International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 15 (2017) pp. 5000-5004

Proposed Architecture of MongoDB-Hive Integration

Subita Kumari [1] and Pankaj Gupta[2]

[1]Research Scholar, Computer Science and Engineering Department,
University Institute of Engineering and Technology, Mahrishi Dayanand University, Rohtak, India.

Orcid ID : 0000-0001-7100-9326

[2]Professor, Computer Science and Engineering Department,
Vaish College of Engineering, Mahrishi Dayanand University, Rohtak, India.

Orcid ID : 0000-0002-5548-9195

Abstract

There is tremendous growth in heterogeneous and unstructured data in last few years. Various NoSQL databases have been developed to store and query this humongous data. MongoDB is prevailing document oriented database among NoSQL databases. MongoDB has been chosen among other technologies because of its ability to work with variety of latest as well as conventional technologies. It provides drivers for almost every development language. This paper summarizes work of various authors who have compared and integrated MongoDB with other big data technologies. In most cases, MongoDB gives better performance in terms of various parameters. Further, after carrying out critical analysis , an architecture is being proposed to integrate MongoDB with Hive. It utilizes the SQL like features of Hive for SQL familiar users.

Keywords: MongoDB ; Mongo-Hadoop Connector ; Hive; Hadoop Ecosystem

INTRODUCTION

Storage in large scale data centers is changing drastically. Now a days, data is stored not only at high cost enterprise server systems but also at the nodes that are to a certain extent storage, partly application and to some degree storage. These nodes originate new technology like Hadoop, Hbase, MongoDB and cloud storage. This paper deals with some of these technologies like MongoDB, Hive, Mongo-Hadoop Connector and Hadoop Ecosystem. Each of these is explained briefly.

A. MongoDB

MongoDB is a document-store database developed by 10gen. It is made up of a set of databases, which are further composed of multiple collections. Each collection consists multiple document. Every document is a JSON structure composed of key-value pairs. MongoDB is highly scalable and available data store having great performance.

B. Hadoop Ecosystem

Hadoop is an open-source integrated software package which stores and processes large datasets in a distributed manner. Its two major modules are map-reduce and Hadoop Distributed File System (HDFS). Map-Reduce is a parallel and distributive programming paradigm for processing bulk amount of heterogeneous and unstructured data on clusters of inexpensive and easily available hardware. Hadoop sends the map-reduce program to datasets stored on commodity hardware. Hadoop Distributed File System is used to store and process the datasets. The Hadoop Ecosystem is combination of Hadoop along with different sub-tools such as Pig, Sqoop and Hive that are used to help Hadoop modules. Map-Reduce operations of Hadoop can be executed using various mechanisms. The conventional method uses Java map-reduce program for semi-structured, structured and unstructured data. The scripting method uses map-reduce for semi structured and structured data using Pig. To process structured data using Hive map-reduce operations are implemented with the help of Hive Query Language.

C. Hive

Hive is a data warehouse solution to analyse structured data in Hadoop. It rests on top of Hadoop to recapitulate very large sets of data and makes analysis and querying easy. Facebook is originator of Hive. But Apache Software Foundation developed it further as open source software under the name Apache Hive. Hive is not a relational database but a language for row-level updates and real-time queries. It is a design for On Line Transaction Processing. It provides SQL type interface for querying, called Hive Query Language (HiveQL). It stores schema in a database and process data into HDFS. This paper proposes an architecture in which Hive processes data from MongoDB as well.

D. MongoDB-Hadoop Connector

The MongoDB connector for Hadoop is a plug-in which allows MongoDB to be used as an input or output or Hadoop tasks. It is developed to provide flexibility and performance to users. The connector makes it easy to integrate MongoDB with other parts of the Hadoop Ecosystem like Pig , Spark and Hive.

LITERATURE REVIEW

Alexandru Boicea et.al [1] concluded that if rapidness and flexibility of database is the major concern one can rely on MongoDB. If the quickness of the database is not the main worry, and there is want of relations between the tables and the collections, one can rely on the traditional solution, Oracle Database. Elif Dede et.al [2] experimentally found that when analysis of data stored in MongoDB needs to be done , the MongoDB connector for Hadoop is a clever way to use Hadoop for scalability. As the data volume grows in size , improved performance can be achieved if the output of the analysis could be written to HDFS rather than back to MongoDB. The Mongo-Hadoop plug-in enhance performance roughly five times compared to using MongoDB's inherent map-reduce implementation.

Veronika Abramova et.al [3] stated that MongoDB is CP i.e. Consistency and Partition tolerance on the other hand Cassandra is PA i.e. Consistency and Availability. MongoDB uses master-slave architecture while Cassandra uses peer-to-peer replication. After running different workloads to read/update for analysing performance authors found that Cassandra is faster than MongoDB for update operations and provides lower execution time irrespective of database size used. Zachary Parker et.al [4] stated that MongoDB is certainly a good option for users who need a flexible database structure. It could be a good alternate for larger data sets where less complex queries need to be executed and schema constantly changes. MongoDB works well on the complex queries except queries involving aggregate functions. This is because MongoDB has to use the map-reduce functions to create aggregate functions. Authors also found that MongoDB has a better run time performance as compared to SQL data base.

Cheng Dai et.al [5] introduce the main concepts used in MongoDB and Hadoop, then combine the merits of both to build a high performance storage structure which relies on commodity computer clusters. This infrastructure combines analytical capability of Hadoop and horizontal scalability of MongoDB. Ashish Thusoo et.al [6] present Hive, a software package built on top of Hadoop. HiveQL , which is SQL-like declarative language , is used to express queries in Hive. The queries are further compiled into map-reduce jobs that are executed using Hadoop. HiveQL also allows developer to convert build-in map-reduce scripts into queries. Hive maintains metadata about all tables using system catalogue. This enables Hive to interface with standard reporting tools.

Sanjeev Dhawan et.al [7] successfully created a map-reduce job using Pig and then analyzed a large database to get required results. Also a map-reduce job was created using Hive and then analyzed a big database to get results. Final results show that the analysis performed by using both the technologies takes almost same time. Robin Henricsson [8] show that MongoDB is significantly faster than CouchDB for insertion operation as well as for querying, when used with their respective Python libraries. Additionally MongoDB is more space efficient than CouchDB.

Sumita Barahmand et.al [9] analyze scalability features of MongoDB and HBase for processing interactive social networking actions using a benchmark. These simple and basic actions write and read a small amount of data. Speedup and scale up is being quantified with a multi-node deployment of both databases. It is observed that each databases scales linearly, their speedup is restricted by the resources of only a few nodes. In case of MongoDB the impact of this limitation is decreased and speedup is improved by use of replicated shards for query processing. But this generates a very small amount of stale data.

Ruxandra Burtica et.al [10] develop an application which keeps track of a keyword over several social networking websites and keeps the aggregated data about keyword in NoSQL databases. After testing various NoSQL databases, it is concluded that the most appealing solution for the above mentioned application is achieved by using MongoDB. It provides Partitioning and Consistency but drops Availability. Safety measures have been developed by authors to ensure the announcement when something crashes in the system to fullfill lack of availability feature. HBase is extravagant for the needs of application. Cassandra needs restructuring of data to get key-value records. Redis is quite fast and has a lot of appealing features.

Subita Kumari et.al [11] explain theoretical differences between CouchDB and MongoDB. MongoDB supports consistency and partitioning tolerance but relaxes availability. On the other hand , CouchDB supports availability and partitioning tolerance but relaxes consistency.

Ayush et.al [12] evaluate performance of NoSQL databases MongoDB and Cassandra by integrating them with Hadoop. Hadoop-MongoDB integration is not read efficient while MongoDB is very read efficient. Hadoop-Cassandra integration is very stable against writes for exponentially growing data. For fault tolerance and scalability also Hadoop-Cassandra has shown better results.

Barkha Jain et.al [13] have executed various query implementation on Hive for huge datasets by using join , indexing and lateral view. Map-reduce functionalities of HDFS supported Hive to process bigger and unstructured

datasets. Conventional RDBMS databases are highly optimized as compared to Hive. Authors have examined the execution time of the queries run on the Hive as well as query run on RDBMS and analysed that response time of the queries in Hive is better as compared to RDBMS for larger data sets. Ramon Lawrence [14] has developed Unity. It is a virtualization and integration system which allows SQL queries to run over Relational as well as NoSQL databases. The additional virtualization layer allows converting SQL queries to NoSQL APIs. Unity engine gives as good performance as MySQL in running queries. MongoDB cannot execute joins but with the help of this virtualization engine collections can be joined in MongoDB efficiently.

CRITICAL ANALYSIS

New tools and technologies are being introduced in the industry by software giants as well as individuals to deal with huge amount of data which is being generated because of evolution of internet. Many NoSQL databases and analytics tools have been studied and analyzed by various researchers throughout the world. A comparison of the work of different researchers has been presented in the tabular form (Table 1).

Table I: Critical Analysis

S.No.	Authors	Publication	Title of the Paper	Technologies Compared	Conclusion
1	Alexandru Boicea, Florin Radulescu, Laura Ioana Agapin	IEEE,2012	MongoDB vs Oracle - database comparison	MongoDB, Oracle	When rapidness and flexibility is major concern rely on MongoDB.
2	Elif Dede , Daniel Gunter , Madhusudhan	ACM, 2013	Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis	MongoDB , Hadoop	Mongo connector for Hadoop is a clever way to use Hadoop for analysing data of MongoDB.
3	Veronika Abramova, Jorge Bernardino	ACM, 2013	NoSQL Databases: MongoDB vs Cassandra	MongoDB , Cassandra	Cassandra is faster than MongoDB for update operations
4	Zachary Parker, Scott Poe, Susan Vrbsky	ACM, 2013	Comparing NoSQL MongoDB to an SQL DB	MongoDB , SQL Server	MongoDB gives better run time performance as compared to SQL.
5	ChengDai, YanYe, TaijunLiu, JingjingZheng	Trans Tech Publications,2013	Design of High Performance Cloud Storage Platform Based on Cheap PC Clusters using MongoDB and Hadoop	MongoDB , Hadoop	Combines horizontal scalability of MongoDB and analytical capability of Hadoop
6	Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao	ACM,2009	Hive – A Petabyte Scale Data Warehouse Using Hadoop	Hadoop,Hive	Queries written in HiveQL are converted into map-reduce programs which are executed by Hadoop
7	Sanjeev Dhawan, Sanjay Rathee	AIJRSTEM,2013	Big Data Analytics using Hadoop Components like Pig and Hive	Pig ,Hive	Analysis performed by using both the technologies takes almost same time.
8	Robin Henricsson	BTH ,2011	A comparison of performance in MongoDB and CouchDB using a Python interface	MongoDB , CouchDB	MongoDB is significantly faster than CouchDB for insertion operation as well as for querying.
9	Sumita Barahmand, Shahram Ghandeharizadeh and Jia Li	DLTR,2014	On Scalability of Two NoSQL Data Stores for Processing Interactive Social Networking Actions	MongoDB , Hbase	Both databases scale super linearly but their speedup is restricted by the resources of only some nodes. MongoDB reduces the impact of this limitation by use of secondary shards for query processing.
10	Ruxandra Burtica, Eleonora Maria Mocanu, Mugurel Ionuț Andreica, Nicolae Țapuș	IEEE,2012	Practical application and evaluation of no-SQL databases in Cloud Computing	Redis, MongoDB, Cassandra and HBase	Most optimal solution for the mentioned problem is achieved using MongoDB.
11	Subita Kumari , Pankaj Gupta	RG Education Society (INDIA) , 2015	Document Store NoSQL Databases	MongoDB , CouchDB	MongoDB supports CP and CouchDB supports AP.
12	Ayush , Seema Bawa	Elseveir , 2014	Performance Analysis of NoSQL Databases Having Hadoop Integration	Hadoop-Cassandra , Hadoop-MongoDB	For read/write and scalability Hadoop-Cassandra integration has shown better results as compared to Hadoop-MongoDB integration.

13	Barkha Jain , Manish Km. Kakhani	Krishi Sanskriti Publications, 2015	Query Optimization in Hive for Large Datasets	Hive , HDFS , RDBMS	Response time of the queries run on Hive is better as compared to queries run on RDBMS for larger data sets.
14	Ramon Lawrence	IEEE,2014	Integration and Virtualization of Relational SQL and NoSQL Systems including MySQL and MongoDB	MySQL ,MongoDB , Unity Virtualization Engine	By using MongoDB, Unity engine gives as good performance as MySQL in running join queries.

PROPOSED ARCHITECTURE FOR MONGODB-HIVE CONNECTOR

It is very much clear from the critical analysis that MongoDB is very fast , scalable and high performance NoSQL database among its current competitors. MongoDB is a document oriented NoSQL database that powers the real time , online operational applications and serves end-users as well as business processes. But when it comes to analytics Hadoop overpowers MongoDB because Hadoop is inherently designed for storing , analyzing and processing huge volumes of data which are further distributed over a cluster of inexpensive servers and commodity storage. There is need to enable MongoDB to use analytics model of Hadoop to enhance its analytical capabilities for its operational processes. Along with Hadoop map-reduce and HDFS the Hadoop Ecosystem also has Pig, Hive and other family of languages which are expressive and high-level and let user write their own custom code.

The proposed integration of MongoDB and Apache Hive is achieved with the help of Mongo-Hadoop Connector. With the help of Mongo-Hadoop connector, Hadoop consumes data from MongoDB, mix it with data from other sources to generate complicated analytics and results are stored back to MongoDB to serve operational processes. The proposed architecture to connect MongoDB and Hive is depicted in Fig. 1.

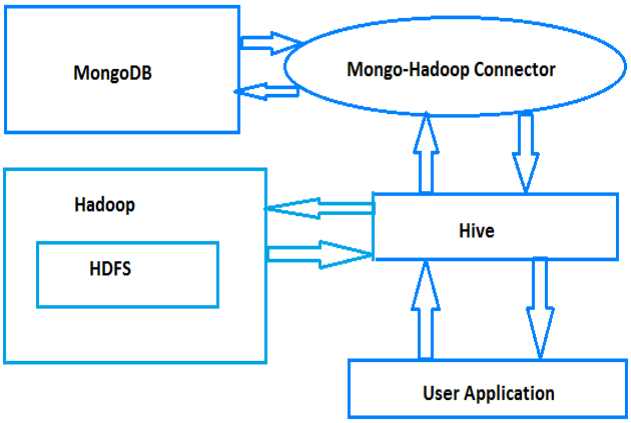

Figure 1. Mongo-Hive Integration

MongoDB supports its native map-reduce in JavaScript, but writing map-reduce is a difficult task. Joining multiple collections in MongoDB is almost impossible although it supports the aggregation pipeline, but it is very hard to use and no custom code available for it. Hive, which is a part of Hadoop Ecosystem as an independent module , looks similar to SQL query language . It converts SQL queries into various map-reduce stages and generates output accordingly. Hive lets user join and cross multiple collections of MongoDB because it helps in churning of data in SQL style. MongoDB can use the custom features of Hive with the help of Mongo-Hadoop Connector. Mongo-Hadoop connector allows creating MongoDB-based Hive tables and these can be queried like HDFS-based Hive tables. User can write queries in SQL like interface and these are converted in map-reduce functions by Hive to import/export data from MongoDB , making user job pretty easy.

CONCLUSION AND FUTURE WORK

Work of various authors who have compared and integrated MongoDB with other big data technologies have been summarized in this paper. A quick look of comparison of various new era technologies is being presented in tabular form. In most cases, MongoDB gives better performance whether used in isolation or in integration with other traditional and latest technologies. An architecture is being proposed to integrate MongoDB with Hive. MongoDB can use the custom features of Hive with the help of Mongo-Hadoop Connector. As a future work , this architecture can be tested using large data sets to check enhancement in analytic capabilities of MongoDB. Also plug-in for different databases or tools can be developed to integrate with MongoDB to utilize its awesome capabilities.

REFERENCES

[1] Alexandru Boicea, Florin Radulescu, Laura Ioana Agapin, "MongoDB vs Oracle - database comparison" , Third International Conference on Emerging Intelligent Data and Web Technologies 2012 (IEEE).

[2] Elif Dede , Daniel Gunter , Madhusudhan ,"Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis" , ScienceCloud'13(ACM), June 17, 2013.

[3] Veronika Abramova, Jorge Bernardino,"NoSQL Databases: MongoDB vs Cassandra", Conference C3S2E(ACM), July 10–12, 2013.

[4] Zachary Parker , Scott Poe ,Susan V. Vrbsky ," Comparing NoSQL MongoDB to an SQL DB", ACMSE'13, April 4-6, 2013.

[5] ChengDai, YanYe, TaijunLiu, JingjingZheng," Design of High Performance Cloud Storage Platform Based on Cheap PC Clusters using MongoDB and Hadoop", Applied Mechanics and

Materials Vols 380-384,2013.

[6] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu, Raghotham Murthy ," Hive – A Petabyte Scale Data Warehouse Using Hadoop", VLDB Endowment, ACM, August 24-28, 2009, Lyon, France.

[7] Sanjeev Dhawan, Sanjay Rathee," Big Data Analytics using Hadoop Components like Pig and Hive ", American International Journal of Research in Science, Technology, Engineering & Mathematics ,pp. 88-93,March-May, 2013.

[8] Robin Henricsson," A comparison of performance in MongoDB and CouchDB using a Python interface", Bachelor thesis BTH ,2011.

[9] Sumita Barahmand, Shahram Ghandeharizadeh and Jia Li , "On Scalability of Two NoSQL Data Stores for Processing Interactive Social Networking Actions", Database Laboratory Technical Report 2014-11 , March, 2015.

[10] Ruxandra Burtica, Eleonora Maria Mocanu, Mugurel Ionuţ Andreica, Nicolae Tapus," Practical application and evaluation of no-SQL databases in Cloud Computing ", 978-1-4673-0750-5 IEEE,2012.

[11] Subita Kumari , Pankaj Gupta," Document Store NoSQL Databases", International Journal of Artificial Intelligence and Knowledge Discovery Vol.5, Issue 3, July, 2015.

[12] Ayush , Seema Bawa,"Performance Analysis of NoSQL Databases Having Hadoop Integration", ERCICA , Elseveir ,Aug 2014.

[13] Barkha Jain , Manish Km. Kakhani ," Query Optimization in Hive for Large Datasets" , Advances in Computer Science and Information Technology (ACSIT) Volume 2, Number 4; April-June, 2015 .

[14] Ramon Lawrence," Integration and Virtualization of Relational SQL and NoSQL Systems including MySQL and MongoDB ", International Conference on Computational Science and Computational Intelligence (CSCI), IEEE 2014.

[15] Chodorow K (2013) MongoDB: the definitive guide. O'Reilly Media Inc.

[16] "Apache Hadoop," Apache. [Online]. Available: http://hadoop.apache.org/.

[17] Min Chen, Shiwen Mao and Yunhao Liu, "Big Data: A Survey", Springer, School of Computer Science and Technology, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, 430074, China (2014).

[18] Mango DB, "Top 5 considerations when evaluating NoSQL Databases",WhitePaper.

[19] "Apache Hive," http://hortonworks.com / hadoop /hive/,2011-2014