Review of NoSQL Databases MongoDB, CassendraDB, DynamoDB, CoughDB.

Patel Pinanki Bharat,

Department of Computer Science and Technology

Lovely Professional University, Phagwara.

patelpinanki96@gmail.com

Asst. Prof. Sarvesh Chopra,

Department of Computer Science and Technology

Lovely Professional University, Phagwara.

Abstract- In the modern edge world framework, customers or web customers are extending exceptionally well ordered by virtue of that a consistently expanding number of unstructured data's are conveying and eating up over the framework. Moreover, how to keep up those data and improve the openness and flexibility of the limit system transforms into a noteworthy test. Nowadays a part of the NoSQL databases are supported the unstructured data organization and give assorted purposes important to the unstructured data organization e.g. CassandraDB, CoughDB, MongoDB, DynamoDB, etc. MongoDB that is giving the most versatile request capacities with respect to the unstructured data organization stood out from exchange databases like DynamoDB, CassandraDB. The objective is to store a considerable measure of unstructured data into the MongoDB.

Keywords- NoSQL, MongoDB, Unstructured Data, Framework.

1. INDRODUCTION

1.1 Unstructured Data

Unstructured information is a general mark to depict the database that isn't involved in any database or various sort of data structure. Unstructured data can be printed or non-peruser. Unstructured data is created in media like email messages, PowerPoint introductions, Word records, joint effort programming and texts. Unstructured data is produced in media, for example, JPEG pictures, MP3 sound documents, and blaze video records.

1.2 Structured Data

Structured data is information or data that has been composed within a designed store, generally a database, to achieve that its components can be made addressable for more compelling preparing and examination. Structured data alludes to any data that lives in a most levelled inside a files or documents. This incorporates data consist in social database and spreadsheets. Structured data is regularly overseen concerned with Structured Query Language (SQL) – a programming language made for overseeing and questioning data in social database administration frameworks.

Structured Query Language (SQL) privilege inquiries on this sort of organized information inside social databases. The word SD Originated from the name of a typical dialect used to get to DBs, which is called organized question dialect or SQL.

1.3 Structured Data Vs Unstructured data

Fig 1.2: Structured Data Vs Unstructured Data.

- Structured Data is effectively accessible by essential Algorithms. Models incorporate spreadsheets and information from machine sensors.
- Unstructured Data is progressively similar to human dialect. It doesn't fit pleasantly into social databases like SQL, and seeking it dependent on the old algorithms ranges from hard to totally unimaginable.

- Unstructured information (or unstructured data) is data that either does not have a pre-characterized information display or isn't sorted out in a pre-characterized way. Unstructured data is regularly message overwhelming, however may contain information, for example, dates, numbers, and certainties too.

Table 1.1: Comparison between Structured Data and Unstructured data.

	Structured Data	Unstructured Data
Characteristics	1. Pre-defined Data Models 2. Text only 3. Easy to find	1. No Pre-defined Data Model 2. Text, Images, Sound, Video 3. Difficult to find
Resides In	1. Relational Database 2. Data Warehouse	1. Applications 2. NoSQL Database 3. Data Lakes 4. Data Warehouse
Generated By	Humans or Machines	Humans or Machines
Typical Applications	1. Inventory Control 2. ERP System 3. CRM System 4. Airline Reservation System	1. Email Clients 2. Tools for Viewing or Editing Media 3. Word Processing 4. Presentation Software
Examples	1. Dates 2. Phone Numbers 3. Credit Card Number 4. Addresses 5. Transaction Information 6. Customer names	1. Text Files 2. Reports 3. Email Messages 4. Audio Files 5. Video Files 6. Images 7. Surveillance Images

2. Hadoop

Hadoop is the best approach for associations that would prefer not to add burden to their essential stockpiling framework and need to compose conveyed occupations that perform well. MongoDB NoSQL database is utilized in the huge information stack for putting away and recovering one thing at once from extensive datasets while Hadoop is utilized for preparing these expansive informational collections. For associations to keep the heap off MongoDB in the generation database, information preparing is offloaded to Apache Hadoop. Hadoop gives higher request of extent and power for information handling.

2.1 MongoDB connector for Hadoop

- The MongoDB-Hadoop connector utilizes the expert of Hadoop's MapReduce to live application information in MongoDB by extricating values from Big Data – quickly just as productively.
- The MongoDB-Hadoop connector ventures it as 'Hadoop compatible file system' and MapReduce jobs would now be able to be perused specifically from MongoDB, without being replicated to the HDFS. Thus, getting rid of the need of exchanging terabytes of information over the system.
- The "need" of checking whole accumulations has been dispensed with as MapReduce jobs can pass inquiries by methods for channels and can tackle MongoDB's ordering capacities like content hunt, compound, exhibit, Geo-spatial and sparse indexes.
- Perusing and composing back outcomes from Hadoop jobs back to MongoDB so as to help questions and constant operational procedures.

2.2 Application of Hadoop and MongoDB

1. MongoDB is utilized for the operational part – as a continuous information store.
2. Hadoop is utilized essentially for disconnected investigation and preparing of clump information.

3. MongoDB

MongoDB is an open-source document database and leading NoSQL database. MongoDB ideas expected to make and convey an exceptionally adaptable and execution arranged database. MongoDB is an archive database with the versatility and adaptability that need with the questioning and ordering that require. MongoDB is a cross-stage, report arranged database that gives, superior, high accessibility, and simple adaptability. MongoDB takes a shot at idea of accumulation and record. Mongo DB is one of a few database types to emerge in the mid-

2000s under the NoSQL pennant. Rather than utilizing tables and lines as a social database, MongoDB is based on a design of gathering and reports. Like other NoSQL databases, MongoDB underpins dynamic outline plan, from which reports get diverse fields and structures in the gathering. The database utilizes an archive stockpiling and information exchange organize called BSON, which gives a parallel portrayal of records, for example, JSON.

a. MongoDB Architecture

The Database: In straightforward words it very well may be called as the physical compartment for information. Every one of the databases has its very own arrangement of documents on the record framework with numerous databases existing on a solitary MongoDB server.

The Collection: A gathering of database records can be called as an accumulation. What could be compared to the accumulation is a table. The whole gathering exists inside a solitary database. There are no constructions with regards to accumulations. Inside the gathering the different archives can have fluctuated fields yet for the most part the reports inside an accumulation are intended for a similar reason or serving a similar end goal.

The Document: An arrangement of key-esteem sets can be assigned as a record. Records are related with dynamic mappings. The advantage of having dynamic blueprints is that report in a solitary gathering does not must have a similar structure or fields. Additionally, the regular fields in a gathering's report can have fluctuated kinds of information.
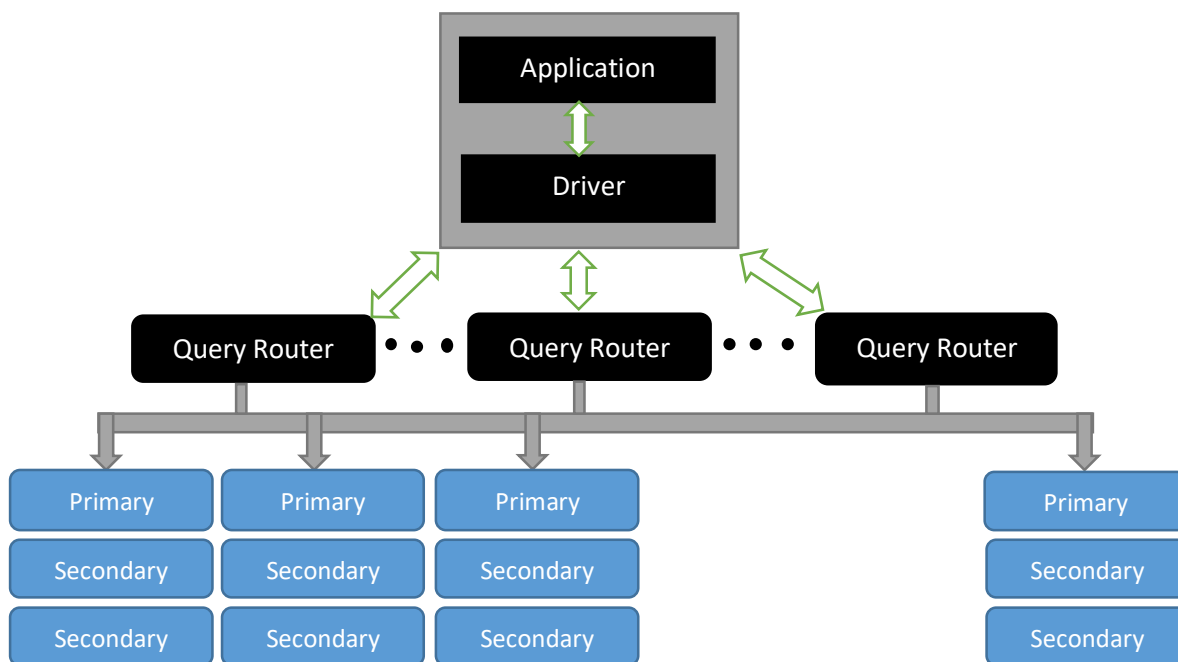

Fig. 3.1: Architecture of MongoDB.

Config Server: Config Server stores the metadata for sharded pack, they are the exceptional mongod events. Note that config server uses arrangement of two phase submit with the ultimate objective to ensure:

☐ Prompt Consistency

☐ Dependability

Shards: At least, 2 "replication set" are known as "shards". MongoDB is a high available in a way that it would constantly be open for taking care of; emphatically no loss of data and no loss of server to trigger assignment. Replication presents overabundance and besides constructs data availability. With having various copies of data present on discrete database servers, replication certainly shields database from loss of single server. It is like manner empowers us to recover from the organization interruptions and gear disillusionment. With these additional copies of data, can submit one for disaster recovery or support.

Query Routers: They are at least one mongos example. The mongos occurrence is the switch for group. Normally you will discover organizations has one mongos example on every application server.

b. Sharding

There are two techniques for tending to framework development:

☐ Vertical scaling.

☐ Horizontal scaling.

Vertical Scaling includes expanding limit of solitary server, for example, utilizing an all the more incredible CPU, including more RAM, or expanding the measure of storage room. Impediments in accessible innovation may confine a solitary machine from being adequately incredible for given remaining burden. Also, Cloud-based suppliers having hard roofs dependent on accessible equipment setups. Accordingly, there is a pragmatic most extreme for vertical scaling.

Horizontal Scaling includes partitioning framework dataset and stack over different servers, adds extra servers to expand limit as required. While, the general speed or limit of a solitary machine not be high, every machine handles subset of general remaining burden, conceivably giving preferred effectiveness over a solitary fast high-limit server. Extending the limit of the sending just requires including extra servers as required, which can be a lower by and large expense than top of the line equipment for a solitary machine. The exchange off is expanded intricacy in framework and upkeep for the sending.

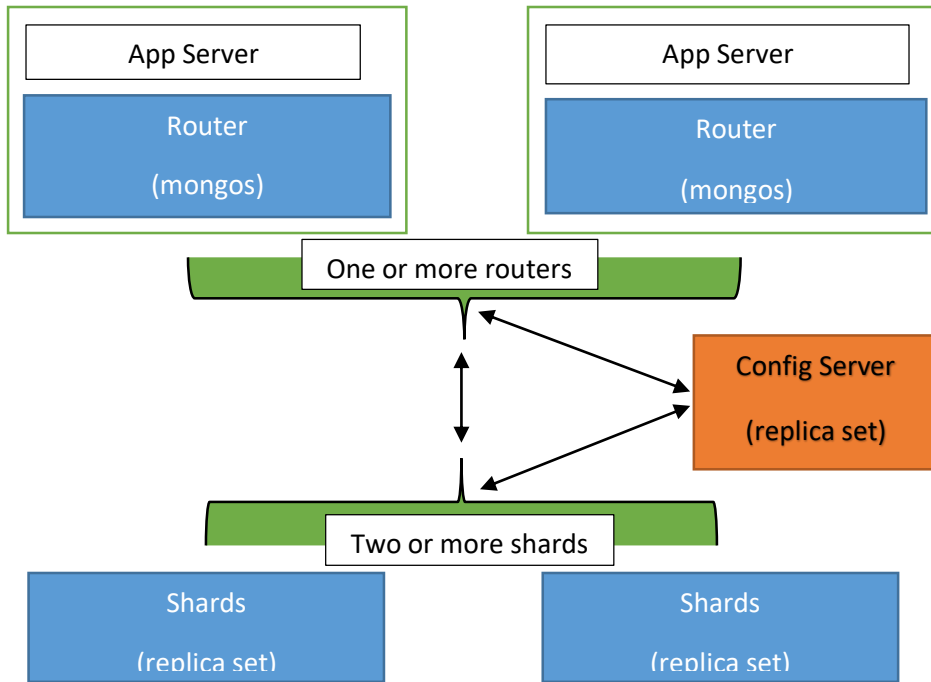MongoDB supports horizontal scaling through sharding.

Fig. 3.2: Sharding in MongoDB.

MongoDB shards information at the gathering level, conveying the accumulation information over the shards in the bunch.

4. LITERATURE SURVEY

Shagufta Praveen, et.al, (2017), discusses various types of data with their aspects with examples, and also represent that the growing data is responsible for the numerous emerging data models and database evolution. Different types of data available in the digital world require separate data models for their storage, processing and analysis. Structured Data are easily entered, stored and analysed. Structured Data is stored in the form of rows and columns which is easily managed with a language called "Structured Query Language (SQL)". Google search is done with the help of structured data. Since starting of the revolution of database network, hierarchical, relational, object relational data model deal with structured data. According to author, the unstructured data is not fir for relational database and in order to make them store, scenario came up with NoSQL database. Today, there are 4 family of NoSQL database: Key-value, Column-oriented, Graph-oriented, and Document-oriented. The author emphasize on the idea that developing information straightforwardly impact its related information models and database advances, it speaks to that huge information idea manages gigantic and huge information as well as it gives another entryway to database examiner and

analysts to deal with different information and information models for survival of new sorts of information in up and coming and present situation.

Cornelia GYORODI, et.al. (2015), they attempt to show a similar investigation of non-social databases and social databases. They chiefly concentrate their introduction on one execution of the NoSQL database innovation, to be specific MongoDB, and make a examination with another execution of social databases, to be specific MySQL. They additionally present the upsides of utilizing a non-social database contrasted with a social database, coordinated in a discussion in the field of individual and proficient advancement. The NoSQL database used to create discussion is MongoDB.

Rajith Kumar S, et.al. (2015), they focuses on examination of Cassandra, MongoDB and HBase which are the most usually utilized NoSQL databases. This examination between NoSQL databases conveys them on yahoo cloud stage which utilizes diverse kinds of virtual machines and group sizes to ponder the impact of various designs. The last outcome demonstrates the execution of databases at various outstanding task at hand levels and the outcome can be contrasted with discover the best among these three databases. This correlation additionally enables clients to pick the most proper database as indicated by their requirements.

Mrs. Anuradha Kanade, et.al. (2015), in this paper they have connected two distinctive demonstrating styles as installing of reports and standardization on accumulations. With the installing include they may confront circumstance where reports develop in size after creation which may corrupt the execution of database. The greatest report measure permitted in MongoDB is restricted. With references they get greatest adaptability than installing however customer side applications must issue follow-up questions to determine the references. The participates for this situation can't be successfully utilized. Consequently, there is requirement for characterizing the methodology of degree of standardization and implanting to show signs of improvement execution in the blended circumstance. The paper examined here demonstrates the variety in the execution alongside the adjustment in the demonstrating style with reference to standardization and installing and it gives the base to discover the degree of standardization and installing for diminishing question execution time.

Sunita Ghotiya, et.al. (2017), this paper provides a writing audit of a portion of the ongoing methodologies proposed by different specialists to move information from social to NoSQL databases. A few analysts proposed components for the concurrence of NoSQL what's more, Relational databases together. This paper gives an outline of systems which can be utilized for

mapping information put away in Relational databases to NoSQL databases. Different strategies for information change and centre layer arrangements are outlined in the paper.

Saran Raj, (2015), the author successfully stored large number of unstructured data in Mongo DB, using continuous hashing algorithms. In this work, various types of unstructured data available in the Scientific Information Resource Division (SIRD) of the Indira Gandhi Centre for Atomic Research (IGCAR) have been shown continuously in MongoDB using the hashing algorithms. The author able handled with the 100GB of data that is available in SIRD, IGCAR.

Harpinder Kaur, et.al, (2015), the author improved the load balancing technique for MongoDB clusters. The load-balancing of servers is vital for storage applications that are mostly read intensive. According to the authors, Traditional balancing methods cannot be trusted for the distributed environment. Then an efficient solution will be developed to balance load on distributed MongoDB groups and ultimately increase their performance in large volumes of loading. The results will be verified by the algorithm implementation. Algorithms will initially monitor all the shards whether they are balanced or loaded or overloaded. Then if a shard is monitored for the overloaded, then according to the revised version of the algorithm, the load is redistributed until the load is balanced.

5. CONCLUSION

Structured and unstructured data are different. Despite their differences, they work in tandem in any effective big data operation. MongoDB enhance the high accessibility, and high adaptability to the information stockpiling. Reliable hashing calculation is chiefly used to store the with low information misfortune. Also, the steady hashing calculation is utilized to do the grouping inside into the MongoDB database. Utilizing the GirdFS innovation the information's ought to be isolated into the shard that could be part into the numerous shards. Utilizing the reliable hashing calculation, the information's are bunched so the recovering will be quick.

6. References

[1]   S. Raj, "Storing of Unstructured data into MongoDB using Consistent Hashing Algorithm," no. January, 2015.

[2]   S. Praveen and U. Chandra, "Influence of Structured, Semi-Structured, Unstructured data on various data models," *Int. J. Isc. Eng. Res.*, vol. 8, no. 12, pp. 67–69, 2017.

[3]   V. Jain, "MongoDB and NoSQL Databases," *Int. J. Comput. Appl.*, vol. 167, no. 10, pp. 975–8887, 2017.

[4]   S. S. Pore and S. B. Pawar, "Comparative Study of SQL & NoSQL Databases," *Ijarcet*, vol. 4, no. 5, pp. 1747–1753, 2015.

[5]   H. Kaur and J. Singh, "Improvement in load balancing technique for MongoDB clusters using data center awareness and ant colony optimization," *Int. J. Appl. Eng. Res.*, vol. 10, no. 11, pp. 29155–29173, 2015.

[6]   M. P. Patel, M. I. Hasan, and H. D. Vasava, "Performance improvement of sharding in MongoDB using k-mean clustering algorithm," vol. 1, no. 5, pp. 1–5, 2014.

[7]   W. Jiang, L. Zhang, X. Liao, H. Jin, and Y. Peng, "A novel clustered MongoDB-based storage system for unstructured data with high availability," *Computing*, vol. 96, no. 6, pp. 455–478, 2014.

[8]   Y. Liu, Y. Wang, and Y. Jin, "Research on the improvement of MongoDB Auto-Sharding in cloud environment," *ICCSE 2012 - Proc. 2012 7th Int. Conf. Comput. Sci. Educ.*, no. Iccse, pp. 851–854, 2012.

[9]   V. J. Dindoliwala and R. D. Morena, "Survey on Security Mechanisms In NoSQL Databases," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 333–338, 2017.

[10]   S. Ghotiya, J. Mandal, and S. Kandasamy, "Migration from relational to NoSQL database," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 263, no. 4, 2017.

[11]   B. Jose and S. Abraham, "Exploring the merits of nosql: A study based on mongodb," *2017 Int. Conf. Networks Adv. Comput. Technol. NetACT 2017*, no. July, pp. 266–271, 2017.

[12]   S. Tiwari, M. Akkalakshmi, and K. Bhagavatula, "Analysis of NoSQL Databases : Mongodb , HBase , Neo4J," no. April, pp. 234–239, 2017.

[13]   L. Kumar, S. Rajawat, and K. Joshi, "Comparative analysis of NoSQL (MongoDB) with MySQL Database International Journal of Modern Trends in Engineering and Research (IJMTER)," *Int. J. Mordern Trends Eng. Res.*, vol. 02, no. 05, pp. 120–128, 2015.

[14]   A. Kanade, A. Gopal, and S. Kanade, "A study of normalization and embedding in MongoDB," *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, pp. 416–421, 2014.

[15]   C. Paper and C. Gy, "A Comparative Study_ MongoDB vs. MySQL.pdf," no. JUNE, 2015.

[16]   D. Damodaran B, S. Salim, and S. M. Vargese, "Performance Evaluation of MySQL and MongoDB Databases," *Int. J. Cybern. Informatics*, vol. 5, no. 2, pp. 387–394, 2016.

[17]   R. Aghi, S. Mehta, R. Chauhan, S. Chudhary, and N. Bohra, "A comprehensive comparison of SQL and MongoDB databases," *Int. J. Sci. Res. Publ.*, vol. 5, no. 2, p. 3, 2015.

[18]   R. K. S and R. R. Mary, "Comparative Performance Analysis of various NoSQL Databases : MongoDB , Cassandra and HBase on Yahoo Cloud Server," no. 4, pp. 265–269,

2017.

[19]   B. Maity, S. Sen, and N. C. Debnath, "Challenges of Implementing Data Warehouse in MongoDB Environment," *J. Fundam. Appl. Sci.*, vol. 10, no. 4S, pp. 222–228, 2018.

[20]   D. Chauhan, "Using the Advantages of NOSQL : A Case Study on MongoDB," no. February, 2017.

[21]   C.-O. Truica, A. Boicea, and I. Trifan, "CRUD Operations in MongoDB," *Proc. 2013 Int. Conf. Adv. Comput. Sci. Electron. Inf.*, no. August, 2013.

[22]   G. Zhao, W. Huang, S. Liang, and Y. Tang, "Modeling MongoDB with relational model," *Proc. - 4th Int. Conf. Emerg. Intell. Data Web Technol. EIDWT 2013*, pp. 115–121, 2013.

[23]   V. Abramova and J. Bernardino, "NoSQL databases: MongoDB vs cassandra," *Proc. Int. C* Conf. Comput. Sci. Softw. Eng. ACM 2013*, pp. 14–22, 2013.

IJSER