International Journal of Engineering & Technology, 7 (2.12) (2018) 344-347

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Handling data analytics on unstructured data using mongo DB

Rajani Kanth Aluvalu [1] *, M.A.Jabbar [2]

[1] Associate professor Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India
[2] Professor Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India
**Corresponding author E-mail: rajanik.rkcet@gmail.com*

Abstract

Nowadays the amount of data generated from various device sources and business transactions is very huge. Most of the transactional, business data generated is unstructured. Business organizations use the data to perform analytics for decision making. Performing Analytics on such huge unstructured data has become a challenge for organizations. Enough tools and techniques both with free ware and proprietary license release are available to handle structured data are available. In earlier systems, unstructured data is converted into structured data and then stored in Database Management System (DBMS) for performing further analytics. This is a time consuming process. As the amount of data being generated is increasing tremendously, it has become impossible to transform huge amounts of data into structured data. In order to perform analytics of the digital data, we require different business processes to handle unstructured data directly and efficiently. In this paper, a skillful mechanism is being proposed to handle unstructured data using MongoDB and perform required analytics. The experimental approach and the results are presented.

Keywords: *ETL; Unstructured Data; Data warehouse; Document Based Model; Mongodb; Big Data.*

1. Introduction

Huge amount of data is being generated by business transactions daily. It becomes very difficult to manage and perform analysis on this bulk data. Generally, the data generated is in Terabytes (TB). We should properly manage this data in order to use it further for research purpose. Big data poses a big challenge as only a limited amount of data is useful which leads to more filtering time. Based on different requirements, there will be a need to sort the data so that it can be reused again for analysis [1]. Big data describes the huge volume of data – both structured as well as unstructured. It is described as the process of collecting and storing huge amount of data for eventual analysis. The data generated is typically in terabytes or zettabytes. Big Data can be characterized by 3Vs namely volume, variety, and velocity. Unstructured data consists of information which is not available in the traditional row-column database. It is exactly the different from structure data which is stored in the form of rows and columns. It becomes a very critical process to extract useful information from unstructured data. So various types of software solutions are needed for finding unstructured data and obtain useful information from it [3]. Digital Data is broadly categorized into structured data, semi-structured data and unstructured data. The below table-1 shows the comparison between all the three kinds of digital data.

In organizations growth of the data has become highly exponential. The major objective of data science systems is to capture data and help the organizations in making business decisions. The biggest challenge in data Science is handling such exponential data from generated from various sources and performing analytics. Organizations are in need of real time analytics to make business decisions. Current day systems depend on OLAP operations for making decisions on real time data. Traditional OLAP operations cannot support handling data analytics in real time. Real Time On-Line Analytical Processing (RT-OLAP) will support organizations in making faster decisions from heterogeneous digital data. Competitive enterprises are in urgent need of RT-OLAP for big data.

Big data analytics explores large amounts of data to determine hidden patterns, customer choices and other important businessrelated information. Data needs to be stored in the Data Warehouse in order to perform further analysis. There is a read-only data in the warehouse and it is being updated on a regular basis. It possesses 4 characteristics known as non-volatile, integrated, subject-oriented and time variant [5].

Approximately 80-90 percent of the generated data is mostly unstructured data which can't be processed directly. ETL process is used in order to deal with the data generated from various sources[7]. ETL stands for Extract-Transform-Load. ETL describes the way in which the data is loaded from the source system and stored in the data warehouse [8]. The process of loading DW, usually referred to as ETL process, can take anywhere from a few hours to several days to complete, and can require many gigabytes of storage to hold the many versions of staging and interface tables. Because of the time taken by this traditional form of ETL, it's usually the case that the data in warehouses and data marts is at least a day or two out of date; in fact, usually it is between a week and a month behind their source systems. This latency as the situation is known, has in the past however not usually been seen as too much of an issue, as DWs were not considered operational, business-critical applications, but rather more as a way for a small group of analysts to do trend analysis

Hadoop is an open source framework used to store distributed data and also performs distributed processing of a very large number of datasets. It can handle the massive amount of data very quickly and efficiently. Various data formats supported by Hadoop are JSON, XML, and other text-based file formats. It is cost effective because it makes use of commodity hardware for storing a huge quantity of data. HDFS makes use of a master and slave architecture in which master contains a single NameNode used for managing the file system metadata and one or more slave DataNodes that actually

stores the data. Hadoop Distributed File System [HDFS] is being used by Hadoop as the primary file system. [11].
MapReduce is the core component of Apache Hadoop software framework. It is a parallel programming used majorly to implement processing and generating large datasets. MapReduce provides enormous scalability among hundreds or thousands of servers residing in a Hadoop cluster. Generally, input, as well as output is stored in a file system. The framework is responsible for scheduling the tasks, monitoring them and it also capable of re-executing the failed tasks. The MapReduce framework comprises of only one master JobTracker and one slave TaskTracker per cluster node. [11].

Table 1: Digital Data Comparison

	Structured Data	Semi-Structured Data	Unstructured Data
Characteristics	- data is stored in the form of rows and columns - conforms to a data model - attributes in a group are the same	- does not conform to any data model but contains tags and elements [metadata] - attributes in a group may not be the same - similar entities are grouped	- not in any particular format or sequence - does not conform to any data model - not easily usable by a program - does not follow any rules or semantics
Sources	- Databases - Spreadsheets - SQL - OLTP systems, etc.	- E-mail - XML - Zipped files - Mark-up languages, etc.	- Web pages - PowerPoint presentations - Videos, Images - Reports - Surveys, etc.

2. Related work

Traditionally, most data warehouses (DW) use a staging approach to loading data into the warehouse fact and dimension tables. Data that needs to be loaded into the warehouse is first extracted from source tables, files, and transactional systems, and is then loaded in batches into interface tables within the staging area. Then, the data within these interface tables is transformed, cleansed, and checked for errors, often with temporary staging tables being created on the way to hold the versions of the source data as it is being processed. This cleansed and transformed data is then loaded into the presentation area.
ETL process involves extraction of data from various sources and performs different processes such as cleaning, formatting and loading data into the data warehouse and performs analytics. While dealing with large sets of data, the traditional ETL process becomes less effective and consumes extra time to perform desired operations [1]. Hadoop is the framework provided by Apache project used for analyzing a large data set. Map Reduce procedure is used along with Hadoop. HDFS stands for Hadoop Distributed File System. HDFS architecture consists of two main nodes: Name Node and the Data Node. The task of a Name Node is to work as a backup node or a checkpoint node as per the requirements. [11].
Data warehouse and ETL process are used by enterprises so that they can focus to improve their performance. It mostly consists of flats files or a database. Also, the data consists of various data types. So in this type of situation ETL process becomes complex and also it requires more time and money. First of all the data is extracted from various sources in ETL process and then it is transformed to desired data format and then it is loaded into a data warehouse. Various ETL tools are available for processing data such as Informatica, SIRIUS and Data Stage and various others. ETL tool allows us to view any changes in the source data and afterward we can make effective changes in another connected data source. [12]
This paper provides a detailed study on the unstructured to distributed data mining. For unstructured data in order to analyze distributed data mining, for generating a new structured data model the NETMARK technique is used which serves as the primary purpose of this paper. We have a variety of methods available to find structured data from unstructured data. Querying unstructured data is a very tough challenge because it requires special knowledge and also a need to implement several techniques on an error-prone basis. A Large quantity of data is generated on a daily basis due to the enormous rise of the computers and information systems. This generated data is difficult to manage and organize because it is in an unstructured format. We are aware of a familiar example of finding the unstructured data known as Google which provides users with an internet search tool. [14].

2.1. Dis-advantages with traditional databases for performing analytics

The major disadvantage with traditional databases lies with their acceptance for a particular schema. We require databases to analyze unstructured and semi structured data. Most databases suffer from scaling, query latencies increases with increase in data. Analyzing realtime data which requires running aggregate queries takes longer time and we are required to place response to such queries in realtime.

2.2. Mongodb for analytics

MongoDB supports all kinds of digital data i.e. Any structure, any format, any source and no matter how often it changes. MongoDB being a No SQL database do not have a SQL interface. NO SQL data bases are developed to support non relational data systems. The analytical engine will be comprehensive and real-time. MongoDB can be used to build scalable databases which can reside and work on commodity hardware in the cloud or available locally in our data center without procuring any additional complex hardware or extra software. The important feature of MongoDB is that it can analyze data of any structure in real time directly within the database and gives the analytical results in real-time without having any expensive data warehouse workloads.
MongoDB uses document model which helps it in storing data of any structure. Most organizations are using MongoDB for analytics as it stores any kind of data and most of the organizations end up in generating unstructured data.

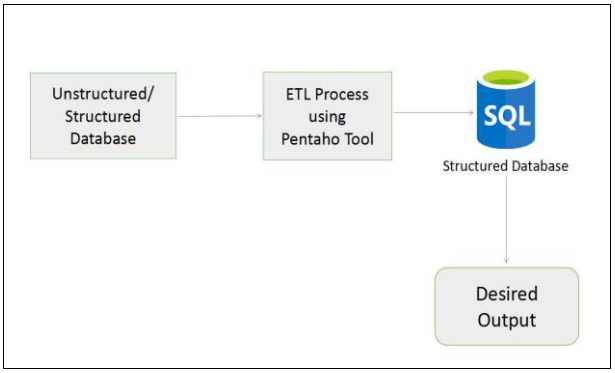

Fig. 1: Existing System Model.

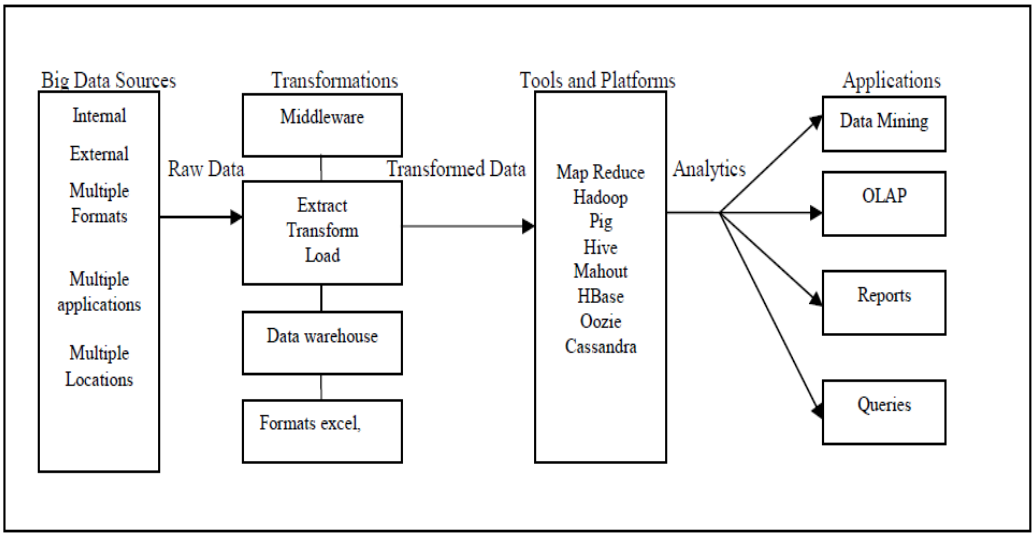

Fig. 2: Traditional Data Analytical Approach.

3. Proposed work

Due to the tremendous usage of internet and its different services, huge quantity of log data is being generated and it becomes very difficult in managing and to process this data.
Relational Database Management System [RDBMS] used to easily manage traditional data. But for this growing amount of data, we require some solutions and tools to handle these large data So we have to perform an ETL process on the data. As shown in figure-1, unstructured data is inserted into MongoDB and ETL procedure is applied to obtain desired output. This procedure will improve the data quality.
SQL Query processing requires structured data .This demands conversion of unstructured and semi-structured data into structured data for performing OLAP operations. Required analytics are performed on converted data by applying proper SQL queries to the file. As per figure - 2, sample data in unstructured format is inserted into MongoDB. Afterward with the use of different Query Evaluation Engine, we find desired results by applying query on the database.

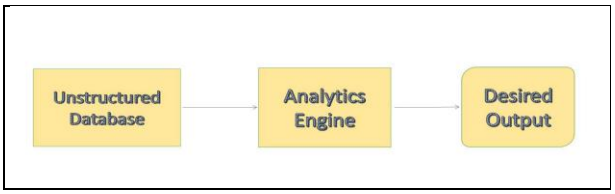

Fig. 2: Proposed System Model.

4. Experiment evolution

Query execution of unstructured data converted to structured data:
For experimental setup, we have loaded various types of digital data as input into MongoDB. We have loaded CSV file, MP3 file and an image file as input data. All the input data files contain unstructured data. We have created a database named gridfs to store data of multiple documents. Below figure -3 shows creation and data insertion into the database. Later we have imported more data into the gridfs collection, after execution. Then we have performed basic commands to arrange the documents in proper sequence. Below figure shows the sequence of command performed.

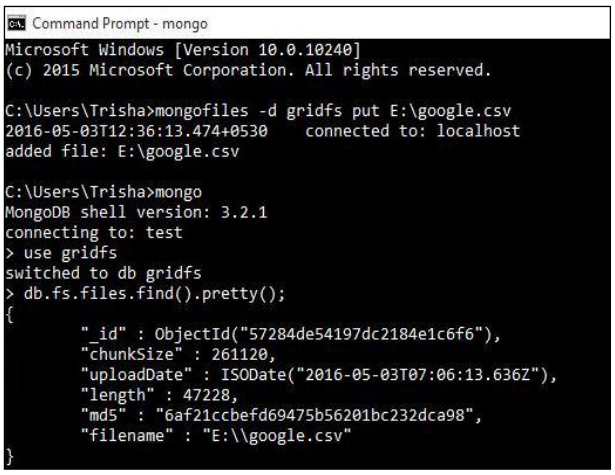

Fig. 3: Creation and Insertion of the Database in Mongodb.

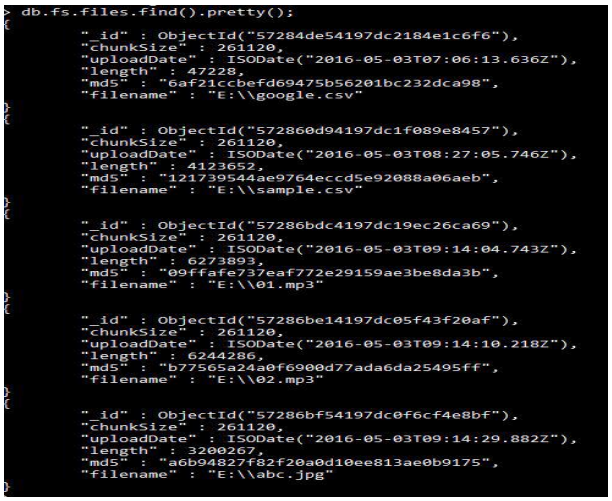

Fig. 4: Result Showing All the Documents in the Collection.

Later we have loaded the imported files in Mysql database. The results are displayed in below figure 5.

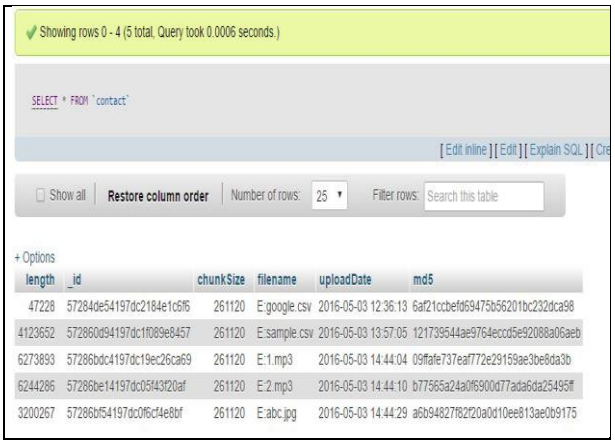

Fig. 5: File Importing.

4.2. Execution time comparison for both the mongoDB and SQL queries

Comparison is done based on the timing constraints from the results derived from SQL and MongoDB queries. Below figure-6 shows the queries performed on data in SQL and MongoDB environment. Table-2 shows the query processing time recorded for SQL and MongoDB. The results show the time taken for query processing in MongoDB is very less and almost negligible when compared to the time taken for execution in SQL environment.

	SQL	MongoDB
Query 1	select c1.filename, c2.filename from contact c1 INNER JOIN contact c2 on c1.md5 = c2.md5 and c1.filename = c2.filename ORDER BY c1.chunkSize	db.fs.files.find().sort({filename:1}).pretty()
Query 2	select SUM(c1.chunkSize + c2.chunkSize) from contact c1 INNER JOIN contact c2 on c1.md5 = c2.md5 and c1.filename = c2.filename ORDER BY c1.chunksize	db.fs.files.aggregate({ $group: { _id:"",chunksize: {$sum:'$chunksize'} } }, { $project: { -id:0, chunksize:'$chunksize' } })
Query 3	select AVG(c1.chunkSize) from contact c1 INNER JOIN contact c2 on c1.md5 = c2.md5 and c1.filename = c2.filename ORDER BY c1.chunkSize	db.fs.files.aggregate([{ _id: "" , chunksize: {$avg: "$chunksize"} }])

a

Fig. 6: Queries of MongoDB SQL.

Table 2: Timing Differences of Mongodb and SQL Queries

	MongoDB Implementation	SQL Implementation
Query1	0.0037	0.0282
Query2	0.0023	0.0247
Query3	0.0016	0.01

5. Conclusions

In this paper, the problems associated with handling various kinds of digital data from heterogeneous data sources and performing analytics on the same are discussed. Our model helps in overcoming the overhead associated with converting unstructured data into structured data and performing real time analytics. Our experimental results had proved that the proposed model is efficient in handling unstructured data and performing required analytics. Proposed model is the need of the hour for various data- intensive organizations.

6. Future work

This paper involves a huge amount of rapidly generating data. So the data in various files and formats are obtained and that needs to convert into a proper format for analytics purpose. For this matter, ETL is used for obtaining scalability and high performance. MongoDB is used for handling unstructured data. Then by using Pentaho Data Integration tool, the data from MongoDB is imported by using procedures of fields and input options. Then connection with some output files such as MS-Access, Microsoft Excel or SQL file is done. While Pentaho transformation is executed, the desired file output is obtained from the input file. To store the result file, HDFS file storage is used. Then the resultant data is queried using a scripting language called Pig Latin. The output obtained will be used for the Big Data Analytics purpose. In near future, there is a possibility of applying some security features and also some administrative access. Now the process can be used in great extent for enterprise as well as business purpose.

References

[1] Kumar P., Gopal M. Ivel, S. [2014]. "Extract Transform and Load Strategy for Unstructured Data into Data Warehouse Using Map Reduce Paradigm and Big Data Analytics", IJIRCCE International Journal of Innovative Research in Computer and Communication Engineering.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] Kumar P., Gopal M. Ivel, S. [2014]. "Extract Transform and Load Strategy for Unstructured Data into Data Warehouse Using Map Reduce Paradigm and Big Data Analytics", IJIRCCE International Journal of Innovative Research in Computer and Communication Engineering, 02[12].

[4] "Challenges and Opportunities with Big Data" by A Community White Paper developed by leading researchers across the United States.

[5] Manoj Manuja Deepak Garg, "Semantic Web Mining of Un-structured Data: Challenges and Opportunities" by International Journal of Engineering [IJE], Volume [5] : Issue [3] : 2011

[6] Subramaniyaswamy V, Vijayakumar V, Logesh R and Indragandhi V, "Unstructured Data Analysis on Big Data using Map Reduce" in Procedia Computer Science 50 [2015] 456 – 465

[7] El-Sappagh, S. H., Hendawi, A. M., Bastawissy, A. H. [2011]. "A proposed model for data warehouse ETL processes", Journal of King Saud University, Computer and Information Sciences, 23[2], 91-104.

[8] Shaker H. Ali El-Sappagh, Abdeltawab M. Ahmed Hendawi , Ali Hamed El Bastawissy, "A proposed model for data warehouse ETL processes" in Journal of King Saud University.

[9] "How is Extraction important in ETL process?" by Sweety Patel Department of Computer Science, Fairleigh Dickinson University, USA, Mrudang D. Pandya Ganpat University, Ganpat Vidyanagar, Mehsana, Gujarat.

[10] "A Review Paper on scope of ETL in Retail Domain" by Satkaur, Anuj Mehta, Research Scholar, S.K.I.E.T. Asst. Prof., S.K.I.E.T. Kurukshetra, Haryana, India Kurukshetra, Haryana, India, in International Journal of Advanced Research in Computer Science and Software Engineering .

[11] White Paper , "Extract, Transform, and Load Big Data with Apache Hadoop" in Big Data Analytics

[12] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System" in 2010 IEEE.

[13] N. Nataraj, Dr. R.V. Nataraj, "Analysis of ETL Process in Data Warehouse" in International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014, ISSN 2091-2730

[14] Satkaur, Anuj Mehta, "Proposed Work on ETL" in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June2013

[15] Padmapriya. G, M. Hemalatha, "A Recent Survey on Unstructured Data to Structured Data in Distributed Data Mining" in Padmapriya et al, Int. J. Computer Technology Applications, Vol 5 [2], 338-344.