Indian Journal of Science and Technology, Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/84173, January 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Big Data Architecture for Capturing, Storing, Analyzing and Visualizing of Web Server Logs

P. Parthiban[1*] and S. Selvakumar[2]

[1]Computer Science and Engineering, Bharath University, Chennai - 600073, Tamil Nadu, India; Parthiban.prathiba@gmail.com
[2]Computer Science and Engineering, GKM College of Engineering and Technology, Anna University, Chennai - 600063, Tamil Nadu, India; sselvakumar@ieee.org

Abstract

Background/Objectives: To propose an Automated model to capture, store, analyze and visualize of Big Data. **Methods/Statistical Analysis:** The web server logs are captured in real time and are stored into NoSQL database. The analysis on web server logs is performed using object-oriented language plus mongodb and finally the analyzed results are visualized in the R environment. **Findings:** To do a real time analysis of web server logs in data center using mongodb. Tracking of client location can be done using GeoIP. This GeoIP location information can be analyzed to find out the total number of visits from different country, real time information about server status of success, failure, warning and fatal errors. Real time updates of exact location of client who access sever and real time information of server status helpful to avoid repeating failure in a system. **Application/Improvements:** For the future work, we have planned to integrate our automated model with stream data processing system and to compare our proposed model with existing framework.

Keywords: Analyzing, Big Data, Capturing, Storing, Visualizing, Web Server Logs

1. Introduction

Big data is collection of huge volumes of data and is categorized into structured, semi-structured and unstructured. Structured data is a representation of rows and columns and it is stored in file formats. Examples for structured data are XML, data warehousing, databases and enterprise resource planning and customer relationship management. Semi-structured data is of the form bibtex file and SGML. Unstructured data is email, video files, audio files, word documents. The web server logs are captured using tools such as Flume and Kafka. Log files are generated at high rates in a server and data is collected using agent-collector methods. The collected data is dumped into the file system such as HDFS. Big data is stored in a distributed manner and should have fault tolerance, sharding, scalability, data availability and high performance. The processing of big data is done in batch processing as well as real time processing.

Few works has been carried out in the past decades in the area of NoSQL. Real time data such as stock prices, online transactions can be stored in the Redis database for analysing and it is disk backed in-memory database. For storing and analysing of log data, HBase is used and it is a column oriented database and can achieve scalability. HBase is integrated on top of Hadoop. CouchDB is easy to use and it is widely used in the Customer Relationship Management and it stores data in JSON documents. The tracking and analysing of real time data, sensor device data and social media analytics can be done using Cassandra.

Capturing big data is done using agent-collector technique and data is collected across collector which is stored in the master node and it is dumped into the HDFS. Storing of big data in HDFS is done in a distributed environment and its features are manage, scalable, availability, performance, sharding, replication and full indexing support. Processing of big data is done either in batch process or via real time process. Analyzing the big data is done for the descriptive analytics and predictive analytics.

* Author for correspondence

To analyze the Big Data using NoSQL Database i.e. MongoDB. In the Relational database, storing, analysing and visualizing of Big Data is Difficult. MongoDB is used to store text, images, audios and videos and can perform analysis on stored data in the database. Advantage of using MongoDB is Dynamic Schema, High Availability, Low-latency and Scalability and can have index on any attributes.

The MongoDB is used for storing of Big Data and analysing is performed using Aggregation operations. MongoDB achieves high availability via Replication and can store huge volumes of data across multiple servers via sharding. MongoDB can use in many use cases such as Fraud Detection, Product Catalogs, User Data Management and Content Management and Delivery.

2. Related Work

The data-driven model is proposed for processing big data and collection of information from different sources for mining and data analysis[1]. Exploring things from unstructured data and finding insights for decision making and business development in the organization[2]. Processing large volumes of data in big data is better compared to traditional databases[3]. Big data is generated from different sources such as Application/server log, cell phone GPS signals, scientific research, telemetry data and transaction records of online purchase and finding insights from the big data is the challenging in the modern world[4]. The method for visualizing of big data is crucial and analyzing is done to improve the method for customizing it to gain attention from the business analyst, data scientists and researchers[5,6].

The processing of events logs to improve the business process is the challenging task in the industry and proposing architecture for analyzing event data to get process flow for better understanding to increase the performance of the business using a distributed processing environment[7]. Analyzing big data in motion is performed using InfoSphere to give better throughput in the short time. Stream data is generated from online purchase and to be analyzed to get quick response for better decision making and creating new opportunities for the organization[8]. Analyzing time series data in real time is the challenging task and is done for predicting the future to increase the performance of the business and visualize the time series data for better understanding of the behavior[9,10].

Extracting knowledge from big data using natural language processing can be done and polarity classification is used for sentiment analysis of unstructured data[11-13]. Extracting frequently occurring patterns from big data is difficult and need to propose an algorithm for mining pattern from large volumes of data[14-16]. The algorithm for extracting knowledge from big data is for creating business opportunities and to make better decisions[17,18]. The top concern in the big data field is the security and an approach is proposed for overcoming this issue. Mining of medical and biological data is increasing at high rates and is to be analyzed using the big data technologies for better throughput and to find out insights[19-22].

Mining of biological data is a complicated task and is needed a processing methods to analyze the biological big data. Biological data may be structured or unstructured and is processed using diagnosis methods. The web services are needed to inter-operate data and applications have been discussed in this paper[23]. To publish the web services in the UDDI Registry for the reference to the consumer and is accessed via HTTP protocols and the data is exchanging using XML. The consumer raises a request in the UDDI registry and the response to the request is the services[24-26].

The web service are accessed via remote server and is accessed by users all over the world and is required to install few software for making use of these web services[27,28]. The Biomedical image analysis is needed by research professionals and to compute the results in the efficient manner is discussed in this paper[29].

3. Automated Model for Big Data Analysis

The goal is to provide an automated model for extracting knowledge from big data and finding insights out of it. The data is generated at very fast rate at the server side and the challenging task is to capture the data in the real time and is to be stored in the database for the analysis of big data and to visualize data in R. The automated model is divided into the following components.

- **Client-Server**: The web server logs are captured in real time using client-server model. The client-server model is implemented for capturing web server logs. The agent module is implemented at the server side for reading the logs in the real time. The collector module is implemented at the client for writing the logs to a file.

- **Mongodb**: Mongodb is used for storing structured and unstructured data. The web server logs are stored into the Mongodb using the mongoimport functionality. Mongodb stores the documents in the form of BSON format. Mongodb is used for analysing of big data. The analysis of web server logs is done using java and mongodb. The analyzed results are stored into the Mongodb for visualization purpose.
- **Java**: Java is used for integrating client-server model with mongodb to analysis the web server logs. The java programming language is linked with mongodb using mongo-java jar, for the analysis of big data in a secured way.
- **R Language**: The analyzed results are visualized in the R environment. R is used for data analysis and statistical computing. Using R environment, we can generate reports for the purpose of future prediction and finding insights out of the reports using the predictive model and statistical model.

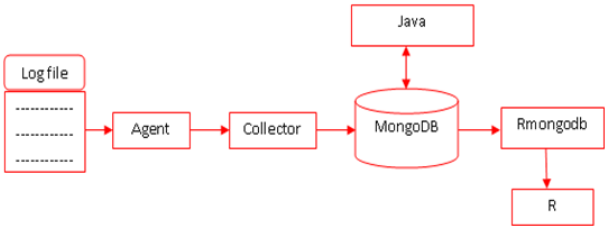

Figure 1. Automated model scheme.

4. Big Data Analysis

The agent is reading the web server logs from the server in the real time and the collector listens to the agent. The collector is continuously reading data from agent and writing the logs into a file. The client-server model is proposed for achieving the goal. The client server model is capable of reading data from multiple servers and writing data to multiple clients.

The Mongodb is integrated with java via mongo-java. jar and is used to store huge volumes of data. Its features are high scalability, high availability of data, replication, sharding, dynamic schema, aggregation framework, mapreduce. It can store text, image, audio and video and analysis can be performed using aggregation framework as well as with mapreduce. It can be integrated with many big data technologies such Hadoop, R, Splunk, GridGain, and Storm.

R language is for study of business data and using programming techniques to create a new opportunities and finding insights for achieving business goals. It is used for descriptive and predictive analysis. R is used for finding trend, outlier and pattern. It is used to generate reports, dashboards, predict future possibilities and to analyze data to find patterns.

4.1 Client-Server Model

The client-server model is implemented using java language and is explained as follows.

4.1.1 Procedure to Create Agent Class

- Create an agent class.
- Declare variables as static fields inside the class.
- The server listens to the request coming from the client.
- Agent is reading a file from the server.
- The server accepts the request from the client.
- The agent is sending the data to collector after reading from the server.

4.1.2 Procedure to Create Collector Class

- Create a Collector class.
- Create a file.
- Collecting data from the agent and writing it to a file.

4.2 MongoDB

To do a real time analysis of web server logs in data center using mongodb. Tracking of client location can be done using GeoIP. This GeoIP location information can be analyzed to find out the total number of visits from different country, real time information about server status of success, failure, warning and fatal errors. Real time updates of exact location of client who access sever and real time information of server status helpful to avoid repeating failure in a system.

4.2.1 Analysis of Log Data using MongoDB

The following are steps to analysing of log data using mongodb.

- Install mongodb
- Create database
- Create collections
- Import documents
- Analyze documents

4.3 R Language

R language is a programming language for statistical analysis, data mining and analysis of data. Integration of mongodb and R is done using a package called rmongodb. The features of R language are in memory processing, predictive model, efficient graphics, have packages for ease of use. Rmongodb is used for big data analysis and is an interface between mongodb and R environment. The analyzed results in R can be used to generate reports, dashboards and patterns.

5. Experimental Results

In this section, we analyzed web server logs using our proposed model.

5.1 Case Study 1: Web Server Logs

In case study 1, the web server logs are captured using client –server model and are imported into mongodb for analysis of logs. Finally analyzed data is given to R environment for visualizing data.

5.1.1 Starting an Agent

The agent is started to listen to the collector is shown in the Figure 2. The agent role is to capture logs in real time and to transfer it to the collector. The agent is reading logs from the server as soon as the logs are generated at the server side. The agent acts as an intermediate between server and collector. The agent can be one or more to capture the big data from different sources.

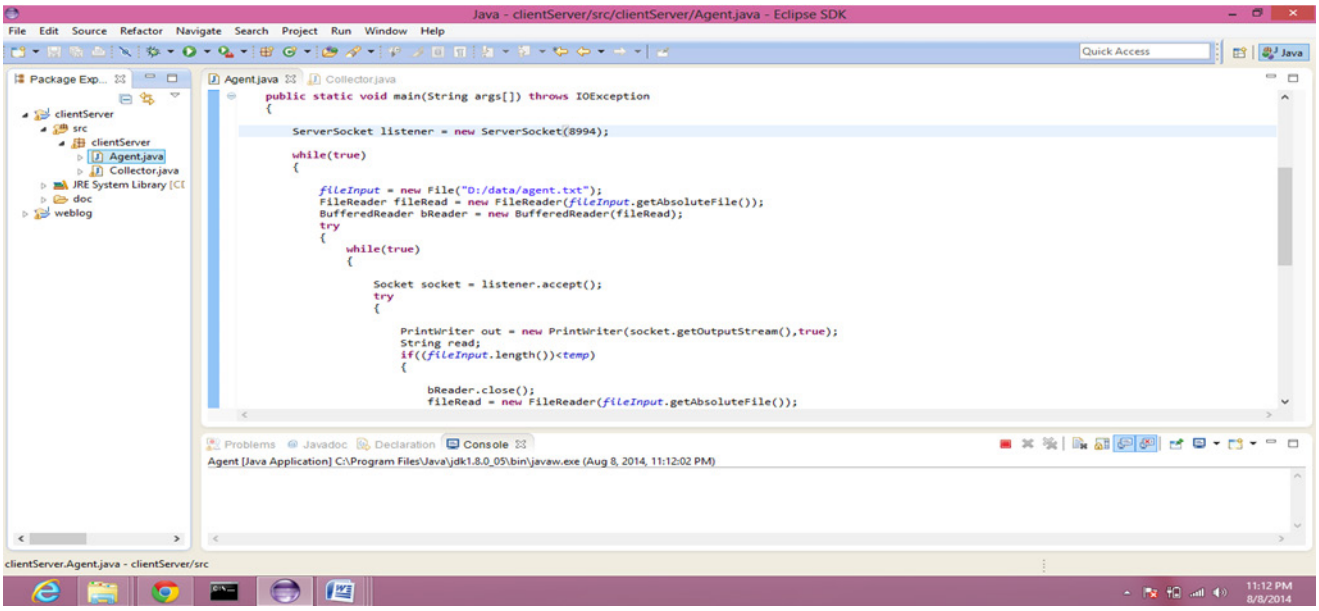

Figure 2. Starting an agent.

5.1.2 Starting a Collector

The collector is started to receive data from the agent and reading data from agent and writing it to a file in real time as soon as the logs are generated. The collector is reading logs continuously and the output is shown in the console as well as writing it into a file. The file is imported into mongodb database using the command in mongodb bin folder. The collector is writing each line from agent and also flushes the data at each time the data is written into a file. The client-server model is implemented in java and is integrated with mongodb for the analysis of big data. The collector module is writing a data to a file and showing the results in the Figure 3. The mongoimport command is used to import the logs from a file and is used for the further processing.

5.1.3 Staring a MongoDB Server

The mongodb server is started in console using the command mongod and also the database path is given to store the data in the respective directory. The mongodb server is started in console and is shown in the Figure 4.

5.1.4 Staring a MongoDB Client

The mongodb client is started and is shown in the Figure 5.

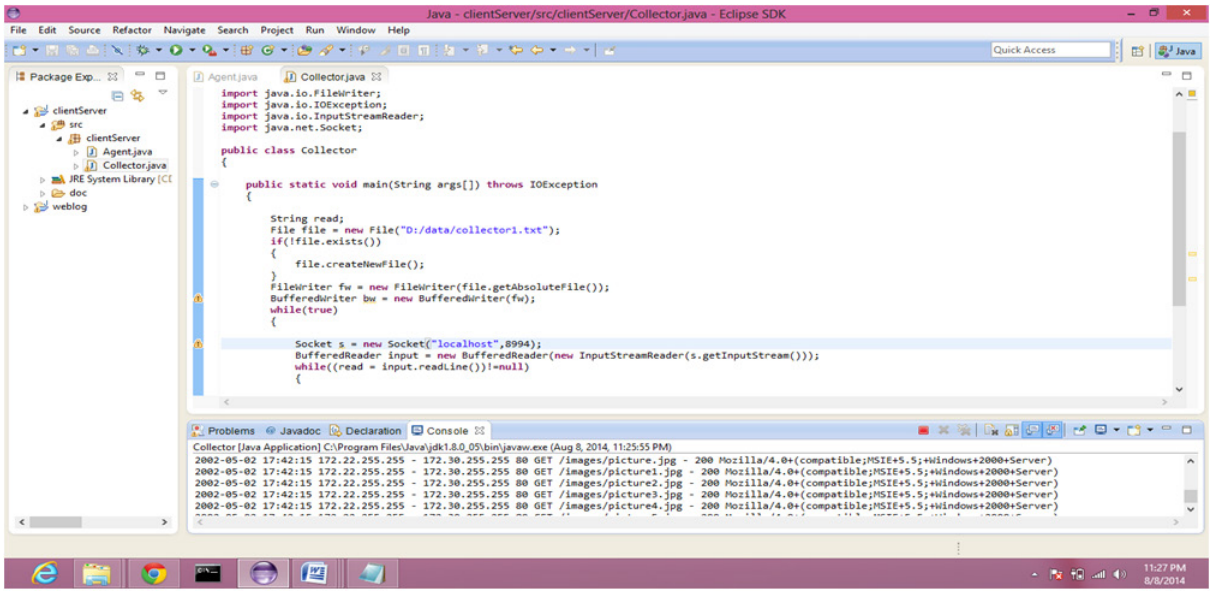

Figure 3. Starting a collector.

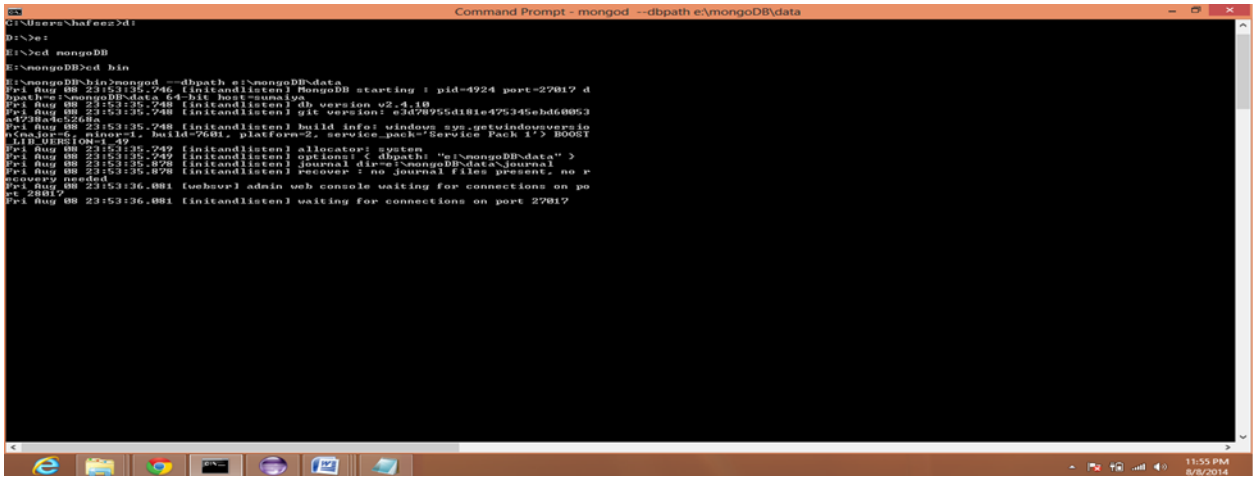

Figure 4. Starting a mongodb server.

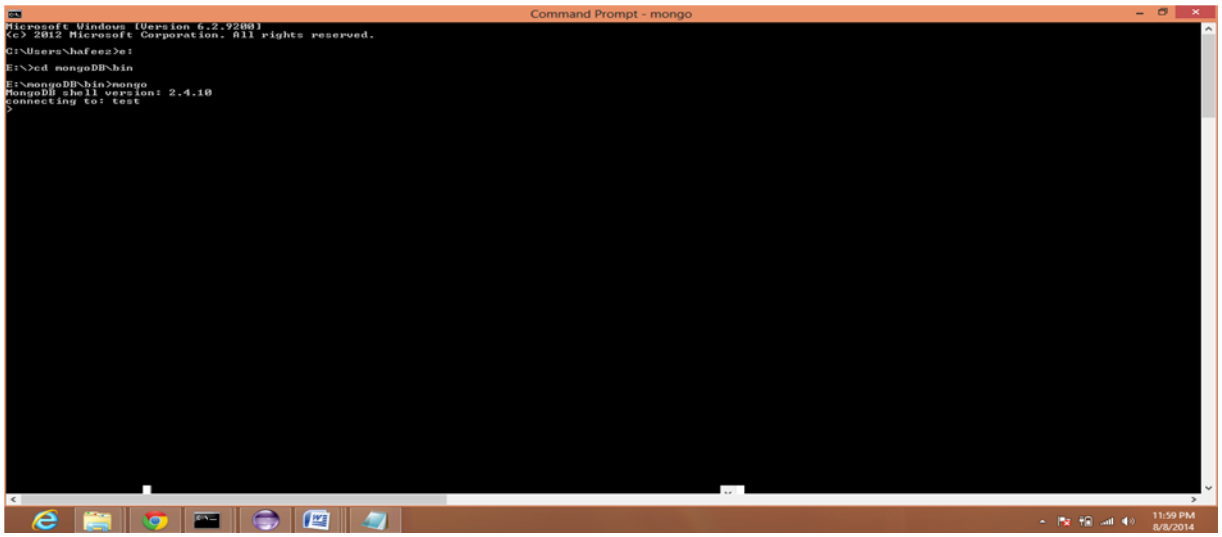

Figure 5. Starting a mongodb client.

5.1.5 Creating Database in MongoDB

The database LogFile is creating using the command use and it is shown in the Figure 6.

5.1.6 Creating a Collection in MongoDB

The collection LogDocument is creating using the command db.createCollection in the mongo shell and is shown in the Figure 7.

5.1.7 Importing a Documents in MongoDB

The documents are imported into collection called LogDocument and is imported using a command mongoimport.

5.1.8 Integrating MongoDB and R using Rmongodb Package

The rmongodb package is installed in the R environment

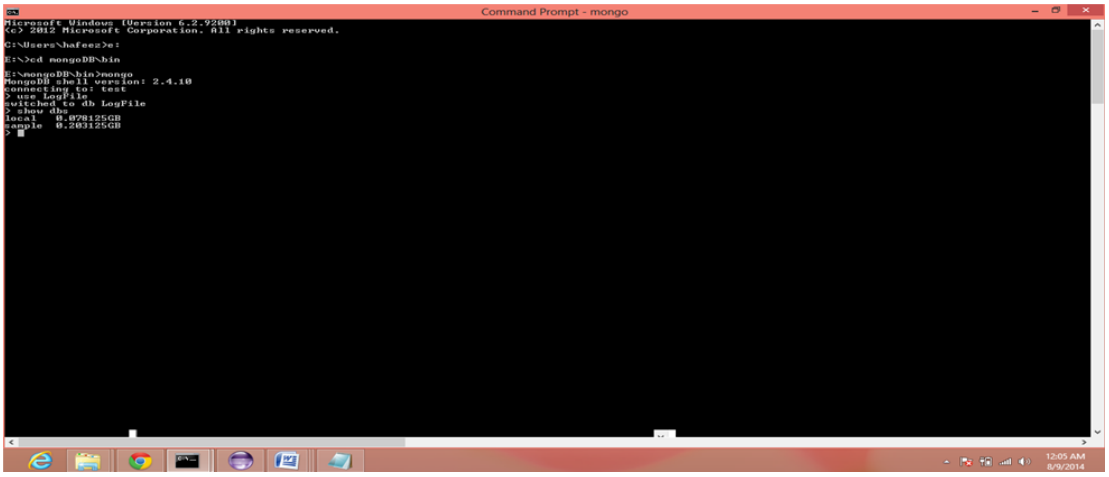

Figure 6. Creating a database in mongodb.

Figure 7. Creating a collection in mongodb.

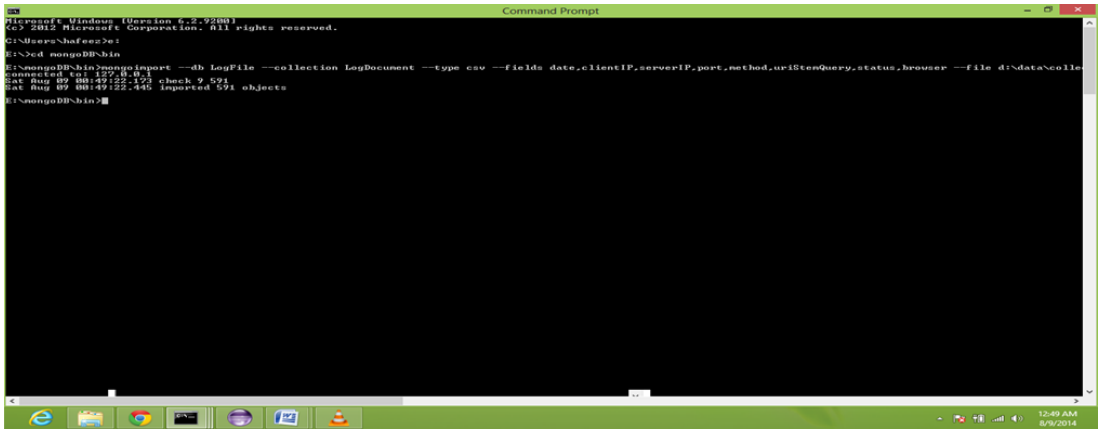

Figure 8. Importing a documents in mongodb.

for connecting mongodb with R to do a data analysis and generate reports and dashboards for predicting the patterns. The Figure 9 shows the connection between mongodb and R.

5.1.9 Data Analysis in R

The query is written to analyze the logs to return the GET method from the collection and is displayed in the R console. The analyzed results are shown in the Figure 10.

The numbers of clients visited to the server per day from different countries are shown in the Figure 11.

The query is written to display number of success and failure status in the Figure 12.

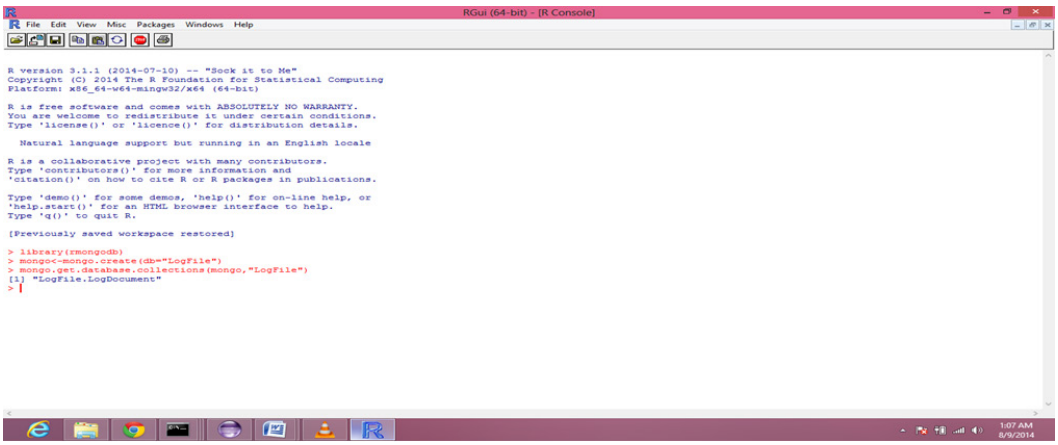

Figure 9. Integrating mongodb and R using rmongodb.

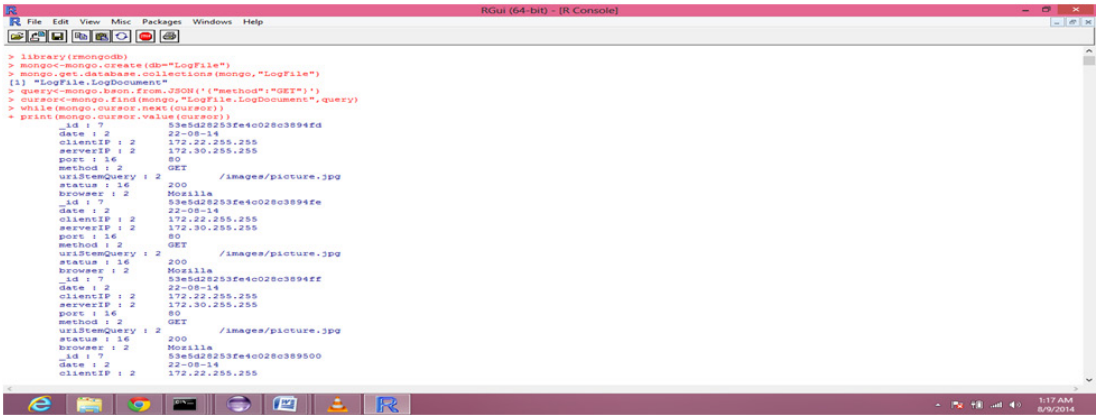

Figure 10. Display documents in R console.rmongodb.

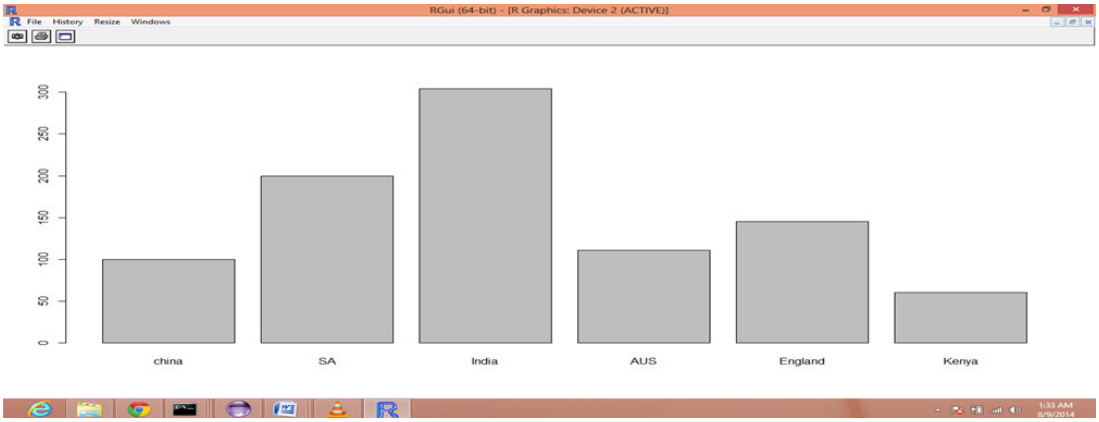

Figure 11. Visitors from different country.

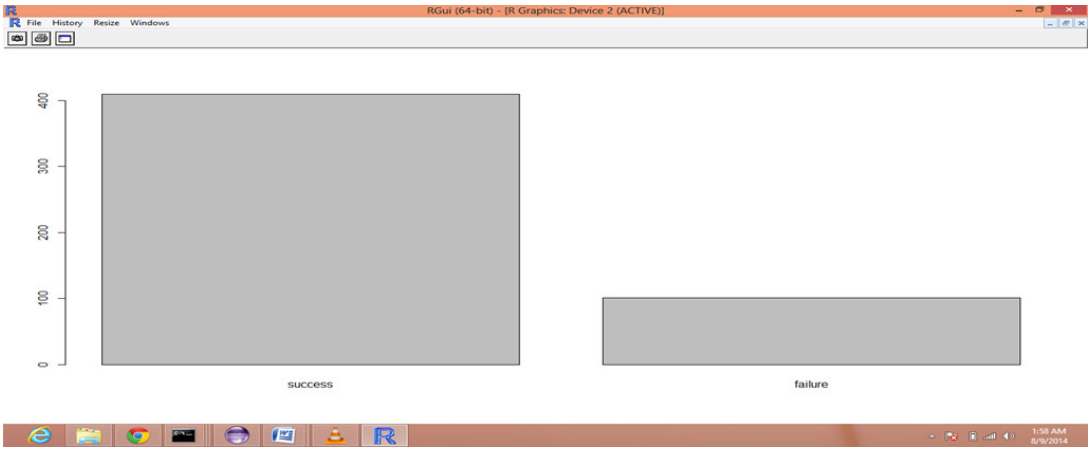

Figure 12. Number of success and failure status.

6. Conclusions and Future Work

The proposed model has following components such as client-server model, mongodb and R. The client-server component is reusable part and the entire component can be integrated with any framework. Mongodb is capable of replicate a logs and is analyzed by our automated model. For the future work, we have planned to integrate our automated model with stream data processing system and to compare our proposed model with existing framework. The cloud computing is moving the large number of application software and database to the distributed data centers, where the data is generating at the high rate and is managing to get insights out of it is still an complicated task. Analysis of image, audio and video is to be done using an efficient processing method and is to be integrated with our proposed model.

7. References

1. Kyoo-Sung N, Doo-Sik. Bigdata platform design and implementation model. Indian Journal of Science and Technology. 2015 Aug; 8(18). doi: 10.17485/ijst/2015/v8i18/75864.
2. Madden S. From databases to Big Data. IEEE Internet Computing. 2012 May; 16(3):4-6.
3. Sagiroglu S, Sinanc D. Big Data: A review. IEEE International Conference on Collaboration and Technology Systems; San Diego, CA. 2013 May 20-24. p. 42-7.
4. Choo J, Park H. Customizing computational methods for visual analytics with big data. IEEE Computer Graphics and Applications. 2013 July-Aug; 33(4):22-8.
5. Zhao Y, Wu J, Liu C. Dache: A data aware caching for big data applications using MapReduce Framework. Tsinghua Science and Technology TUP. 2014 Feb; 19(1):39-50.
6. Baquero VA, Palacios CR. Business process analytics using big data approach. IEEE IT professional. 2013 Nov-Dec; 15(6):29-35.
7. Hirzel M, Gedik B, Silva JG. IBM streams processing language: Analyzing big data in motion. IBM Journal of Research and Development. 2013 May/July; 57(3/4):7:1-11.
8. Biem A, Feng H, Turaga DS. Real-time analytics and management of big time-series data. IBM Journal of Research and Development. 2013 May/July; 57(3/4):8:1-12.
9. Zeng XQ, Li GZ. Incremental partial least squares analysis of big streaming data. Pattern Recognition. 2014 Nov; 47(11):3726-35.
10. Lakshmi M, Sowmya K. Sensitivity analysis for safe grain-storage using Big Data. Indian Journal of Science and Technology. 2015 Apr; 8(S7). doi:10.17485/ijst/2015/v8iS7/71225.
11. Marshall C. Big data, the crowd and me. ACM. 2012 July; 32(3-4):215-26.
12. Sukumar SR, Ferrell RK. Big Data' collaboration: Exploring, recording and sharing enterprise knowledge. ACM. 2013 July; 33(3-4):257-70.
13. Jeong Y-S. Parallel processing scheme for minimizing computational and communication cost of bioinformatics data. Indian Journal of Science and Technology. 2015 July; 8(15). doi:10.17485/ijst/2015/v8i15/76686.
14. Parthiban P. Analysis of Big Data using Mongodb. International Conference on Recent Trends in Engineering and Technology; Oxford Engineering College. 2014.
15. Sherly KK, Nedunchezhian R. A improved incremental and interactive frequent pattern mining techniques for market basket analysis and fraud detection in distributed and parallel systems. Indian Journal of Science and Technology. 2015 Aug; 8(18). doi:10.17485/ijst/2015/v8i18/55109.
16. Bifet A, Gavalda R. Mining frequent closed trees in evolving data streams. ACM Intelligent Data Analysis. 2011:29-48.
17. Dhamodaran S, Sachin KR, Rahul KS. Big Data implementation of natural disaster monitoring and alerting system in real time social network using hadoop technology. Indian Journal of Science and Technology. 2015 Sep; 8(22). doi:10.17485/ijst/2015/v8i22/79102.

18. Kim KW, Park WJ, Park ST. A study on plan to improve illegal parking using big data. Indian Journal of Science and Technology. 2015 Sep; 8(21). doi:10.17485/ijst/2015/v8i21/78274.
19. Kim BS, Kim DY, Kim KW, Park ST. The improvement plan for fire response time using Big Data. Indian Journal of Science and Technology. 2015 Sep; 8(23). doi:10.17485/ijst/2015/v8i23/79198.
20. Park HW, Yeo IY, Jang H, Kim NG. Study on the impact of Big Data traffic caused by the unstable routing protocol. Indian Journal of Science and Technology. 2015 Mar; 8(S5). doi:10.17485/ijst/2015/v8iS5/61480.
21. Tzanis G. Biological and medical Big Data mining. ACM International Journal of Knowledge Discovery in Bioinformatics. 2014; 4(1):15.
22. Nordin MI, Abdullah A, Hassan MI. Goal-based request cloud resource broker in medical application. World Academy of Science, Engineering and Technology-Kuala Lumpur; 2011 Sep 19-20. p. 1-5.
23. Doukas C. An open web services - based framework for data mining of biomedical image data. International Conference on Information Technology and Applications in Biomedicine; Larnaca. 2009 Nov 4-7. p. 1-5.
24. Huang Y, Hu CY, Zhao YW, Ma D. Web-based remote collaboration over medical image using web services. Global Information Infrastructure Symposium (GIIS'09); Hammemet. 2009 June 23-26. p. 1-8.
25. Wei W, Barnaghi PM. Semantic support for medical image search and retrieval. Biomedical Engineering. 2007 p. 2372-5.
26. Chang CC, Lu HM. Integration of heterogeneous medical decision support systems based on web services. IEEE International Conference on Bioinformatics and Bioengineering; Taichung. 2009 June 22-4. p. 415-22.
27. Ali S, Kiefer S. Semantic coordination of ambient intelligent medical devices - A case study. 3rd International ICST Conference on Pervasive Computing Technologies for Healthcare; 2009. p. 1-6.
28. Miri MP, Pooshfam H, Rajeswari M, Ramachandram D. A Web–based framework for distributed medical image processing using Image Markup Language(IML). 3rd UKSim European Symposium on Computer Modeling and Simulation; 2009 Nov 25-27. p. 470-5.
29. Kaldoudi E, Karaiskakis D. A service based approach for medical image distribution in healthcare Intranets. Comput Methods Programs Biomed. 2006 Feb; 81(2):117-27.