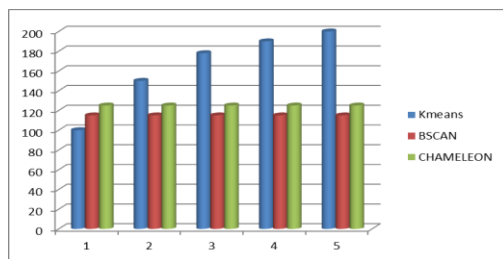
Below we display the results:

#### N=4000 record count of the input data set



Notice that the number of partitions remains the same for DBSCAN & CHAMELEON.

#### N=40000 record count of the input data set



## V. CONCLUSION

It is widely recognised that clustering algorithms are for real world analysis. In this paper we have shown that the DbSCAN and Chameleon algorithms represent the same trends and present the same realtime regardless the partition number, K mean algorithm show a dependency between the realtime and the number of partitions.

This process may be tuned for further use and analysis.

## VI. REFERENCE

**Inmon, W.H.**, 1996. Building the Data Warehouse. Wiley and Sons, NY.

**Jain, A.K.** ; Dept. of Comput. Sci. & Eng., Michigan State Univ., East Lansing, MI, USA ; Topchy, A. ; Law, M.H.C. ; Buhmann, J.M.

**J. MBARKI** Deployment of Partitioning Around Medoids Clustering Algorithm on a Set of Objects Derived from Analytical CRM Data Research Journal of Applied Sciences, Engineering and Technology 7(4): 786-790, 2014 ISSN: 2040-7459; e-ISSN: 2040-7467