International Journal of Innovative Studies in Sciences and Engineering Technology (IJISSET)

ISSN 2455-4863 (Online) www.ijisset.org *Volume: 3 Issue: 8 | August 2017*

A Novel Approach of Using MongoDB for Big Data Analytics

***Ananth G S, +K Raghuveer**

**Assistant Professor, Dept. of MCA, NIE, Mysore*
+Prof.& Head, Dept. of IS & E, NIE Mysore

Abstract: *Everything is "Data" In this current world. Everything generates a lot of data. A simple example of a smart phone - generates loads of data daily like - phone logs, message logs, mails data etc. There are humongous devices today and just imagine each device loading up torts of such data. Likewise there are more than a trillion websites today (There can't be anymore websites that could be created with the IPv4 protocol shortly) Each website puts up data In the form of ZetaBytes or even more. Facebook per se handles more than 30+ Peta bytes of user data. What is small today is big tomorrow. 'Ibis is tailed the world of Big Data. It's not a trend anymore rather a boon and exponential growth of data. Storage of such data and processing is always a problem. Traditional RDRMS based databases wont be of help anymore. We need databases that can store huge data, process them in a faster sequence time and help analysis. In this paper we discuss one such NoSQL based database called MongoDB. We discuss how MongoDB can help solve the problem of big data analytics.*

Keywords: *Analytics, Big Data, MongoDB, NoSQL.*

1. INTRODUCTION - WHAT IS DATA ANALYTICS?

Data analytics is a science. Its an art of examining raw data with a purpose of drawing conclusions about information. It is used in almost all kinds of industries to your household purposes. We purchase a product and store its receipt, we store our t e l ephon e bi l l s , manag e e s s ent i a l commodities bills etc. We mainly manage expenses for a month of our salary. This is data and its analytics to a common man.

2. WHAT IS BIG DATA?

The term big data has apparently been around for decades, and we seem to be doing analytics all day and night. But the point is - its not big, its bigger! We have been watching movies right from the inception of television or from the time of photography and videography began. But just read this line as is: Youtube from Google has users uploading 48 hours of video every single minute of the day[1] Also a message by the computer giant IBM - the year 2012 had generated around 2.5 exabytes of data - which is approximately 2.5 billion Gb[2].

2.1 Characteristics of Big Data - There are more than 4 to 5 V's that characterize Big Data. Gartner characterized into 3 important V's. I. Volume - how much data is generated daily? How huge is the data taken into picture? 2. Variety - big data Is not single data. Its not data of one type or of one extension. Its varied data, unstructured data, its data from different sources. How to analyze such data? How to aggregate an assumption from such data? 3. Velocity - Data is not generated slowly. Its exponential in growth. Its rapid in nature. Its in a flow. Data is in motion. Its real time in nature. How to get desired results from such data? Lastly 4. Value - A data consisting of say volume, variety, velocity leave apart the other Vs - how to get a meaning out of such data? How important is it to store such data? How good is it to process such data? Welcome to the world of Big Data Analytics! [3]

2.2 Processing Big Data - Now here comes the real problem. Processing data that is so huge in nature. How on earth can we store, process, examine, manipulate this huge Zbytes of generated data? Walmart for example handles more than **1** million customer transactions every hour[4] What if we wanted to do analytics to find out - who was that person who used an AXIS Bank Credit card ending with numbers 89.

3. TYPES OF DATA - STRUCTURED V/S UNSTRUCTURED DATA - THE CONCEPT

The beginning world only saw data that was structured in nature. An example of structured data a fable Employee - consisting of fields Nate, ID, Age, Gender and table employee eventually would store multiple tuples. But today the world is majorly consisting of unstructured data. 30 Billion pieces of content is shared on Facebook *every* rnonth[5]. This does not consist of simple data as per the above example. Data is an image, a link, a website, an audio piece, some text message, and lot more. This is where the V - Variety comes into picture.

4. TYPES OF DATABASES - TRADITIONAL RDBMS V / S MODERN NO SQL DATABASES - THE DIFFERENCE.

The concept of a database started way back in the ye a r 1965 when IBM started IMS (Information Management Systems), 1970 when Codd published a relational model paper. The language of databases SQL was invented in the year 1975. Google published **a** white paper that lead to the calling of the term Big Data called the "Big Table" sometime in 2002-03. But what are some of the differences between a traditional database to a NoSQL database? Here is the answer - RDBMS is completely structured w a y o f storing data . While the NoSQL is unstructured way of storing the data. And another main difference Is that the amount of data stored mainly depends on the Physical memory of the system. While in the NoSQL you don't have any such limits as you can scale the S y s t e m h o r i z o n t a l l y . "Extremely large datasets are often event based transactions that occur in chronological order. Examples are weblogs, shopping transactions, manufacturing data from assembly line devices, scientific data collections, etc. These types of data accumulate in large numbers every second and can take a RDBMS with all of its overhead to its knees. But for OLTP processing, nothing beats the combination of data quality and performance of a well designed RDBMS." [6]

4.1 The concept of NoSQL - NoSQL is a very broad term and typically is referred to as meaning "Not Only SQL: and usually means that the database is nor a relational database, which have been very popular the last decades.

4.2 Why NoSQL - the reason why NoSQL has been so popular the last few years is mainly because, when a relational database grows out of one server, it is no longer that easy to use. In other words, they don't scale out very well in a distributed system. All of the big sites Google. Yahoo, Facebook and Amazon have lots of data and store the data in distributed systems for several reasons. It could be that the data doesn't fit on one server, or their ace requirements for high availability.

NoSQL, also follows the CAP Theorem. The properties of a distributed system can be described by the CAP Theorem. Of the three properties you can only have at most two[7]: Consistency Availability Tolerance to network partitioning Examples of NoSQL based databases -Currently there are more than 250 NoSQL databases[8]. Many of them are based on separate user requirements. For example for wide tolumn store families - we have Hadoop, MapR, Apache Cassandra, Google CloudData, et c. For document s torage - we have ElasticSearch, MongoDB, Couchbase server, Couch DB, SequiaDB, Terastore etc. For key value multiple store we have DynamoDB, Riak, R e d i s - etc.

5. MONGODB AS THE CHOICE OF USAGE - WHY MONGODB?

Inventing the wheel is not necessary unless the traditional systems do have some problems. The traditional system of RDBMS had enough challenges: 1. Volume and new characteristics -RDBMS systems could be scaled horizontally but not vertically, many of the servers were commodity based in nature, could not be used during distributed usage or for purposes like applications dealing with cloud computing. 2. Data variety and volatility - Going on and on it became extremely difficult to store data of various types. As Gartner rightly told - 3Vs or the 4Vs of big data and its processing was ext r eme ly di ff i cul t . 3. RDBMS wa s a transaction based model unfit for distributed systems. The MongoDB is a child product of the company 10Gen – In the year 2007 Eliot Horowitz and Dwight Merriman were tired of reinventing the wheel, they started the company 10Gen. The MongoDB development began. In the year 2009 there was an initial release of MongoDB to the community.

5.1 **MongoDB** characteristics -

These are some of the characteristics of MongoDB -

1. It can store different products in the form of collections.

2. It is horizontally scalable.

3. Its flexible to use.

4. Data is mapped to a key.

5. Data is stored on disk in a column oriented fashion.

6. Its predominantly hash based indexing concept.

7. Data can be partitioned by range or consistent hashing.

8. These are all characteristics of the NoSQL. databases.

9. Data is stored in the form of documents. Either as XML or JSON.

10. It is great for grinding through data.

5.2 Installation of MongoDB -

Our work environment for this paper was a Linux environment with Linux Mint 17 LMDE as the

distribution. The MongoDB was run on a single cluster mode with hardware being Intel Pentium i5, 4GB RAM.

As the root user Installing MongoDB using Synaptic Package manager of Linux Mint: Search for mongodb and right click to install it.

5 .3 Installing Mong oDB us ing CLI (Command line interface):

$ sudo apt-get install mongodb – select all dependencies.

To start the service of MongoDB

$*sudo service mongodb start*

List of all commands – db.help() inside the mongodb client (here CLI)

5.4 **MongoDB data modelling – MongoDB database creation -**

>use mydb

To check the currently used database

>db

To check the databases list

>show dbs

To create a collection – What is a collection **in MongoDB?** See below table

Table 1: *Difference between RDBMS and MongoDB*

RDBMS	MongoDB
Database	Database
Table	Collection
Row	Document

So, a table in a traditional **RDBMS** is a collection in MongoDB. Likewise a Row in RDBMS is a document in MongoDB as per the above **table 1.**

Lets create a collection and insert a document into it -

The command(s) goes as follows:

>db.colleges.insert({"college_name":"Hassan, MCE"})

>db.colleges.insert({"college_name":"NIE"})

>db.colleges.insert({"college_name":"SJCE"})

To list only the collections present in the

database "mydb":

>show collections

colleges

system.indexes

To view data created -

>db.colleges.find()

{

"_id":ObjectID(7d178ad8902c),

"college_name":"Hassan MCE"

}

{

"_id": ObjectId(7d178ad8903d),

"college_narne":"NIE"

}

{

"Objectld(7df78ad8904e).

"college_name":"SJCE"

}

5.5 MongoDB Datatypes

MongoDB supports many datatypes whose list is given below:

- String: This is most commonly used data type to store the data. String in mongodb must be UTF-8 valid.
- Integer: This type is used to store a numerical value. Integer can be 32 bit or 64 bit depending upon your server.
- Boolean: this type is used to store a boolean (true/ false) value.
- Double: This type is used to store floating point values.
- Min/ Max keys: This type is used to compare a value against the lowest and highest BSON elements.
- Arrays: This type is used to store arrays or list or multiple values into one key.
- Timestamp: timestamp. This can be handy for recording when a document has been modified or added.
- Object: This data type is used for embedded documents.
- Null: This type is used to store a Null value.
- Symbol: This data type is used identically to a string however, it's generally reserved for languages that use a specific symbol type.

- *Date:* This data type is used to store the current date or time in UNIX time format. You can specify your own date
- time: by creating object of pate and passing day, month, year into it.
- Object ID: This datatype is used to store the document's Ill.
- Binary data: This datatype is used to store bina'y data. :
- Code: This datatype is used to store javascript code into document.
- Regular expression: This datatype is used to store regular expression.

6. MONGODB AND BIG DATA - IS IT ANALYTICALLY POSSIBLE?

Let us start with an example.

Let us see if we can insert a blog post or a document data into MongoDB[9]

```
>db.post.insert([
{
title: 'MongoDB Overview',
description: 'MongoDB is no sql database', by: 'tutorials point',
url: 'http://www.tutorialspoint.corn;
tags: ['mongodb'., 'database', 'NoSQL', likes: 100
{
title: 'NoSOI. Database',
description: 'NoSQL database doesn't have tables',
by: 'tutorials point',
url: http://www.tutorialspoint.com,
tags: ['mongodb', 'database', 'NoSQL'],
likes: 20,
comments:[
{
user:'userl',
message: 'My first comment',
dateCreated: new Date(2013,11,10,2,35),
like: 0
}
]
}
])
```

But was this possible with a traditional RDBMS? Physically and virtually not as easy as this one.

6.1 The Gartner's 3 Vs and can MongoDB solve it?

Just for a quick recall — the Gartner 3Vs are Volume, Variety and Value We shall start with **Variety** -

6.2 MongoDB — Document Oriented and Schema free data

Take a look at Figure – 1:

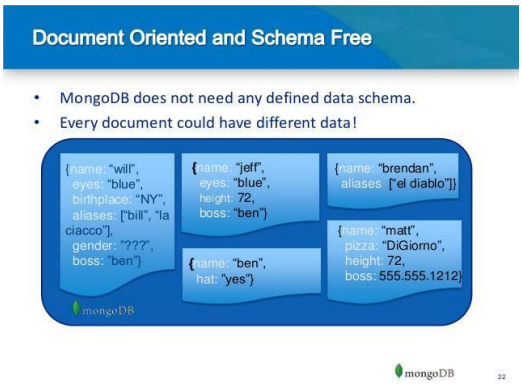

Fig 1: *MongoDB document oriented and schema free data [10]*

It clearly mentions that each document could have different types of data. So the problem of Variety is solved. *Not just* different types of data but even various data types are available in MongoDB for solving the problem of Variety.

6.3 Volume — is one of the biggest problems of Big Data. So how does MongoDB solve it?

Recall that traditional RDBMS could not be scaled vertically. NoSQL databases had this feature. MongoDB is one such database.

MongoDB — Characteristics to solve the problem of 'Volume of Big Data:

1. Provides atomic document operations.

2. Supports secondary indexes alongside hashing concept.

3. A concept called Sharding — also called partitioning for both vertical and horizontal

4. Replication of data across clusters distributed over a common or distributed network systems.

6.4 MongoDB — Aggregates, statistics and Analytics for Big Data

Supports both direct and native Map/Reduce concepts. MongoDB also has an aggregation framework It has support for Hadoop.

6.5 MongoDB Hadoop support

MongoDB can intersect its indexing with the brute force parallelization of Hadoop

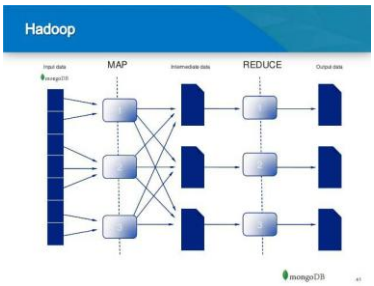

Fig 2: *shows usage of the Hadoop MongoDB connector*

Using the concept of Hadoop — data that is spread across multiple clusters can be accessed and processed for data analytics. Also MongoDB sharding provides scalability functionalities.

Thus the biggest problem of volume is solved to some extent. See Future work.

6.6 Lastly value for both money and value of data for using MongoDB.

Data is anytime precious. As per our previous problem statement who *are the* users using AXIS bank credit card ending with some number — what if there is some fraud transaction? Is storing data valuable? Yes indeed.

Creating applications using analytics for value of data is something that cannot be covered with view point of MongoDB. But yes of course does it make sense to use a database like MongoDB? How far is it better than other NoSQL databases or traditional databases for deployment purposes? See Fig 3:

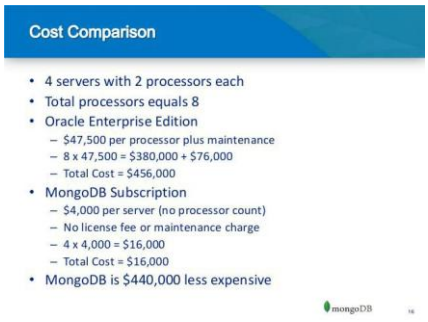

Fig 3: *Cost comparison between Oracle Enterprise Edition v/s MongoDB subscriptions*

The image clearly compares a traditional and also partially NoSQL Oracle Enterprise Edition and MongoDB subscription for a year. MongoDB is extremely inexpensive as compared to Oracle version.

Also the community edition of MongoDB allows a user to work with products upto an extent where he can manage things on his own. If in case — a support is required only then he may opt for a MongoDB subscription. This is the beauty with respect to value for Money. Managing a user's precious data — but when its free — is it something not interesting?

6.7 MongoDB and third party integrations:

MongoDB has 3rd party integration functionality with various data science tools like Apache Storm, Apache Spark, Apache Kafka, with regular SQL apart from the Hadoop functionality that was mentioned earlier[11]

7. CONCLUSION

This paper was able to partially solve the 3Vs of Gartner with respect to Big Data using a NoSQL. Database MongoDB. MongoDB is a freeware, has extremely flexible options, is community bound and is a product that is being asked as a must know in the world of Big Data. Also MongoDB could integrate with many third party sources for creating an environment that solves most of the Big Data issues.

8. FUTURE WORK

There were few parts in this research where the solution to the problem statement is incomplete. One such example is using Map/Reduce with MongoDB. The solution is not fully derived from our side. The content that MongoDB can handle is something huge. But we were able to process data that is both simple and many a times in text nature. All different kinds of data types need to be ventured upon. The concept of sharding could have been more clearly explained.

Also MongoDB concurrently works with data that is both distributed and real lime. But the real time analytics integration - with current tools like Apache Sparka and Apache Storm alongside Apache Kafka is — a place that is not much discussed, Th e work is still in progress[12].

ACKNOWLEDGMENT

This work was undertaken from the Department of PGSCEA and Department of ISE, NIE, Mysore. We are grateful to the Principal and Management of NIE for supporting this work.

REFERENCES

[1] http://wikibon.org/blog/big-data-statistics/

[2] http://www.bbc.com/news/business-2638305 8

[3] http://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data

[4] http://www.sas.com/resources/whitepaper/wp_46345.pdf

[5] http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

[6] http://www.researchgate.net/post/What_is_the_difference_between_RDBMS_and_NoSQL 2

[7] http://dba.stackexchange.com/questions/5/what-are-the-differences-between-nosql-and-a-traditional-rdbms

[8] http://nosql-database.org/

[9] http://tutorialpoints.com

[10] http://slideshare.com/slide/mongo_db_bigdata.634521

[11] Learning Apache Kafka 2nd Ed, Nishant Garg, Packt Publishers

[12] Big Data, Mining and Analytics, Stephan Kudyba, Auerbach Publications, CRC Press.