Prioritized Property-Value based Data Modelling for Big Data

Sai Jyothi Bolla, S. Jyothi, Y. Rajesh

Abstract:- ***In the present era, as technology is emerging widely data storage is also increasing its volume or space of storage enormously; which is the current buzz defined as Big Data. Existing Big Data modelling includes mostly in handling structured data but no defined approach was designed for modelling Big Data which includes structured, semi-structured and unstructured data. Among the existing challenges on Big Data, the most imperative challenge is modelling Big Data. This paper proposes a generic modelling approach for modelling Big Data. The effectiveness of this innovative approach is sensed by modelling oncology data using MongoDB. This modelling facilitates ease analytics and is independent of context.***

Keywords: Big Data, Data Analytics, Data Modelling, MongoDB

I. INTRODUCTION

The era of Data namely Data Science is emerging enormously in all technical and scientific sectors. Instances of various sources of such data like Facebook, Wikipedia, Google, etc. includes information of variable types like text, images, videos etc. Over decades with the passage of time volume of Big Data is increasing enormously; with an average increase of 2.5 quintillion bytes every day [1]. In a survey done in 2012, concluded that 90% of Big Data produced today was generated in the last couple of years, this volume of data ranged from few terabytes to petabytes [1]. This drastic increase in the volume of data confronts several challenges, among those, the major are storage capacity, big data modelling and analytics [13]. Among these Big Data Modelling plays an inevitable role in the era of Big Data; because if Big Data is modelled storage structures in the physical medium can be easily designed and analytics can be performed with less effort. Using traditional data modelling methods big data cannot be handled, because Big Data comprises of several heterogeneous data which may be structured, semi-structured and unstructured [3]. This intense of quick increase of Big Data with a wide range and varieties of data has created several ambiguities in Big Data Modeling before performing analytics. A finest Big Data Model will intensify the empathy of humanoids socially and economically for the best excellence.

The following section includes literature review which infers the evolution history of Big Data Modelling and the key route to our proposed modelling. The next section is followed by the detailed review of proposed data modelling and its implementation. This is continued with strengths of proposed modelling and reasons for considering document-based data modelling in the proposed model. The last section includes the conclusion and future enhancements.

II. LITERATURE REVIEW

Big Data is expanding exponentially [6] from various resources [4] contains diverse types of information [5]. To perform analytics on Big Data, data in Big Data must be organized and this organization needs the best processing methods. Most of the data processing methods earlier were legacy [7]. Cloud computing for few years has suggested the solution for computing over such Big Data but couldn't be independent of bulk data which is stressed at a node in the network with breakdowns [8]. To overcome this Big Data must be organized semantically using a data-driven approach which is essential for better analytics in all ecosystems [9].

The traditional data modelling approaches are not feasible for modelling Big Data. This is due to the existence of non-relational data of multiple types with various formats which don't fit into traditional data models. This is because as data in big data is evolving drastically, modelling of big data cannot be structured. This resulted in getting deposited with structured data, semi-structured data and unstructured data in major percentage. This resulted in an evolution of "Non-Relational Databases" for modelling Big Data [14]. In 2007 Amazon was the first to use non-relational database notion for storing its business data [11]. Table 1 shows the approaches and perspective over different types of databases [12].

Table 1: Approaches And Perspectives Of Over Various Types Of Databases

Approaches	Operational	Decision Support	Big Data
Data Modelling Perspective	ER and Relational Models	Star Schema and OLAP Models	Key-Value, Document, Wide-Column and Graph
	RDBMS	DW	Big Data-Based Systems
Data Analytics Perspective	OLTP	OLAP	Multiple Classes (Batch-oriented processing, stream-processing, OLTP and Interactive ad-hoc queries)

Revised Manuscript Received on July 18, 2019.

Sai Jyothi Bolla, Department of Computer Science,Sri Padmavathi Mahila Viswavidyalayam,Tirupati, Andhrapradesh, India.(E-mail: saidilipyerram@gmail.com)

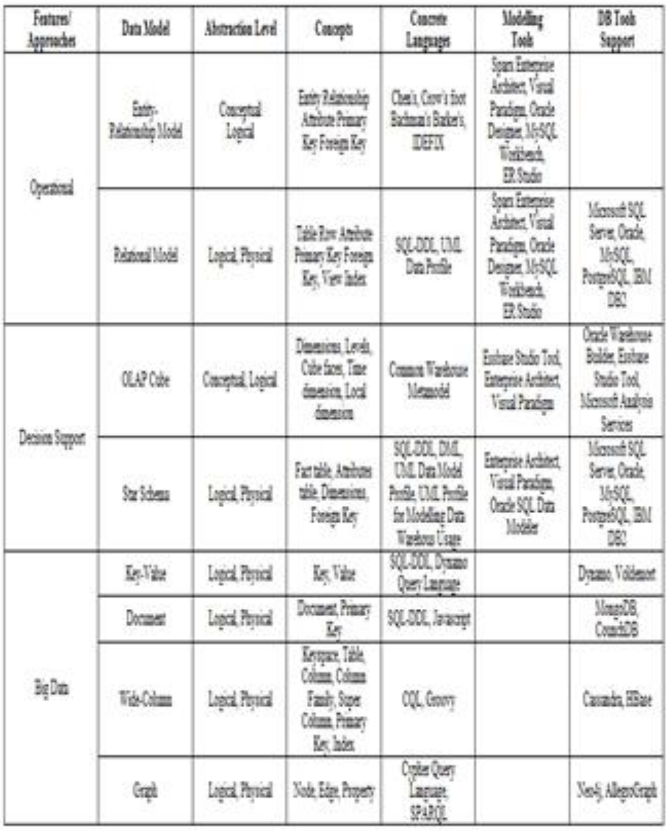
S. Jyothi, Department of Computer Science,Sri Padmavathi Mahila Viswavidyalayam, Tirupati, Andhrapradesh, India.(E-mail: Jyothi.spmvv@gmail.com)

Y. Rajesh, Department of Information Technology,Vasireddy Venkatadri Institute of Technology,Guntur, Andhrapradesh,India. (E-mail: rajesh.yelchuri@gmail.com)

Retrieval Number: I11450789S219/19©BEIESP
DOI : 10.35940/ijitee.I1145.0789S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 2 explores data modelling of operational databases, decision support systems and big data [12]. This reveals about abstraction levels, tools used for modelling data and the supporting tools for extracting worthy knowledge. This table also reveals that operational databases with few data models have designing tools for data modelling but doesn't have supporting tools for extracting knowledge from databases, for example, ER Model. Whereas decision support databases have supporting tools for extracting knowledge along with data modelling tools of various types. Table 2 explores a thought-provoking notion about Big Data modelling. This infers that most of the tools which are being used and using are supporting tools for extracting knowledge but there are no designing tools for modelling Big Data. There are NoSQL Data Models for Big Data like key-value, document, wide-column and graph-based [15]. These data models have both logical and physical representation of data which can be supported by many tools [12]. All supporting tools for big data are of type NoSQL. These tools never identify relations among objects.



Features/ Approaches	Data Model	Abstraction Level	Concepts	Concrete Languages	Modelling Tools	DB Tools Support
Operational	Entity-Relationship Model	Conceptual Logical	Entity Relationship Attribute Primary Key Foreign Key	Chen's, Crow's foot Bachman's Barker's, IDEFIX	Sparx Enterprise Architect, Visual Paradigm, Oracle Designer, MySQL Workbench, ER Studio	
	Relational Model	Logical, Physical	Table Row Attribute Primary Key Foreign Key, View Index	SQL-DDL, UML Data Profile	Sparx Enterprise Architect, Visual Paradigm, Oracle Designer, MySQL Workbench, ER Studio	Microsoft SQL Server, Oracle, MySQL, PostgreSQL, IBM DB2
Decision Support	OLAP Cube	Conceptual, Logical	Dimensions, Levels, Cube faces, Time dimension, Local dimension	Common Warehouse Metamodel	Essbase Studio Tool, Enterprise Architect, Visual Paradigm	Oracle Warehouse Builder, Essbase Studio Tool, Microsoft Analysis Services
	Star Schema	Logical, Physical	Fact table, Attributes table, Dimensions, Foreign Key	SQL-DDL, DML, UML Data Model Profile, UML Profile for Modelling Data Warehous Usage	Enterprise Architect, Visual Paradigm, Oracle SQL Data Modeler	Microsoft SQL Server, Oracle, MySQL, PostgreSQL, IBM DB2
Big Data	Key-Value	Logical, Physical	Key, Value	SQL-DDL, Dynamo Query Language		Dynamo, Voldemort
	Document	Logical, Physical	Document, Primary Key	SQL-DDL, Javascript		MongoDB, CouchDB
	Wide-Column	Logical, Physical	Keyspace, Table, Column, Column Family, Super Column, Primary Key, Index	CQL, Groovy		Cassandra, HBase
	Graph	Logical, Physical	Node, Edge, Property	Cypher Query Language, SPARQL		Neo4j, AllegroGraph

Table 2: Summary Of Data Modeling Over Various Approaches.

Figure 1.1 specifies the relationship of increase between the data size and the processing capacity available annually; the calculations behind this juxtaposition are based upon the truth presented by Hilbert and López, 2011 [10] in their published work in Science. It can be observed how state-of-the-art computation capacity is lagging behind in front of the speedily leading data size, which is available to discover worthless information patterns.

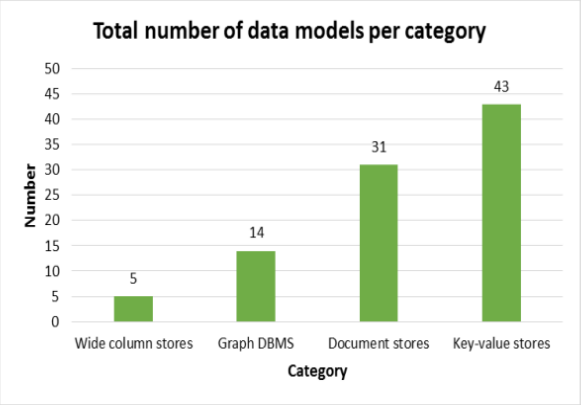


Figure 1.1: Total number of data models per category

As conflicting to their peers, graph data models exceedingly weigh relations and provides the visual illustration of information, Therefore , they are treated as more user-friendly. The graph data models are especially useful for such scenarios where the affairs in data have more weightage when compared to data itself . Such examples include forensic investigations conclusion , traversal and social network representation etc,.

Figure 1.1 disclose a number of data models which are available today for every individual category of the NoSQL, specified in this paper. NoSQL data models deals with the data at very considerably faster rate than relational data models, which are frequently utilized in profitable domains that need higher care for the dealings processed. So, the relational data models strictly obey ACID (atomicity, consistency, isolation, durability) constraints on each piece of data and it effects the speed of relational data models. A good number of the most important NoSQL models are flexible to be used to convene the requested needs and regularly don't pass through ACID conditions and keep away from all the technical inhabitant requirements that relational data models have, therefore better performance. In opposing, for those applications that need great precision, NoSQL data models may undergo overheads.

The rising and falling quantity of Big Data of Internet actions and sensory progressions, and the growing requirements for Big Models for ultra-high-dimensional problems put fabulous stress on Big Data Analytics. It is exceedingly unproductive; if possible, to utilize such big data consecutively in a group or scholastic manner in distinctive Analytics. As Big Data includes structured, semi-structured and unstructured data, a highly scalable and generalized approach for modelling Big Data is desirable.

III. PROPOSED MODEL

The proposed model is a Prioritized Property-Value based Data Model (PPVDM) for Big Data. This involves in scraping information from various resources of single context or single ecosystem. Not all resources are used for scraping information. Every echo system will have its own set of properties. Among these properties, only the properties which influence the behaviour of its ecosystem are considered. These properties are sorted according to their priorities. Priorities for these properties are finalized under the supervision of an expert in that ecosystem. After sorting such properties, values related to such properties are scraped from several resources of same echo system for several distinct objects. This results in a Semi-Structured representation of Big Data which is sparse.

At instance information of side effects of various cancer drugs, heterogeneous data [16] are scraped from cancer.gov, uihc.org, cancerresearchuk.org using Java JSOUP third-party API for prioritized properties for this ecosystem. Using DOM, CSS and Jquery methods of JSOUP information is scraped. The data found in PDFs are extracted by using PDFBOX third party API. After scraping the information is written into the comma separated file using JAVA JAVACSV third party API. Analytics done on this file is clumsy. This is due to; most of the values in the file will be missing. If these values are considered, analytics performed over such file will never result in the optimal extraction of analysis. To overcome this disadvantage a sparse representation of this extracted CSV file is exported into Prioritized Property-Value based documents, is included in the proposed Big Data model. This is done using MongoDB. MongoDB is a document store in which a collection holds different documents. Each document represents an object in the ecosystem. Every object in the echo system has its own properties. The number of properties, content and size of the object can differ from one object to another. All objects in a collection are the similar or related type. MongoDB does not enforce a schema for collections. It uses dynamic schema. This semantics is suitable for echo systems which comprise properties which are variable. These semantic considers properties which are non-null besides ignores properties with nulls; a sparse representation of data. Properties having multiple values are treated as multiple values under common property whereas in other implementing environments multiple values of a single property are devised into a single concatenated value which makes analytics legacy. The detailed block diagram is shown in Figure 4.1.

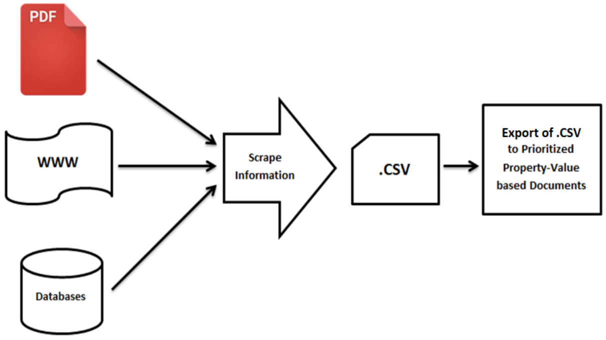
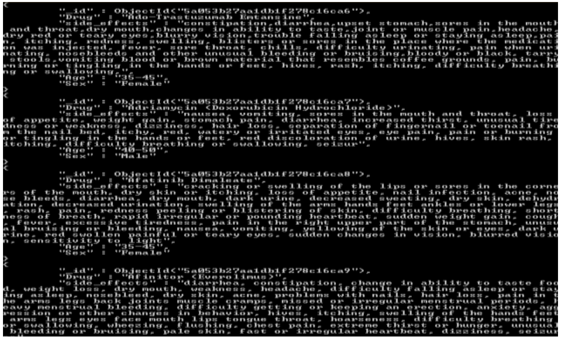

PDF
WWW
Databases
Scrape Information
.CSV
Export of .CSV to Prioritized Property-Value based Documents

Figure 4.1 Process of Property Value Data Model

The scope of this model building approach using various big data sources into one framework to provide Physician centric decision supportive system which enhances vast experience comes to our fingertips along with latest research innovations.

IV. IMPLEMENTATION & RESULTS

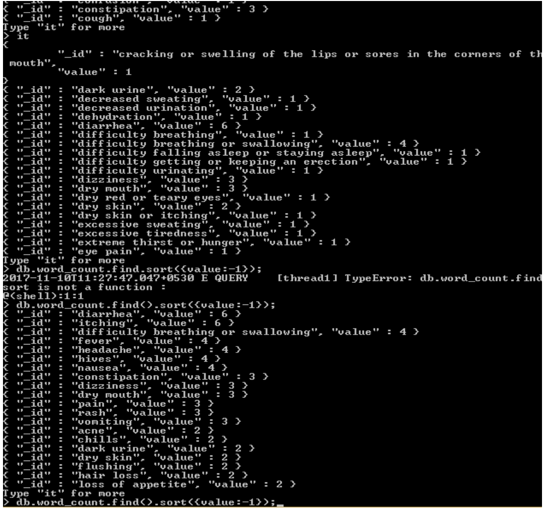
For implementing this Prioritized Property-Value based Data Modelling for Big Data, information about Cancer-related drugs which had side effects and no effects are scraped. Initially, information is extracted into .CSV. This extracted file is exported to property value based documents using MongoDB to convert the data into Flexible Semi-Structured Data Documents. The following screenshot Figure 4.2 reveals the representation of scraped information according to Priority Property-Value.



```
{
        "_id" : ObjectId("5a053b27aa1db1f278c16ca6"),
        "Drug" : "Ado-Trastuzumab Emtansine",
        "side_effects" : "constipation,diarrhea,upset stomach,sores in the mouth
 and throat,dry mouth,changes in ability to taste,joint or muscle pain,headache,
dry red or teary eyes,blurry vision,trouble falling asleep or staying asleep,pai
n, itching, redness, swelling, blisters or sores in the place where the medicati
on was injected, fever, sore throat, chills, difficulty urinating, pain when uri
nating, nosebleeds and other unusual bleeding or bruising,bloody or black, tarry
 stools,vomiting blood or brown material that resembles coffee grounds, pain, bu
rning or tingling in the hands or feet, hives, rash, itching, difficulty breathi
ng or swallowing,",
        "Age" : "35-45",
        "Sex" : "Female"
}
{
        "_id" : ObjectId("5a053b27aa1db1f278c16ca7"),
        "Drug" : "Adriamycin (Doxorubicin Hydrochloride)",
        "side_effects" : "nausea, vomiting, sores in the mouth and throat, loss
of appetite, weight gain, stomach pain, diarrhea, increased thirst, unusual tire
dness or weakness, dizziness, hair loss, separation of fingernail or toenail fro
m the nail bed, itchy, red, watery or irritated eyes, eye pain, pain or burning
or tingling in the hands or feet, red discoloration of urine, hives, skin rash,
itching, difficulty breathing or swallowing, seizur",
        "Age" : "40-50",
        "Sex" : "Male"
}
{
        "_id" : ObjectId("5a053b27aa1db1f278c16ca8"),
        "Drug" : "Afatinib Dimaleate",
        "side_effects" : "cracking or swelling of the lips or sores in the corne
rs of the mouth, dry skin or itching, loss of appetite, nail infection, acne, no
se bleeds, diarrhea, dry mouth, dark urine, decreased sweating, dry skin, dehydr
ation, decreased urination, swelling of the arms hands feet ankles or lower legs
, rash, pain, redness peeling or blistering of skin, difficulty breathing, short
ness of breath, rapid irregular or pounding heartbeat, sudden weight gain, cough
, fever, excessive tiredness, pain in the right upper part of the stomach, unusu
al bruising or bleeding, nausea, vomiting, yellowing of the skin or eyes, dark u
rine, red swollen painful or teary eyes, sudden changes in vision, blurred visio
n, sensitivity to light",
        "Age" : "35-45",
        "Sex" : "Female"
}
{
        "_id" : ObjectId("5a053b27aa1db1f278c16ca9"),
        "Drug" : "Afinitor (Everolimus)",
        "side_effects" : "diarrhea, constipation, change in ability to taste foo
d, weight loss, dry mouth, weakness, headache, difficulty falling asleep or stay
ing asleep, nosebleed, dry skin, acne, problems with nails, hair loss, pain in t
he arms legs back joints muscle cramps, missed or irregular menstrual periods, h
eavy menstrual bleeding, difficulty getting or keeping an erection, anxiety, agg
ression or other changes in behavior, hives, itching, swelling of the hands feet
 arms legs eyes face mouth lips tongue throat, hoarseness, difficulty breathing
or swallowing, wheezing, flushing, chest pain, extreme thirst or hunger, unusual
 bleeding or bruising, pale skin, fast or irregular heartbeat, dizziness, seizur
```

Figure 4.2 Result of scraped data cropped into MongoDB

In supportive of this proposed Big Data Modelling an experiment is programmed in extracting knowledge about the most prominent side effects which is recurrent for most of the cancer drugs. The following screenshot Figure 4.3 reveals about this experiment.



```
{ "_id" : "constipation", "value" : 3 }
{ "_id" : "cough", "value" : 1 }
Type "it" for more
> it
{
        "_id" : "cracking or swelling of the lips or sores in the corners of th
 mouth",
        "value" : 1
}
{ "_id" : "dark urine", "value" : 2 }
{ "_id" : "decreased sweating", "value" : 1 }
{ "_id" : "decreased urination", "value" : 1 }
{ "_id" : "dehydration", "value" : 1 }
{ "_id" : "diarrhea", "value" : 6 }
{ "_id" : "difficulty breathing", "value" : 1 }
{ "_id" : "difficulty breathing or swallowing", "value" : 4 }
{ "_id" : "difficulty falling asleep or staying asleep", "value" : 1 }
{ "_id" : "difficulty getting or keeping an erection", "value" : 1 }
{ "_id" : "difficulty urinating", "value" : 1 }
{ "_id" : "dizziness", "value" : 3 }
{ "_id" : "dry mouth", "value" : 3 }
{ "_id" : "dry red or teary eyes", "value" : 1 }
{ "_id" : "dry skin", "value" : 2 }
{ "_id" : "dry skin or itching", "value" : 1 }
{ "_id" : "excessive sweating", "value" : 1 }
{ "_id" : "excessive tiredness", "value" : 1 }
{ "_id" : "extreme thirst or hunger", "value" : 1 }
{ "_id" : "eye pain", "value" : 1 }
Type "it" for more
> db.word_count.find.sort({value:-1});
2017-11-10T11:27:47.047+0530 E QUERY    [thread1] TypeError: db.word_count.find
sort is not a function :
@(shell):1:1
> db.word_count.find().sort({value:-1});
{ "_id" : "diarrhea", "value" : 6 }
{ "_id" : "itching", "value" : 6 }
{ "_id" : "difficulty breathing or swallowing", "value" : 4 }
{ "_id" : "fever", "value" : 4 }
{ "_id" : "headache", "value" : 4 }
{ "_id" : "hives", "value" : 4 }
{ "_id" : "nausea", "value" : 4 }
{ "_id" : "constipation", "value" : 3 }
{ "_id" : "dizziness", "value" : 3 }
{ "_id" : "dry mouth", "value" : 3 }
{ "_id" : "pain", "value" : 3 }
{ "_id" : "rash", "value" : 3 }
{ "_id" : "vomiting", "value" : 3 }
{ "_id" : "acne", "value" : 2 }
{ "_id" : "chills", "value" : 2 }
{ "_id" : "dark urine", "value" : 2 }
{ "_id" : "dry skin", "value" : 2 }
{ "_id" : "flushing", "value" : 2 }
{ "_id" : "hair loss", "value" : 2 }
{ "_id" : "loss of appetite", "value" : 2 }
Type "it" for more
> db.word_count.find().sort({value:-1});_
```

Figure 4.3 Result of most frequent side effects among cancer drugs

Pseudo code used for performing the above analytics is:

```
conn = new Mongo();
db = conn.getDB("users");
var map = function() {
    var summary = this.side_effects;
    if (summary) {
        summary = summary.toLowerCase().split(",");
        for (var i = summary.length - 1; i >= 0; i--) {
            if (summary[i])  {
              emit(summary[i].trim(), 1);
            }
        }
    }
};
var reduce = function( key, values ) {
    var count = 0;
    values.forEach(function(v) {
        count +=v;
    });
    return count;
}
db.drugs.mapReduce(map, reduce, {out: "word_count"})
```

A. Reasons for selecting MongoDB for Prioritized Property-Value based Data Modelling

The main reason for selecting MongoDB for modelling Big Data in the proposed model:

a. MongoDB loads .CSV files as documents. Each document contains only those properties which are non-null.

b. Sparse representation of data

c. This paper pins MongoDB to be the optimal database for proposed work based on its performance when compared with other popular databases like DynamoDB, Couchbase, CouchDB, RethinkDB and OrientDB. This performance evaluation is done on (i) data insertion and (ii) data retrieval

(i) Data Insertion

MongoDB proves to be the optimal database for inserting data, as it takes less amount of time shown in Figure 4.4.

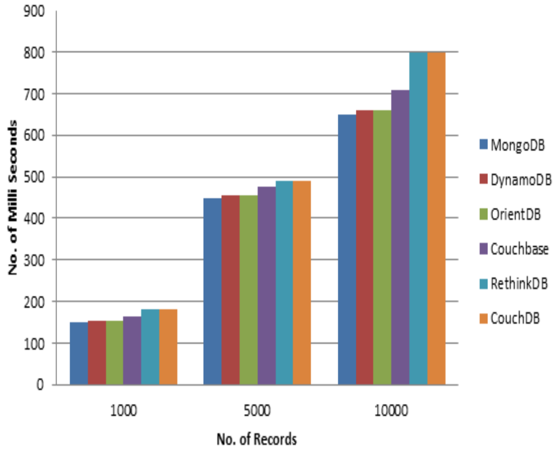


Figure 4.4: Insertion Time

(ii) Data Retrieval

MongoDB takes less amount of time for retrieving data when compared with other databases shown in Figure 4.5.

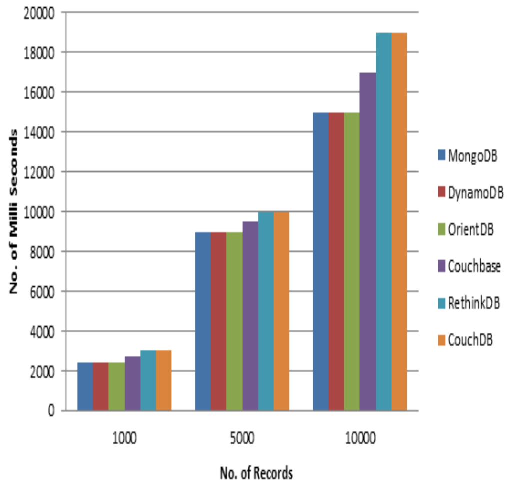


Figure 4.5: Retrieval Time

B. Strengths of Proposed Model

a. The biggest strength of this proposed model is flexible.

b. Most of the data modelling methods for Big Data includes null values. The proposed model doesn't consider null values. This is due to integrated involvement of MongoDB in proposed work.

c. Most of the data modelling methods for Big Data provides a dense representation of data because null values are considered but not handled before performing analytics. The proposed modelling doesn’t include null values in data representation which is a sparse representation; this reduces the effort for performing analytics. This is because of two reasons: (1) Prioritized Property-Value Document representation of Data (2) Excluding null values before modelling data. This doesn't effect Data Analytics.

d. Performing analytics over this model takes less effort when compared to analytics which is row organized and dense.

V. CONCLUSION

Big Data is bulk which includes structured, semi-structured and unstructured data. Handling Big Data for performing analytics over it is the most challenging chore which relies on modelling Big Data. While comparing Data Analytics and Data Modelling; Data Modelling plays a vital role in Data Analytics. The more flexible is the Big Data Modelling results in the most effective Data Analytics. The proposed Big Data Modelling is flexible because the proposed model is Prioritized Property-Value Data Modelling. This proposed modelling can be used in any echo systems and can perform analytics over it with less effort.

VI. REFERENCES

1. . IBM, 2012. *What is Big Data ? Bringing Big Data to the enterprise.* [Online] http://www-01.ibm.com/software/data/bigdata/

2. Furner, J. (2003). Little Book, Big Book: Before and After Little Science, Big Science: A Review Article, Part I. Journal of Librarianship and Information Science 35 (2): 115–125. doi:10.1177/0961000603352006. Retrieved 2014-02-09.
3. Weiss, R., Zgorski, L. J. (2012*). Obama Administration Unveils "BIG DATA" Initiative: Announces $200 Million In New R&D Investments*. [Online] http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf
4. Villars, R. L., Olofson, C. W. and Eastwood, M. (2011). Big Data: What it is and why you should care. *IDC White Paper*. Framingham, MA: IDC.
5. Robert, H. (2012). It's time for a new definition of Big Data. *MIKE2.0: The open source standard for Information Management*. [Online] http://mike2.openmethodology.org/
6. Watters, A. (2010). The Age of Exabytes: Tools and Approaches for Managing Big Data (Website/Slideshare). *Hewlett-Packard Development Company*.
7. Carino, F., and Sterling, W. M. (1998). Method and apparatus for extending existing database management system for new data types. U.S. Patent No. 5,794,250.
8. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. Nature Rev. Genet. 11, 647–657.
9. Bell, G., Hey, T., Szalay,A. (2009). Beyond the data deluge, Science 323 (5919), 1297–1298.
10. Hilbert, M., López, P. (2011). The world's technological capacity to store, communicate, and compute information, Science 332 (6025), 60–65.
11. Leavitt, N. (2010). Will NoSQL databases live up to their promise?. Computer, 43(2), 12-1
12. André Ribeiro, Afonso Silva, Alberto Rodrigues da Silva, "Data Modeling and Data Analytics: A Survey from a Big Data Perspective", Journal of Software Engineering and Applications, 2015, 8, 617-634 Published Online December 2015 in SciRes. http://www.scirp.org/journal/jsea http://dx.doi.org/10.4236/jsea.2015.812058
13. Idrees, S.M., Alam, M.A. & Agarwal, P. Int. j. inf. tecnol. (2018). https://doi.org/10.1007/s41870-018-0185-1
14. Karla Saur, Tudor Dumitras¸ and Michael Hicks, "Evolving NoSQL Databases Without Downtime", arXiv:1506.08800v3 [cs.DB] 25 Apr 2016.
15. 15.https://pdfs.semanticscholar.org/773e/9e98d42f395864baecf6e87a9c7ded1f36e6.pdf
16. Baloch, Z., Shaikh, F.K. & Unar, M.A. Int. j. inf. tecnol. (2018) 10: 241. https://doi.org/10.1007/s41870-018-0116-1