A COMPARATIVE ANALYSIS OF DOCUMENT-ORIENTED BIG DATA DATABASES

[1]Shikha Sangal, [2]Anita Ganpati
[1]M.Tech CS Student, [2]Professor,
[1]Department of Computer Science,
[1]Himachal Pradesh University, Shimla, India.

Abstract: Now-a-days companies are starting to realize the importance of using more data in order to support decision for their strategies. Big Data refers to penta-bytes or exabytes of data. Big Data describes a holistic information management strategy that includes and integrates many data types of data and data management alongside traditional data. Big data is a phrase that defines the large volume of data both structured and unstructured that it is so hard to process using traditional database system performance and software techniques. In the present scenario the big databases have become prominently important because it have facility to generate large amount of data for social media daily analysis and multimedia, etc. big data is an enormous area to be researched. The main objective of this research study, is to evaluate performance of document-oriented Big Data Databases namely CouchDB, Couchbase and MongoDB using YCSB; in term of average runtime and throughput with increasing of number of records and operations. Each database has its own specific use and no one fit for the requirements. The research methodology describes the tool used to compare the document-oriented databases on various features. Most of NoSQL database provide high scalability and high availability at cost compromise consistency.

Keywords: *CouchDB, Couchbase and MongoDB, Big Data.*

1. INTRODUCTION

Big data refers to the enormous amount, diverse sets of information that develop at rising rates. It encompasses the quantity of information, the speed or momentum or velocity at which it is created and collected, and the variety or scope of the data points being covered [27]. Big data often comes from multiple sources and arrives in multiple formats. Twitter, Face book, Amazon, Verizon, Macy's are all companies. Day by day, we create more than two quintillion bytes of data (2 EB), and it is expected that more than 90% of the data has been generated in the last few years alone1 KB = 1024 Bytes, 1 MB = 1024 KB, 1 GB = 1024 MB, 1 TB = 1024 GB ~ 1,000,000 MB, 1 PB = 1024 TB ~ 1,000,000 GB ~ 1,000,000,000 MB, 1 EB = 1024 PB ~ 1,000,000 TB ~ 1,000,000,000 GB ~ 1,000,000,000,000 MB such big amounts of data [1].
We have implemented this research work in different sections, Section 1 contains Introduction, Section 2 describes Big Data Databases, Section 3 depicts the comparison of MongoDB, CouchDB and CouchBase, Section 4 depicts the analysis and final results, Section 5 depicts the conclusion.

2. BIG DATA DATABASES

Now-a-days companies are starting to realize the importance of using more data in order to support decision for their strategies. Big Data refers to penta-bytes or exabytes of data. Big Data describes a holistic information management strategy that includes and integrates many data types of data and data management alongside traditional data. Big data is a phrase that defines the large volume of data both structured and unstructured that it is so hard to process using traditional database system performance and software techniques. In the present scenario the big databases have become prominently important because it have facility to generate large amount of data for social media daily analysis and multimedia, etc. big data is an enormous area to be researched. It consists of a group of programs that manipulate the database. The DBMS accepts the request for data from an application and instructs the operating system to provide the specific data. There are different types of databases in Database Management Systems which are Relational DBMS, SQL DBMS, and Object-Oriented DBMS. SQL (Structured Query Language) programming used to insert, search, update and delete database data. NoSQL: this db is an approach to design such databases that can accommodate a variety of data models. NoSQL databases are useful to a large set of data [2].
Big data databases are going with very fast speed because of their thrilling features like more flexibility and scalability, schema-free architecture, comfortable replication support, simple API consistent/base (not ACID), support for big data and more.NoSQL database system has the Key-value database, Document-oriented database, and Column-oriented database and Graph database. In this review paper, only three Document-Oriented big data databases are compared namely MongoDB, CouchDB and CouchBase [2]. In this research work we have used three type of document-oriented databases which are given below:

MongoDB Database: This database is an open-source DB which uses a document-oriented data model a non-structured query language. It is one of the most powerful NoSQL systems and databases around, today. MongoDB is a cross-platform and open-source document-oriented database, a kind of NoSQL database [11]. MongoDB shuns the relational database's table-based structure to adapt JSON-like documents that have dynamic schemas which it calls BSON. This makes data integration for certain types of applications more rapid and easier. MongoDB is built for scalability, high availability and performance from single server deployment to large and complex multi-site infrastructures. MongoDB database has many features such as indexing, schema-less database, replication, performance, Ad-hoc Queries, Sharding, etc. in NoSQL databases [14].

CouchDB Database: CouchDB is an open-source document-oriented NoSQL database, implemented in Erlang. CouchDB uses multiple formats and protocols to store, transfer, and process its data, it uses JSON to store data, JavaScript as its query

language using Map Reduce, and HTTP for an API. CouchDB implements a form of multisession concurrency control (MVCC) so it does not lock the database file during writes [7]. CouchDB can do on-the-fly document transformation as well as current real-time change notifications, making web application development easier. It specializes in Availability and Partition Tolerance (AP) but can finally be consistent through minor work [8].

Couchbase Database: Couchbase database, initially known as Membase, is an open-source, distributed (shared-nothing architecture) multi-model NoSQL document-oriented database software tie together that is optimized for interactive applications [2]. These applications may provide many concurrent users by creating, storing, and retrieving, aggregating, manipulating and presenting data. The replication, indexing, persistence, and scalability feature of Couchbase database. Couchbase is the merger of two popular NoSQL technologies:

1) Membase, which provides persistence, replication, and sharding to the high-performance in Memcached technology.
2) CouchDB, which pioneers the document-oriented model based on JSON [2].

3. COMPARISON FEATURES

CouchDB, Couchbase and MongoDB are document-oriented big data databases that have their API, data structure and data storage, system performance features in a system. NoSQL databases are highly available and scalable. A NoSQL DBMS stores each item individually with a unique key. Additionally, a NoSQL database does not require a structured schema that defines each table and the related columns. This provides a much more flexible approach to storing data than a relational database [18]. These databases used for data operations on data that are created, read, update and delete.

4. RESULTS

In Table 4.1, the apache 2.0, AGPL license type systems are project CouchDB, MongoDB, Couchbase databases. C, C++, Erlang, and JavaScript have implemented languages in CouchDB, Couchbase and MongoDB databases. Data storage, data models, data operations and system orientation are provided by all databases. In Table 4.1, it describes the various general features of document-based databases namely CouchDB, Couchbase and MongoDB.

Table 4.1: Comparative Analysis of CouchDB, Couchbase and MongoDB based on General Features [2, 5, 7, 8, 10, 11, 16, 17, 19, 20, 21, 23, 24, 25 and 26]

Sr. No.	Databases⇒ Features⇓	CouchDB	Couchbase	MongoDB
1.	**Data-Sets (real-time processing)**	Semi-structured document	Semi-structured document	Semi-structured document
2.	**API**	HTTP/REST	Key-value/Restful	HTTP REST
3.	**Data Model**	Document, schema free model	Document, key-value	Document
4.	**System orientation**	Document-JSON	Document-oriented	Document-oriented
5.	**Language**	Erlang	C, C++, Go and Erlang	C++, JavaScript
6.	**Owner**	Apache	Couchbase, inc	Adobe
7.	**Platform**	Android, Window	Open telecom	Windows, Linux. OS.
8.	**License**	Apache 2.0	Apache	AGPL
9.	**Data Storage**	B-tree	B-tree	B-tree
10.	**Map Reduce**	Yes	Yes	Yes
11.	**In memory capability**	No	Yes	Yes
12.	**Data Operations**	CRUD, Futon, and CURL	CRUD	CRUD
13.	**GUI**	Futon		Compass

It is evident from table 4.1, that each database has its API as HTTP/REST and Key-Value. They have different GUI in a system for data processing and doing the data operation known as CRUD. The above-mentioned databases support the map-reduce in the system and provide the system orientation. Each database has its platform such as Android, Windows, Linux, and OS, etc. these databases are document-based databases and schema-free.

Similarly, other system performance features are given below in Table 4.2.

Table 4.2: Comparative Analysis of CouchDB, Couchbase and MongoDB based on system performance features [4, 6, 9,12,13,14,15,19,20,21,22,23,24,25 and 26]

S. No.	Databases⇒ Features ⇓	CouchDB	Couchbase	MongoDB
1.	**Concurrency**	MVCC, Yes	Pessimistic, yes	Locks, yes
2.	**Query Method**	CURL	N1QL	Ad-hoc
3.	**Locking**	MVCC used/no	Yes	No
4.	**Multiple servers**		Yes	yes
5.	**GridFS**	No	No	Yes
6.	**Query language**	JavaScript	N1QL	JavaScript
7.	**Fault tolerance**	Yes	Yes	No
8.	**Data-portioning techniques**	Sharding	Sharding	Sharding
9.	**Notification**	Yes	Yes	Yes
10.	**Performance**	High	High	High
11.	**Scalability**	High	High	High
12.	**Availability**	High	High	High
13.	**Consistency**	Yes	Yes	Yes
14.	**Replication**	Yes	Yes	Yes
15.	**Security**	Still going on	Still going on	Still going on
16.	**Durability**	Yes	Yes	Yes
17.	**Auto-Sharding**	No	Yes	Yes
18.	**Memcached-compatible**	No	Yes	
19.	**Automatic-failure**	No	Yes	Yes
20.	**Index support**	Yes	Yes, Document	Yes, geospatial
21.	**Secondary indexes**	Yes	Yes	Yes
22.	**Aggregation**		Yes	Yes

Table 4.2, the concurrency control MVCC is provided by Couchbase, CouchDB, and locks is provided by MongoDB and ACID is provided by MySQL. Replication is provided by all databases. Performance and scalability also provided by all three databases which are above mentioned. Durability and index support also provided by databases. MongoDB supports GridFS in a system. Query language N1QL and JavaScript used in databases. Sharding data partitioning technique is used in all three databases which are above mentioned. These databases provide Memcached-compatible, index support, auto-sharding and consistency in the system for better results. Data is replicated and retrieved intelligently by these databases. They provide the concurrency and locks are used in a system. From the table 4.1 and 4.2 we can analyze that MongoDB is better than other 2 used databases. MongoDB supports more platforms than CouchDB and CouchBase as mentioned in Table 4.1. On the other hand, CouchDB supports Erlang language, CouchBase supports C, C++, Go and Erlang languages and MongoDB supports C++ and JavaScript languages.

5. CONCLUSION

Big Data is data but with giant size. Big Data is a term used to illustrate a collection of data that is vast in size and yet rising exponentially with time. Big data refers to the enormous amount, diverse sets of information that develop at rising rates. Big data is a phrase that defines the large volume of data both structured and unstructured that it is so hard to process using traditional database system performance and software techniques. In the present scenario the big databases have become prominently important because it have facility to generate large amount of data for social media daily analysis and multimedia, etc. big data is an enormous area to be researched.

The objectives of this research work are: 1) To have deeper understanding of big data. 2) To compare and analyze different document-oriented databases performance.

In this research work we have compared and examined the three databases namely: CouchDB, Couchbase and MongoDB based on system performance and features. So, from this research work we have concluded that each system has its own specific purpose and each analyze different document-oriented databases have different flaws. And, MongoDB performance is the best for maximum features.

6.Future Scope

In this research work we have compared and examined the three databases namely: CouchDB, Couchbase and MongoDB. But in future, we will compare and analyze more document-oriented databases like RavenDB, Terrastore, and OrientDB, etc can be compared. Since only document databases are considered. This work did a quantitative comparison of three document-oriented big data databases.

REFERENCES

[1] Sridhar All, (2018). "Big Data Analytics with Hadoop 3: Build highly effective analytics solutions to gain valuable insight into your big data.
[2] Guy Harrison. (2016). Next Generation Databases: NoSQL, NewSQL, and Big Data: What every professional needs to know about the future of databases in a world of NoSQL and Big Data.
[3] Carlos Coronel, (2012) "Database System: Design, implement on, and management".
[4] Boicea, A., Radulescu, F., & Agapin, L. I. (2012). MongoDB vs. Oracle - Database Comparison. 2012 Third International Conference on Emerging Intelligent Data and Web Technologies. DOI:10.1109/eidwt.2012.32
[5] Gajendran Santhosh Kumar. (2012) "A Survey on NoSQL Databases". University of Illinios
[6] Jing Han, Haihong E, Guan Le, & Jian Du. (2011). "A Survey on the NoSQL database". 2011 6th International Conference on Pervasive Computing and Applications.
[7] Sitalakshmi Venkatraman, (2016) "SQL versus NoSQL Movement with Big Data Analytics". I.J. Information Technology and Computer Science, 2016, 12, 59-66 Published Online in MECS (http://www.mecs -press.org/)
[8] Rupali Kaur, Jaspreet Kaur Sahiwal, (2019) "A review of comparison between NoSQL Databases: MongoDB and CouchDB". International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S, March 2019
[9] Satyadhyan Chickerur, (2015) "Comparison of Relational Database with Document-Oriented Database (MongoDB) for Big Data Applications". 2015 8th International Conference on Advanced Software Engineering & Its Applications
[10] NoSQL with MongoDB in 24 Hours
[11] Next Generation Databases: NoSQL and Big Data
[12] Pramod J. Sadalage "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence". 1st Edition
[13] John Zablocki, (2015) "Couchbase Essentials". Kindle Edition
[14] Cattell, Rick. (2011) "Scalable SQL and NoSQL data stores." ACM SIGMOD Record 39.4
[15] Lourenço, J. R., Abramova, V., Vieira, M., Cabral, B., & Bernardino, J. (2015). "NoSQL Databases: A Software Engineering Perspective. Advances in Intelligent Systems and Computing".
[16] Andreas Meier, (2019) "SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management". Paperback – August 29, 2019.
[17] Ian Robinson, "Graph Databases NEW OPPORTUNITIES FOR CONNECTED DATA". 2nd edition.
[18] Nayak, Ameya, Anil Poriya, and Dikshay. (2013) "Type of NoSQL databases and its comparison with relational databases," International Journal of Applied Information Systems 5.4
[19] Brown, Martin C. (2012) "Getting Started with CouchDB: Extreme Scalability at Your Fingertips," O'Reilly Media, Inc.
[20] Giama, Alex. (2017) "MASTERING MongoDB 3.x".first ed. vol.1, PACKT PUBLISHING LIMITED.
[21] Kristina Chodorow and Michael Dirolf. (2010) "MongoDB: The Definitive Guide"
[22] Maribel Yasmina Santos, "Data Models in NoSQL Databases for Big Data Contexts"
[23] Cornelia Gy_rödi, Robert Gy_rödi, George Pecherle, Andrada Olah, (2015) "A Comparative Study: MongoDB vs. MySQL". 2015 13th International Conference on Engineering of Modern Electric Systems (EMES)
[24] https://blog.eduonix.com>database, Accessed on 12th October 2019 at 3:20 AM
[25] https://dbengines.com/en/system/CouchDB%3BCouchbase%3BMong DB, Accessed on 23rd October 2019 at 5 AM
[26] https://assist-software.net/blog/couchbase-vs-couchdb-vs-mongodb, Accessed on 26th October 2019 at 7 PM
[27] https://www.guru99.com/what-is-big-data.html, Accessed on 9th November 2019 at 12 PM.