JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

A Comparitive Study on Mongo and Cassandra Database For Data Clustering

R. Sasikala

Department of Computer Application, National College, Trichy, India

Available online at: www.ijcseonline.org

Abstract— Databases provide data storage, extraction and manipulation by using SQL language. It has emerged as a backend to support Big Data applications. It is mainly characterized by horizontal scalability, schema-free data models, and easy cloud deployment. There are various NoSQL databases and the performance varies with different types based on node capacity, number of cores, replication actor, and different workloads. Hence, it is important to compare them in terms of their performance and verify how the performance is related to the different database. This paper focuses on comparison of Cassandra, MongoDB and HBase which are the most commonly used NoSQL databases. This comparison between NoSQL databases deploys them on yahoo cloud platform which uses different types of virtual machines and cluster sizes to study the effect of different configurations. The final result shows the performance of databases at different workload levels and the result can be compared to find out the best among these two databases. In this paper, the comparison of two data bases which are mongo db and Cassandra db algorithm are used to produce the result which is the best db for future data base.

Keywords— BigData, MongoDB, Cassandra db, Virtual machine

I. INTRODUCTION

Databases are considered as a vital part of the organization. It is being used all over the globe. Originally, relational database were used which helped in the storage, extraction and manipulation of large volumes of data. However, with the constant growth of data, relational databases have their own limitations. To overcome the limitations, a new database model was developed with additional features, known as NoSQL database (Non-relational database). These databases are much more efficient and were not limited to scalability and storage. NoSQL database emerged as a breakthrough technology and it is used as a complement to relational database

MongoDB is a flexible and scalable document oriented data store with dynamic schemas, autosharing, built-in replication and high availability, full and flexible index support, rich queries, aggregation. Mongo DB follows a master-slave approach, and it has an automatic failover feature where if a master server goes down, MongoDB can automatically failover to a backup slave and promote the slave to a master. Master-slave replication is the most general replication mode supported by MongoDB, very flexible for backup, failover, read scaling. Shading is MongoDB's approach to scaling out. Shading allows you to add more machines to handle increasing load and data size horizontally without affecting your application. Shading refers to the process of splitting data up and storing different portions of the data on different machines. The basic concept behind MongoDB's shading is to break up collections into smaller chunks. These chunks can be distributed across shards so that each shard is responsible for a subset of total data set.

Apache Cassandra in a nutshell is an open source, peer-peer distributed database structural design, decentralized, easily scalable, fault tolerant, highly available, eventually consistent, schema free, column oriented database. Cassandra has a peer to- peer distribution model, such that any given node is structurally identical to any other node that is, there is no "master" node that acts differently than a "slave" node. There is no need to store a value for individual column every time a new entity is stored. A cluster is a container for time signature spaces. A time signature space is the outermost container for data in Cassandra, but it's perfectly fine to create as many key spaces as the application needs.

II. LITERATURE REVIEW

Rick Cattell [1] has done a comprehensive survey on Scalable SQL and NoSQL databases. In this paper, He classify these system on their data model, consistency control, data storage, durability, availability, query support, and other dimensions into key-value, document, extended record and relational.

Bogdan George [3] has done a evaluation between several NoSQL databases with comment and notes. This term paper is trying to comment on the various NoSQL systems and to make a comparison based on qualitative and quatitative point of view between Cassandra, Hbase and MySQL.The quantitative evaluation criteria or view based on two sets, one related to size(number of records/rows/document store, number of node in an installation) and other related to performance(Read and write latency in both write and read intensive environment).. These systems cannot be used interchangeable for solving any type of problem, but choose between the two types of databases for a given instance.

Jing han et al. [4] has done a survey on NoSQL database. This thesis describe the background, basic characteristics, data model of classifies NoSQL databases according to the CAP theorem be the mainstream NoSQL databases and on the basis of properties to help enterprises to choose NoSQL. Based on the above knowledge of the mainstream NoSQL databases companies decide whether to use NoSQL. In their study observed that companies need to consider the following options when deciding which properties NoSQL are Data Model, CAP Support, Multi Data Center Support, Capacity, Performance, Query API, Reliability, Data Persistence, Rebalancing and Business Support.

Santhosh Kumar Gajendran [5] has done a survey on nosql database. The goal of is to understand the present need that have led to the evolution of NoSQL databases, why relational database.In their study, common concepts underlying these databases and how they compromise on ACID properties to achieve high scalability and availability. The NoSQL databases Dynamo, voldemort, CouchDB, MongoDB, BigTable, HBase and Cassandra based on License type, concurrency control, data storage and replication are surveyed . Each database and its implementation has strengths at addressing specific enterprise or cloud concerns such as being easy to operate, providing a flexible data model, high availability, high scalability and fault tolerance.

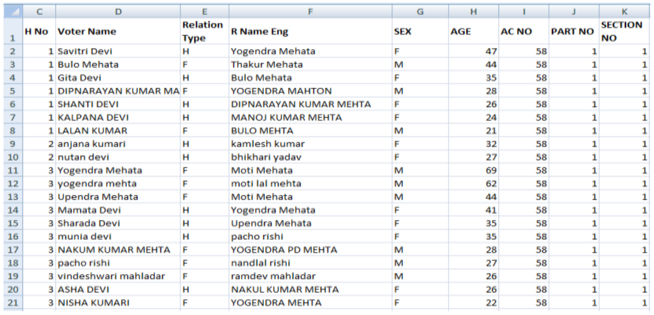
Manoj V [6] has done a comparative study on NoSQL databases are Cassandra, MongoDB and Hbase on basis of architecture and working. The parameter of study are classification, architecture, availability, data model, partitioning and evaluation of Cassandra as industry use case. In their study that MongoDB fits for use cases with document, document search and aggregation functions are mandate. HBase suits the scenarios in which hadoop map reduce is useful for bulk read and load operations and offers optimized read performance with hadoop platform. Cassandra can be used for applications requiring faster writes and high availability.

III. METHODOLOGY

3.1 DATASET

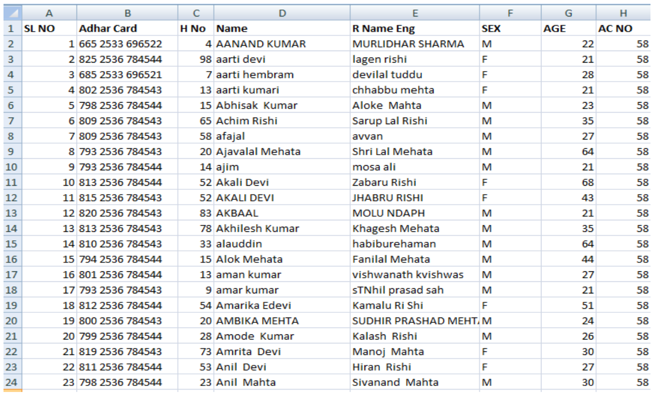
A group of related sets of data that is composed of separate element but can be manipulated as a unit by a computer.In a database, for example, a data set might contain a collection of business data like calculating PH value, comparing with .NET Software. The record itself can be considered a data set can bodies of data within it related to a particular type of information, such as sales data for a particular corporate department. A Vorter, Adhaar card dataset has been used with 800 items each for analysis. A set of association rules are obtained by applying K-Means, Navie bayes and decision tress. By analyzing the data, and giving different execution time, memory space we can obtain different number of rules.

Voter id



	C	D	E	F	G	H	I	J	K
1	H No	Voter Name	Relation Type	R Name Eng	SEX	AGE	AC NO	PART NO	SECTION NO
2	1	Savitri Devi	H	Yogendra Mehata	F	47	58	1	1
3	1	Bulo Mehata	F	Thakur Mehata	M	44	58	1	1
4	1	Gita Devi	H	Bulo Mehata	F	35	58	1	1
5	1	DIPNARAYAN KUMAR MA	F	YOGENDRA MAHTON	M	28	58	1	1
6	1	SHANTI DEVI	H	DIPNARAYAN KUMAR MEHTA	F	26	58	1	1
7	1	KALPANA DEVI	H	MANOJ KUMAR MEHTA	F	24	58	1	1
8	1	LALAN KUMAR	F	BULO MEHTA	M	21	58	1	1
9	2	anjana kumari	H	kamlesh kumar	F	32	58	1	1
10	2	nutan devi	H	bhikhari yadav	F	27	58	1	1
11	3	Yogendra Mehata	F	Moti Mehata	M	69	58	1	1
12	3	yogendra mehta	F	moti lal mehta	M	62	58	1	1
13	3	Upendra Mehata	F	Moti Mehata	M	44	58	1	1
14	3	Mamata Devi	H	Yogendra Mehata	F	41	58	1	1
15	3	Sharada Devi	H	Upendra Mehata	F	35	58	1	1
16	3	munia devi	H	pacho rishi	F	35	58	1	1
17	3	NAKUM KUMAR MEHTA	F	YOGENDRA PD MEHTA	M	28	58	1	1
18	3	pacho rishi	F	nandlal rishi	M	27	58	1	1
19	3	vindeshwari mahladar	F	ramdev mahladar	M	26	58	1	1
20	3	ASHA DEVI	H	NAKUL KUMAR MEHTA	F	26	58	1	1
21	3	NISHA KUMARI	F	YOGENDRA MEHTA	F	22	58	1	1

Aadhar id



	A	B	C	D	E	F	G	H
1	SL NO	Adhar Card	H No	Name	R Name Eng	SEX	AGE	AC NO
2	1	665 2533 696522	4	AANAND KUMAR	MURLIDHAR SHARMA	M	22	58
3	2	825 2536 784544	98	aarti devi	lagen rishi	F	21	58
4	3	685 2533 696521	7	aarti hembram	devilal tuddu	F	28	58
5	4	802 2536 784543	13	aarti kumari	chhabbu mehta	F	21	58
6	5	798 2536 784544	15	Abhisak Kumar	Aloke Mahta	M	23	58
7	6	809 2536 784543	65	Achim Rishi	Sarup Lal Rishi	M	35	58
8	7	809 2536 784543	58	afajal	avvan	M	27	58
9	8	793 2536 784543	20	Ajavalal Mehata	Shri Lal Mehata	M	64	58
10	9	793 2536 784544	14	ajim	mosa ali	M	21	58
11	10	813 2536 784544	52	Akali Devi	Zabaru Rishi	F	68	58
12	11	815 2536 784543	52	AKALI DEVI	JHABRU RISHI	F	43	58
13	12	820 2536 784543	83	AKBAAL	MOLU NDAPH	M	21	58
14	13	813 2536 784543	78	Akhilesh Kumar	Khagesh Mehata	M	35	58
15	14	810 2536 784543	33	alauddin	habiburehaman	M	64	58
16	15	794 2536 784544	15	Alok Mehata	Fanilal Mehata	M	44	58
17	16	801 2536 784544	13	aman kumar	vishwanath kvishwas	M	27	58
18	17	793 2536 784543	9	amar kumar	sTNhil prasad sah	M	21	58
19	18	812 2536 784544	54	Amarika Edevi	Kamalu Ri Shi	F	51	58
20	19	800 2536 784543	20	AMBIKA MEHTA	SUDHIR PRASHAD MEHT	M	24	58
21	20	799 2536 784544	28	Amode Kumar	Kalash Rishi	M	26	58
22	21	819 2536 784543	73	Amrita Devi	Manoj Mahta	F	30	58
23	22	811 2536 784544	53	Anil Devi	Hiran Rishi	F	27	58
24	23	798 2536 784544	23	Anil Mahta	Sivanand Mahta	M	30	58

3.2 ALGORITHM

Naïve Bayesian Classifier:

Theorem with well-built autonomy assumption between the feature. It is a highly scalable require a figure of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

The probabilities applied in the Naïve Bayes algorithm are calculated according to the Bayes' Rule. The probability of the likelihood of some conclusion S , given some evidence or observation T, where a dependence relationship between S and T, denoted as P(S|T) , can be calculated based on Eq. 1

$$P(S \mid T) = \frac{P(T \mid S) * P(S)}{P(T)}$$

The Bayes Naive classifier selects the most likely classification Vnb given the attribute values a1; a2; : : : an. This results in:

Vnb = argmaxvj2V P(vj)YP(aijvj) (1)

We generally estimate P(aijvj) using m-estimates:

P(aijvj) = nc + mp

―――――

n+m

Where:

n = the number of training examples for which v = vj
nc = number of examples for which v = vj and a = ai
p = a priori estimate for P(aijvj)
A. m = the equivalent sample size

K Means Clustering:

K-means clustering aim to partition n explanation into k clusters in which each observation belong to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

K means clustering with example

The steps of the K-means algorithm are written below:
1. Initialization: Randomly *K* data points are chosen to initialize the cluster centers.
2. Nearest-neighbor search: Each data point is assigned to the cluster center that is closest to it.
The distance from the data vector to the centroid is calculated using the following equation.

$$d(z_p, a_j) = \sqrt{\sum_{k=1}^{d} (z_{pk} - a_{jk})^2}$$

Where d is the dimension of the data vector, z_p is the centroid of cluster p and a_j is the data vector.
3. Mean update: New cluster centers are calculated finding the mean of the input vectors assigned to a particular cluster.

IV. RESULT AND DISCUSSION

A Vorter, Adhaar card dataset has been used with 800 items each for analysis. A set of association rules are obtained by applying K-Means, Navie bays and decision tress. By analyzing the data, and giving different execution time, memory space we can obtain different number of rules. During analysis it found that Genetic is much faster for large number of transactions as compare to K-means. It takes less time to generate frequent item sets. We work on mogodb, cassendra data which contains transactions. All the results are collected from Pentium Dual core processor with 1. 73GHz speed and 1 - GB RAM

Mango DB

MongoDB is a flexible and scalable document oriented data store with dynamic schemas, autosharing, built-in replication and high availability, full and flexible index support, rich queries, aggregation

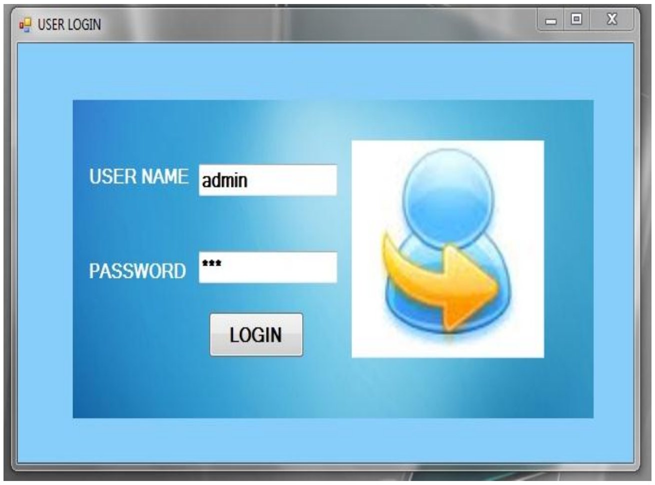
Cassandra DB

Apache Cassandra in a nutshell is an open source, peer-peer distributed database structural design, decentralized, easily scalable, fault tolerant, highly available, eventually consistent, schema free, column oriented database.

Data sets contain voter id and aadhar card.

Preprocess:

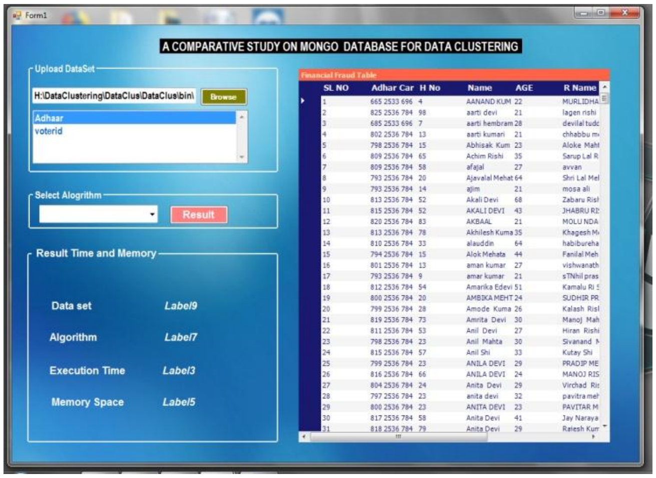
A preprocessor is a program that processes its input data to produce output that is used as input to another program. The output is said to be a preprocessed form of the input data, which is often used by some subsequent programs like compilers. The amount and kind of processing done depends on the nature of the preprocessor; some preprocessors are only capable of performing relatively simple textual substitutions and macro expansions, while others have the power of full-fledged programming languages.



USER NAME admin

PASSWORD ***

Figure 4.1 : Login MainPage



H:\DataClustering\DataClus\DataClus\bin\

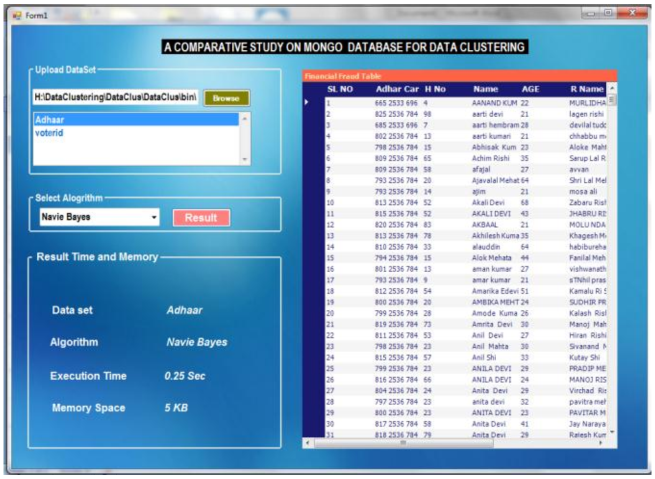
Adhaar
voterid

Data set Label9
Algorithm Label7
Execution Time Label3
Memory Space Label5

Financial Fraud Table

SL NO	Adhar Car	H No	Name	AGE	R Name
1	665 2533 696	4	AANAND KUM	22	MURLIDHA
2	825 2536 784	98	aarti devi	21	lagen rishi
3	685 2533 696	7	aarti hembram	28	devilal tudc
4	802 2536 784	13	aarti kumari	21	chhabbu m
5	798 2536 784	15	Abhisak Kum	23	Aloke Maht
6	809 2536 784	65	Achim Rishi	35	Sarup Lal R
7	809 2536 784	58	afajal	27	avvan
8	793 2536 784	20	Ajavalal Mehat	64	Shri Lal Mel
9	793 2536 784	14	ajim	21	mosa ali
10	813 2536 784	52	Akali Devi	68	Zabaru Risl
11	815 2536 784	52	AKALI DEVI	43	JHABRU RI
12	820 2536 784	83	AKBAAL	21	MOLU NDA
13	813 2536 784	78	Akhilesh Kuma	35	Khagesh M
14	810 2536 784	33	alauddin	64	habibureha
15	794 2536 784	15	Alok Mehata	44	Fanilal Meh
16	801 2536 784	13	aman kumar	27	vishwanath
17	793 2536 784	9	amar kumar	21	sTNhil pras
18	812 2536 784	54	Amarika Edevi	51	Kamalu Ri
19	800 2536 784	20	AMBIKA MEHT	24	SUDHIR PR
20	799 2536 784	28	Amode Kuma	26	Kalash Risl
21	819 2536 784	73	Amrita Devi	30	Manoj Mah
22	811 2536 784	53	Anil Devi	27	Hiran Rishi
23	798 2536 784	23	Anil Mahta	30	Sivanand
24	815 2536 784	57	Anil Shi	33	Kutay Shi
25	799 2536 784	23	ANILA DEVI	29	PRADIP ME
26	816 2536 784	66	ANILA DEVI	24	MANOJ RIS
27	804 2536 784	24	Anita Devi	29	Virchad Ris
28	797 2536 784	23	anita devi	32	pavitra meh
29	800 2536 784	23	ANITA DEVI	23	PAVITAR M
30	817 2536 784	58	Anita Devi	41	Jay Naraya
31	818 2536 784	79	Anita Devi	29	Rajesh Kum

Figure 1. Figure 4.2 Upload Adhaar card Mongo dataset



H:\DataClustering\DataClus\DataClus\bin\

Adhaar
voterid

Navie Bayes

Data set Adhaar

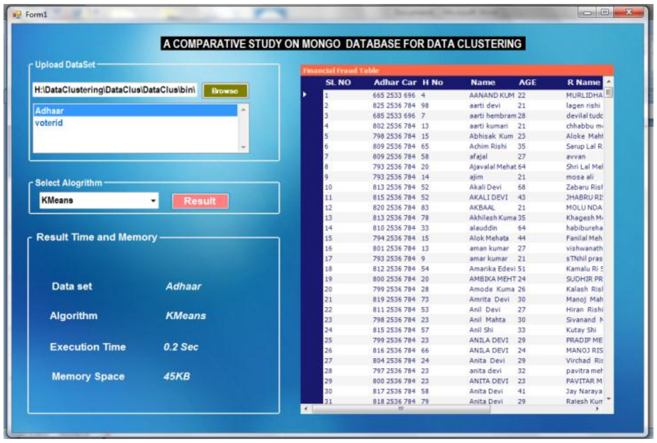
Algorithm Navie Bayes

Execution Time 0.25 Sec

Memory Space 5 KB

SL NO	Adhar Car	H No	Name	AGE	R Name
1	665 2533 696	4	AANAND KUM	22	MURLIDHA
2	825 2536 784	98	aarti devi	21	lagen rishi
3	685 2533 696	7	aarti hembram	28	devilal tudc
4	802 2536 784	13	aarti kumari	21	chhabbu m
5	798 2536 784	15	Abhisak Kum	23	Aloke Maht
6	809 2536 784	65	Achim Rishi	35	Sarup Lal R
7	809 2536 784	58	afajal	27	avvan
8	793 2536 784	20	Ajavalal Mehat	64	Shri Lal Mel
9	793 2536 784	14	ajim	21	mosa ali
10	813 2536 784	52	Akali Devi	68	Zabaru Risl
11	815 2536 784	52	AKALI DEVI	43	JHABRU RI
12	820 2536 784	83	AKBAAL	21	MOLU NDA
13	813 2536 784	78	Akhilesh Kuma	35	Khagesh M
14	810 2536 784	33	alauddin	64	habibureha
15	794 2536 784	15	Alok Mehata	44	Fanilal Meh
16	801 2536 784	13	aman kumar	27	vishwanath
17	793 2536 784	9	amar kumar	21	sTNhil pras
18	812 2536 784	54	Amarika Edevi	51	Kamalu Ri S
19	800 2536 784	20	AMBIKA MEHT	24	SUDHIR PR
20	799 2536 784	28	Amode Kuma	26	Kalash Risl
21	819 2536 784	73	Amrita Devi	30	Manoj Mah
22	811 2536 784	53	Anil Devi	27	Hiran Rishi
23	798 2536 784	23	Anil Mahta	30	Sivanand M
24	815 2536 784	57	Anil Shi	33	Kutay Shi
25	799 2536 784	23	ANILA DEVI	29	PRADIP ME
26	816 2536 784	66	ANILA DEVI	24	MANOJ RIS
27	804 2536 784	24	Anita Devi	29	Virchad Ri
28	797 2536 784	23	anita devi	32	pavitra mel
29	800 2536 784	23	ANITA DEVI	23	PAVITAR M
30	817 2536 784	58	Anita Devi	41	Jay Naraya
31	818 2536 784	79	Anita Devi	29	Rajesh Kum

Figure 4.3 Mongo Dataset using Navie bayes in Time & Memory space



H:\DataClustering\DataClus\DataClus\bin\

Adhaar
voterid

KMeans

Data set Adhaar

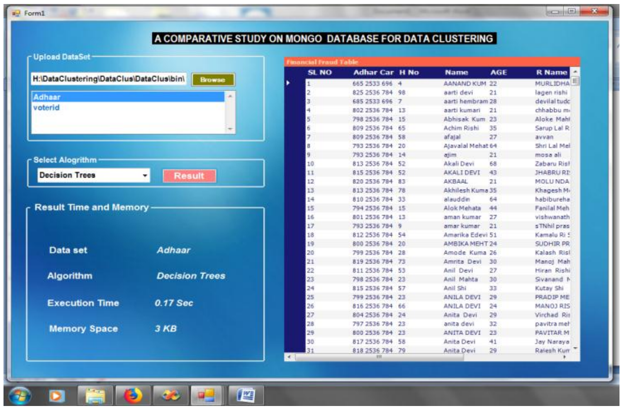
Algorithm KMeans

Execution Time 0.2 Sec

Memory Space 45KB

SL NO	Adhar Car	H No	Name	AGE	R Name
1	665 2533 696	4	AANAND KUM	22	MURLIDHA
2	825 2536 784	98	aarti devi	21	lagen rishi
3	685 2533 696	7	aarti hembram	28	devilal tudc
4	802 2536 784	13	aarti kumari	21	chhabbu m
5	798 2536 784	15	Abhisak Kum	23	Aloke Maht
6	809 2536 784	65	Achim Rishi	35	Sarup Lal R
7	809 2536 784	58	afajal	27	avvan
8	793 2536 784	20	Ajavalal Mehat	64	Shri Lal Mel
9	793 2536 784	14	ajim	21	mosa ali
10	813 2536 784	52	Akali Devi	68	Zabaru Risl
11	815 2536 784	52	AKALI DEVI	43	JHABRU RI
12	820 2536 784	83	AKBAAL	21	MOLU NDA
13	813 2536 784	78	Akhilesh Kuma	35	Khagesh M
14	810 2536 784	33	alauddin	64	habibureha
15	794 2536 784	15	Alok Mehata	44	Fanilal Meh
16	801 2536 784	13	aman kumar	27	vishwanath
17	793 2536 784	9	amar kumar	21	sTNhil pras
18	812 2536 784	54	Amarika Edevi	51	Kamalu Ri S
19	800 2536 784	20	AMBIKA MEHT	24	SUDHIR PR
20	799 2536 784	28	Amode Kuma	26	Kalash Risl
21	819 2536 784	73	Amrita Devi	30	Manoj Mah
22	811 2536 784	53	Anil Devi	27	Hiran Rishi
23	798 2536 784	23	Anil Mahta	30	Sivanand M
24	815 2536 784	57	Anil Shi	33	Kutay Shi
25	799 2536 784	23	ANILA DEVI	29	PRADIP ME
26	816 2536 784	66	ANILA DEVI	24	MANOJ RIS
27	804 2536 784	24	Anita Devi	29	Virchad Ri
28	797 2536 784	23	anita devi	32	pavitra mel
29	800 2536 784	23	ANITA DEVI	23	PAVITAR M
30	817 2536 784	58	Anita Devi	41	Jay Naraya
31	818 2536 784	79	Anita Devi	29	Rajesh Kum

Figure 4.4 Adhaar Card Mongo Dataset using K-means in Time & Memory space



H:\DataClustering\DataClus\DataClus\bin\

Adhaar
voterid

Decision Trees

Data set Adhaar

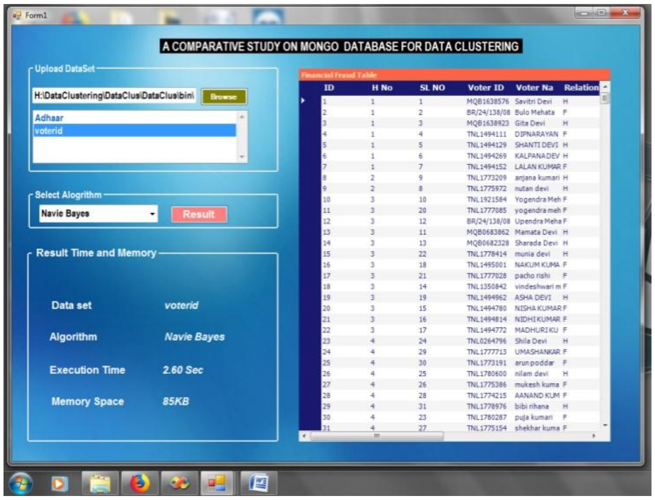
Algorithm Decision Trees

Execution Time 0.17 Sec

Memory Space 3 KB

SL NO	Adhar Car	H No	Name	AGE	R Name
1	665 2533 696	4	AANAND KUM	22	MURLIDHA
2	825 2536 784	98	aarti devi	21	lagen rishi
3	685 2533 696	7	aarti hembram	28	devilal tudc
4	802 2536 784	13	aarti kumari	21	chhabbu m
5	798 2536 784	15	Abhisak Kum	23	Aloke Maht
6	809 2536 784	65	Achim Rishi	35	Sarup Lal R
7	809 2536 784	58	afajal	27	avvan
8	793 2536 784	20	Ajavalal Mehat	64	Shri Lal Mel
9	793 2536 784	14	ajim	21	mosa ali
10	813 2536 784	52	Akali Devi	68	Zabaru Risl
11	815 2536 784	52	AKALI DEVI	43	JHABRU RI
12	820 2536 784	83	AKBAAL	21	MOLU NDA
13	813 2536 784	78	Akhilesh Kuma	35	Khagesh M
14	810 2536 784	33	alauddin	64	habibureha
15	794 2536 784	15	Alok Mehata	44	Fanilal Meh
16	801 2536 784	13	aman kumar	27	vishwanath
17	793 2536 784	9	amar kumar	21	sTNhil pras
18	812 2536 784	54	Amarika Edevi	51	Kamalu Ri S
19	800 2536 784	20	AMBIKA MEHT	24	SUDHIR PR
20	799 2536 784	28	Amode Kuma	26	Kalash Risl
21	819 2536 784	73	Amrita Devi	30	Manoj Mah
22	811 2536 784	53	Anil Devi	27	Hiran Rishi
23	798 2536 784	23	Anil Mahta	30	Sivanand M
24	815 2536 784	57	Anil Shi	33	Kutay Shi
25	799 2536 784	23	ANILA DEVI	29	PRADIP ME
26	816 2536 784	66	ANILA DEVI	24	MANOJ RIS
27	804 2536 784	24	Anita Devi	29	Virchad Ri
28	797 2536 784	23	anita devi	32	pavitra mel
29	800 2536 784	23	ANITA DEVI	23	PAVITAR M
30	817 2536 784	58	Anita Devi	41	Jay Naraya
31	818 2536 784	79	Anita Devi	29	Rajesh Kum

Figure 4.5 Adhaar Card Mongo Dataset using Decision Trees in Time & Memory space



H:\DataClustering\DataClus\DataClus\bin\

Adhaar
voterid

Navie Bayes

Data set voterid

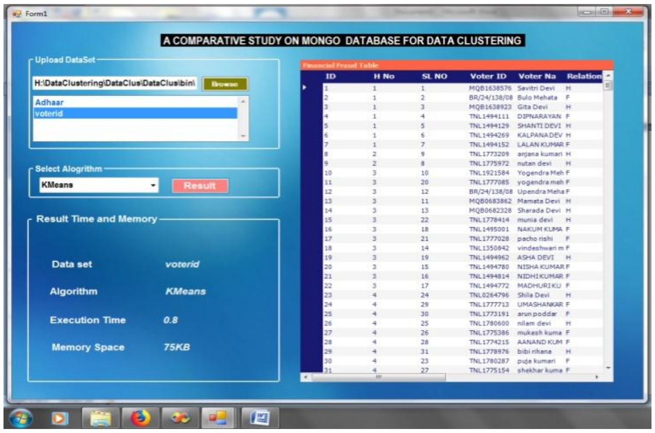
Algorithm Navie Bayes

Execution Time 2.60 Sec

Memory Space 85KB

ID	H No	SL NO	Voter ID	Voter Na	Relation
1	1	1	MQB1638576	Savitri Devi	H
2	1	2	BR/24/138/08	Bulo Mehata	F
3	1	3	MQB1638923	Gita Devi	H
4	1	4	TNL1494111	DIPNARAYAN	F
5	1	5	TNL1494129	SHANTI DEVI	H
6	1	6	TNL1494269	KALPANA DEV	H
7	1	7	TNL1494152	LALAN KUMAR	F
8	2	9	TNL1773209	anjana kumari	H
9	2	8	TNL1775972	nutan devi	H
10	3	10	TNL1921584	Yogendra Meh	F
11	3	20	TNL1777085	yogendra meh	F
12	3	12	BR/24/138/08	Upendra Meha	F
13	3	11	MQB0683862	Mamata Devi	H
14	3	13	MQB0682328	Sharada Devi	H
15	3	22	TNL1778414	munia devi	H
16	3	18	TNL1495001	NAKUM KUMA	F
17	3	21	TNL1777028	pacho rishi	F
18	3	14	TNL1350842	vindeshwari m	F
19	3	19	TNL1494962	ASHA DEVI	H
20	3	15	TNL1494780	NISHA KUMAR	F
21	3	16	TNL1494814	NIDHI KUMAR	F
22	3	17	TNL1494772	MADHURI KU	F
23	4	24	TNL0264796	Shila Devi	H
24	4	29	TNL1777713	UMASHANKAR	F
25	4	30	TNL1773191	arun poddar	F
26	4	25	TNL1780600	nilam devi	H
27	4	26	TNL1775386	mukesh kuma	F
28	4	28	TNL1774215	AANAND KUM	F
29	4	31	TNL1778976	bibi rihana	H
30	4	23	TNL1780287	puja kumari	F
31	4	27	TNL1775154	shekhar kuma	F

Figure 4.6 Voter ID Mongo Dataset using Navie Bayes in Time & Memory space



H:\DataClustering\DataClus\DataClus\bin\

Adhaar
voterid

KMeans

Data set voterid

Algorithm KMeans

Execution Time 0.8

Memory Space 75KB

ID	H No	SL NO	Voter ID	Voter Na	Relation
1	1	1	MQB1638576	Savitri Devi	H
2	1	2	BR/24/138/08	Bulo Mehata	F
3	1	3	MQB1638923	Gita Devi	H
4	1	4	TNL1494111	DIPNARAYAN	F
5	1	5	TNL1494129	SHANTI DEVI	H
6	1	6	TNL1494269	KALPANA DEV	H
7	1	7	TNL1494152	LALAN KUMAR	F
8	2	9	TNL1773209	anjana kumari	H
9	2	8	TNL1775972	nutan devi	H
10	3	10	TNL1921584	Yogendra Meh	F
11	3	20	TNL1777085	yogendra meh	F
12	3	12	BR/24/138/08	Upendra Meha	F
13	3	11	MQB0683862	Mamata Devi	H
14	3	13	MQB0682328	Sharada Devi	H
15	3	22	TNL1778414	munia devi	H
16	3	18	TNL1495001	NAKUM KUMA	F
17	3	21	TNL1777028	pacho rishi	F
18	3	14	TNL1350842	vindeshwari m	F
19	3	19	TNL1494962	ASHA DEVI	H
20	3	15	TNL1494780	NISHA KUMAR	F
21	3	16	TNL1494814	NIDHI KUMAR	F
22	3	17	TNL1494772	MADHURI KU	F
23	4	24	TNL0264796	Shila Devi	H
24	4	29	TNL1777713	UMASHANKAR	F
25	4	30	TNL1773191	arun poddar	F
26	4	25	TNL1780600	nilam devi	H
27	4	26	TNL1775386	mukesh kuma	F
28	4	28	TNL1774215	AANAND KUM	F
29	4	31	TNL1778976	bibi rihana	H
30	4	23	TNL1780287	puja kumari	F
31	4	27	TNL1775154	shekhar kuma	F

Figure 4.7 Voter ID Mongo Dataset using Kmeans in Time & Memory space

Table 1.Comparison of two dataset using k-means, NB algorithms & Decision Trees

Dataset	K-Means Time Taken (in secs.)	NB Time Taken (in secs.)	DT Time Taken (in secs.)
Adhaar Card	0.2	0.25	0.17
Voter ID	0.8	2.60	1.45

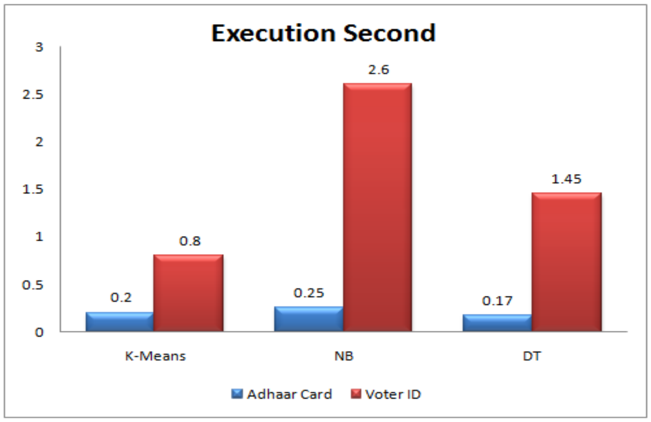
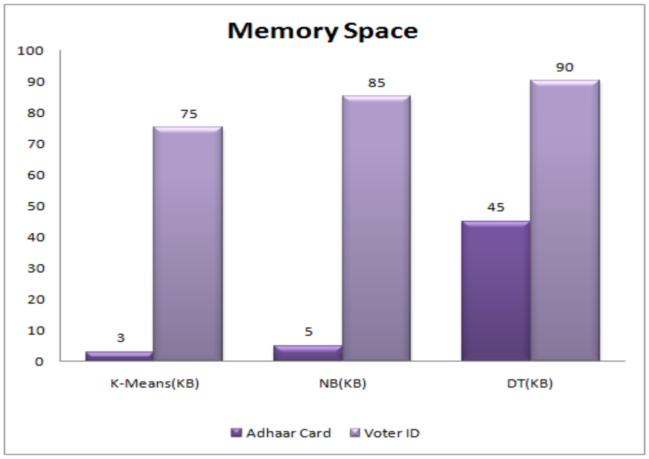


Figure 4.8 comparisons of two dataset using k-means, NB algorithms & decision trees execution time

Table 2.Comparison of two dataset using k-means, NB algorithms & decision trees memory space

Dataset	K-Means Memory Space	NB Memory Space	DT Memory Space
Adhaar Card	3	5	45
Voter ID	75	85	90

Figure 8.1.2 comparison of two dataset using k-means, NB algorithms & decision trees memory space



V. CONCLUSION

In conclusion, the project work proposed a method which combines SQL database which belongs to the relational group of database systems and MongoDB being a NoSQL database to store and manage big data. The result obtained it is understandable that system can be used for storage space and administration of big eliminating the weaknesses in both databases. This project produce the best result in "k-means algorithm" used in mongodb and Cassandra db while comparing other algorithm like naïve bayseian and k means clustering algorithm.

VI. FUTURE ENHANCEMENT

MongoDB has newly come into existence, whereas the standard SQL language has been over years and, therefore if we merge the two we can use the features of both the database. Although, NoSQL (MongoDB) has the advantage of horizontal expansion, but for complex SQL requests, it cannot support them very well. For the Query based on KEY/VALUE and massive data storage requirements, NOSQL is a very worth doing choice for me and all other developers and organizations who's developed big applications.

REFERENCES

[1] Venkat N Gudivada,Dhana Rao,Vijay V Raghavan,"Nosql systems for Big Data Management "IEEE 2014,DOI 10.1109/SERVICES .2014.42,pp:190-197.

[2] Thomas Sandholm,Dongman Lee,"Notes on Cloud Computing Principles"in Journal of Cloud Computing:Advances,Systems and applications,springer 2014.

[3] Divyakant Agarwal,Sudipto Das,Amr EI Abbadi, "Bigdata and Cloud Computing:Current State and Future opportunities",ACM 2011.

[4] Nani Fadzlina Naim, Ahmad Ihsan MohdYassin, Wan Mohd Ameerul Wan Zamri, Suzi Seroja Sarnin, "Mysql Database for storage of finger print data" IEEE 2011, DOI 10.1109/UKSIM.2011.62,pp:293-298.

[5] Sudhanshu Kulshreshta, Shelly Sachdeva, "Performance for Data Storage-DB4o and Mysql Databases", IEEE 2014.

[6] Mehul Nalin Vora, "Hadoop-HBase for Large Scale Data" , IEEE 2011,pp:601-605.

[7] Gansen Zhao, Weichai Huang, ShunlinLiang, Yong Tang, "Modelling MongoDB with Relational Model", IEEE 2013,DOI 10.1109/EIDWT.2013.25,pp:115-121.

[8] Shalini Ramanathan, Savita Goel, Subramanian Alagumlai, "Comparison of Cloud Database: Amazon‟s SimpleDB and Google‟s BigTable" ,in IEEE 2011 and International Journal of Computer Science Issues(IJCSI), Vol 8,Issue 6,No 2,Nov 2011,ISSN:1694-0814.

[9] Jing Han, Hai Hong E, Guan Le, Jian Du, "Survey on Nosql Databases" IEEE 2011, pp:363-366.