ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Data Modeling for Product Data Using Cassandra and Mongodb

G. Rajalakshmi[1], A. Ajitha[2], N. Mohan Prabhu[3]

M.E. Computer Science and Engineering, Mount Zion College of Engineering and Technology, Lena Vilakku, Thirumayam Tk, Pudukkottai, Tamil Nadu 622507, India.

Assistant Professor, Department of Computer Science and Engineering, Mount Zion College of Engineering and Technology, Lena Vilakku, Thirumayam Tk, Pudukkottai, Tamil Nadu 622507, India.

Assistant Professor, Department of Computer Science and Engineering, Mount Zion College of Engineering and Technology, Lena Vilakku, Thirumayam Tk, Pudukkottai, Tamil Nadu 622507, India.

ABSTRACT: In the present situation, purchasing has turned out to be on the web. Knowing the quality and unwavering quality of the item is a repetitive errand. Online audits are assisting customers to choose which items can be bought and furthermore organizations to comprehend the purchasing conduct of buyers. We propose to facilitate the circumstance by putting away the current audits in Mongo DB and Cassandra and examining them utilizing R.

KEYWORDS: Cassandra, MongoDB, Multiple interconnections, Distributed Environment, NoSQL.

I. INTRODUCTION

E-commerce is a transaction of buying or selling online. Online advertising, also called online marketing or internet advertising or web advertising is form of marketing and advertising which uses the internet to deliver promotional marketing messages to consumer. Consumers view online advertising as an unwanted distraction with few benefits and have increasingly turned to ad blocking for a variety of reasons.

The computational investigation of feelings, suppositions communicated in content is called Sentiment examination/conclusion mining. Generally, conclusions are communicated on anything including items, administrations, people, associations, occasions, points. The term complaint indicates an objective element to be remarked upon. A question has a parts set and a traits set, each having its own sub-segments and characteristics.

So question can be progressively disintegrated in light of a portion of connection. Conclusions pick up significance because of the way that at whatever point individuals decide, they have to hear others' sentiments and it is the same for associations. In any case, just constrained computational reviews on conclusions existed before the Web as there was minimal obstinate (content with suppositions/estimations) accessible. Introductory research began with slant and subjectivity grouping, regarding the issue as a content arrangement issue. Right now, Sentiment order manages obstinate record, (for example, item surveys) and characterizes a sentence as communicating positive or negative suppositions.

Subjectivity grouping figures out if a sentence is subjective/objective. Be that as it may, some genuine applications require a point by point examination as clients need to know sentiments on subjects[3]. A sentiment holder is a man/association communicating supposition. In item audits and web journals, assessment holders are for the most part the post's creators. Assessment holders are vital in news articles as they frequently plainly express that the individual/association holds a particular feeling.

Positive and negative are called sentiment introductions. Suppositions bear a noteworthy part in basic leadership. At the point when individuals need to settle on a decision, they are hear others' feelings and in the event that it includes devouring profitable assets like time as well as cash, individuals emphatically depend on associates' past

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

encounters. The move from a read-just to a read compose Web gave individuals new apparatuses permitting them to make/share their own particular substance, thoughts, and assessments with a great many individuals on the World Wide Web in an opportune and cost-productive way. The opportunity to catch popular's suppositions about item inclinations, get-togethers, political developments, promoting efforts, and organization methodologies made high enthusiasm for both mainstream researchers and business world.

II. SYSTEM OBJECTIVES

- To analyze from the product data the best companies and best brand.
- To make good decision on the purchases.
- To analyze the good and bad information.
-

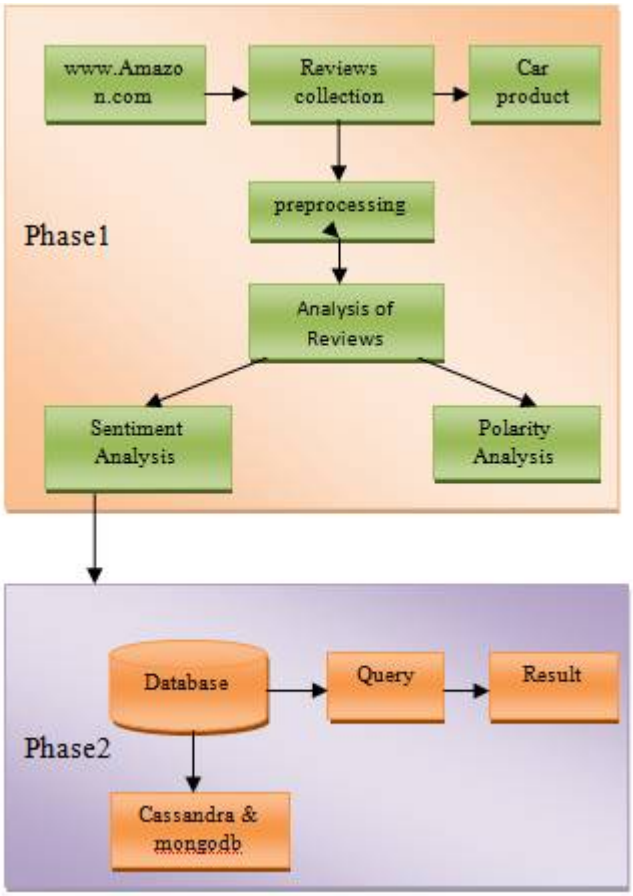

Fig.1 System Architecture Design

III. SYSTEM OVERVIEW

The customer reviews are extract aspect and mine whether given is positive or negative opinion.A review is given as input to data preprocessing[6]. Next, to analyze the sentiment and polarity from reviews.The product reviews are store to the databases.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Huge accumulations of shopper audits for items are currently accessible on the Web. These audits contain rich stubborn data on different items. They have turned into a profitable asset to encourage customers in comprehension the items before settling on buying choices, and bolster makers in grasping shopper conclusions to viably enhance the item offerings.

Notwithstanding, such audits are regularly sloppy, prompting to trouble in data route and learning procurement. It is wasteful for clients to accumulate popular sentiments on an item by perusing all the buyer audits and physically breaking down feelings on every survey. In this venture which intends to naturally recognize vital item angles from online shopper audits. The essential perspectives are recognized by two perceptions: the vital parts of an item are generally remarked by a substantial number of customers; and buyers' suppositions on the vital angles significantly impact their general feelings on the item. Specifically, given purchaser surveys of an item, we first recognize the item angles by ISODATA grouping and decide customers' suppositions on these viewpoints by means of an opinion classifier.

To decided the extremity of the client surveys of items. Framework plays out the item construct assessment mining with respect to the given surveys and the element insightful abridged outcomes produced by the framework will be useful for the client in taking the choice. Perspective based assessment mining is essential since these days everybody is occupied and they don't have an opportunity to peruse all the positive or negative audits in the event that somebody simply needs to think about some element of the item.

IV. REVIEWS COLLECTION

Opinion is person viewpoint about an object whereas mining is extraction of knowledge from facts or raw data. Thus; in another word it is a technique which detects intelligent information from data accessible on web. The people who express their opinion almost based on User Generated Content e.g. review sites, forums, discussion group, blogs, product etc. Based on above web site, we can collect user reviews about cars. The reviews may be Standard English format or Tamil language. If the reviews are Tamil are tanglish means converted into standard format using *GOOGLE API services.*

Online consumer reviews are prior to purchasing a product, while many firms use online consumer reviews as an important resource in their product development, marketing, and consumer relationship management. In this module we can collect consumer reviews for car products. The car products reviews are collected from the website amazon.com.1382 reviews are collected.

V. PREPROCESSING

In this module, we can eliminate stop words and stemming words[6]. In corpus linguistics, part-of-speech[6], also called grammatical tagging or words-category disambiguation, the process of word in a text as corresponding to a particular part of speech, based on both its definition and its context-i.e., its relationship with adjacent and related words in phrase, sentence, or paragraph. In computing, stop words are words which are filtered out before or after processing[6] of natural language data (text).

Though stop word is usually refer to the most common words in a language, there is no single universal list of words used by all natural language processing tools, and indeed not all even use such a list. In computational linguistics, a stem is the part of the word never changes even when morphologically inflected, and a lemma is the base form of the word.

Stemming words[6] are also removed from user reviews. Preprocessing is used to construct the structured data. In this module removing stop words from data collection eliminating stemming words[6] from review database. The reviews are imported then removing the stop words and stemming words[6].

VI. STOP WORD REMOVAL

Stop word removal[6] is used to remove unwanted words in each review sentence. Words like is, are, was etc .Reviews are stored in text file which is given as input to stopword like is, are, was etc. Reviews are stored in text file

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

which given as input to stop word removal. Stop words are collected and stored in a text file. Stop word is removed by checking against stop words list.

Stemming Words

Stemming[6] is used to form root word of a word. A Stemming algorithm[6] reduce the words ”longing”, ”longed” and ”longer” to the root word, ”long”. It consist many algorithms like n-gramanalysis, Affix stemmers and Lemmatization algorithms. Porter stemmer algorithm is used to form root word for given input reviews and store it in text file.

Analysis of Reviews

The reviews are analyze from two ways

- Sentiment Analysis
- Polarity Analysis
-

Sentiment Analysis

Sentiment analysis[3] is the process of determining whether the piece of writing is positive and negative words it is also known as opinion mining. Sentiment analysis is also known as opinion mining. Opinion mining works based on natural language processing and text analysis. Sentiment analysis[3] is widely applied to reviews and social media for a variety of applications, ranging from marketing customer service. Sentiment analysis aims to determine the attitude of a writer. Sentiment Analysis[3] works based on the identified emotions such as angry, sad, and happy. A sentiment analysis[3] is classifying the polarity Analysis of a given text at the document and sentence.

The sentiment analysis[3] is retrieve from the car product reviews.

Polarity Analysis

Polarity analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level-whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. The polarity analysis is retrieve from the car product reviews.

Term Document Matrix

A term-document matrix is also known as document-term matrix which is also shortly called as TDM. Document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms

Requirements

R Version 2.15.3 with packages like tm, Rstem, sentiment, SnowballC, NLP, ggplot2, word cloud bit32 with all of its dependency packages.

- R Studio Version 0.99.903

MongoDB

MongoDB[5][12] is a a document oriented database that is easily scalable and highly available .It provides high performance and works on the concept of collections and documents. A collection in MongoDB[5][12] holds multiple documents. MongoDB[5][12] has a flexible schema and consequently, the number of field in a document with in collection need not be the same. mongoDB[5][12] is document oriented database has a rich query language. MongoDB[5][12] is available under the Apache license.

The advantage of mongodb is structure of single object is clear, no complex join deep queries on document using a document based query, conversion of application object to database object not needed.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Cassandra

Cassandra[11] is an open source NoSQL[1] distributed database written in java. it is an ideal fit to maintain huge amount of structure as well as unstructured data because of its ability to scale elastically as well as linearly. Because of Cassandra is linear elasticity, the performance increase with the increase in the number of nodes in the cluster.

Easy data distribution by replicating data across several data center is supported by Cassandra[11]. Cassandra[11] also support ACID properties like the relational database and has very fast write speed. Cassandra[11 falls into the category of being a key-value database because it is eventually consistent like Amazon's.

VII. EXISTING SYSTEM

In recent internet application is which have focused on detecting the polarity of the text, our text classifier helps users distinguish between positive and negative reviews thus assisting the user with opinion mining[11]. This could be very useful for web applications like product web site, where the user has to face large chunks of raw data.

To classify opinion an unsupervised lexicon technique is used to sentiment classification. There are so many user generated opinions on the web for a product; it may be difficult to know how many opinions are positive or negative. It makes tough to take decision about the product purchasing. A sentence level opinion mining is used and it is done by counting based approach which compare the opinion by count method.

To focus on mining product feature that reviewer has to comment. MLBSL technique is used to improvement over manual and automatic-built lexicons like Senti-WordNet[3] is used. To collect a data by the real-time stream there is no language, or any other kind of restriction was made during the streaming process.

In fact, their collection consists of reviews in English languages. It analyzing a solution for sentiment analysis[3] level, using sentence level opinion mining .The classification of opinion mining techniques[11] that conveys user's opinion at various levels. The precise method for predicting opinions enable us, to extract sentiments from the web and foretell online customer's preferences.

Disadvantages

- Manually developed for each domain
- The works do not exploit the fact that majority of the reviews have a lot of domain independent components.

VIII. PROPOSED METHODOLOGY

The proposed system uses customer reviews are extract and mine whether given is positive or negative opinion. Each review is split into individual sentences. A review sentence is given as input to data preprocessing. Next, it extracts aspect in each review sentence .Stop word removal[6], stemming and pos tagging are data preprocessing. Sentiment orientation is used to identify whether it is positive or negative opinion sentence. Then it identifies the number of positive and negative opinions ..

Nowadays, there are the several websites that allow customers to buy and post reviews of purchased products, which results in incremental accumulation of a lot of reviews written in natural language. Moreover, conversance with E-commerce[7] and social media[7] has raised the level of sophistication of online shoppers and it is common practice for them compare competing brands of products before making a purchase. Prevailing factors such as availability of the online reviews and raised end-user expectations have motivated the development of opinion mining systems that can automatically classify and summarize users' reviews. This project proposes an opinion mining system that can be used for sentiment classifications of the user reviews. Feature-based sentiment classification is a multistep process that is involves preprocessing to remove noise, extraction of features and corresponding descriptors, and tagging their polarity. To classify the Analysis of reviews from trained opinion words and the product reviews to users. Product reviews are store to Cassandra[11] and mongoDB[5] databases. The store data can be analyzed.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

VIII. LITERATURE SURVEY

In the year of 2015, the authors "Alexandra Cernian, Valentin Sgarciu and Bogdan Martin" described in paper titled "Sentiment analysis from product reviews using SentiWordNet as lexical resource" such as Customers make their decisions based mostly on the opinion of other consumers. On the other side, companies need quick feedback from their customers in order to adapt to their needs in real time. The effective connection between the two aspects relies on opinion mining tools, which automatically process consumers' reviews and opinions about products or services.

This paper presents to a semantic approach for a sentiment analysis application, which is based on using the SentiWordNet lexical resource. The experimental validation proved that 61% average rate of success of the application.

In the year of 2015, the authors "GauravDubey, GauravDubey, Naveen Kumar Shukla" described into their paper titled "User Reviews Data Analysis using Opinion Mining on Web" such as The product reviews guide the customers and help them in making decisions regarding various available products this makes the automatic extraction, summarization, and tracking of the available opinions very beneficial for the customers looking to buy a product. Through this paper our implementation in mobile domain will be based on three main steps: Applying Part-of speech tagging (POST), Rule-Mining and identifying opinion words, Summarizing and displaying the end results.

In the year of 2015, the authors "Rajni Singh and Rajdeep Kaur" described into their paper titled "Sentiment Analysis on Social Media and Online Review" such as Social media monitoring has been growing day by day so analyzing of social data plays an important role in knowing customer behavior .So analyzing Social data such as Twitter Tweets using sentiment analysis which checks the attitude of User review on movies.

This paper develops a combined dictionary based on social media keywords and online review and also find hidden relationship pattern from these keyword.

IX. EXPERIMENTAL RESULTS

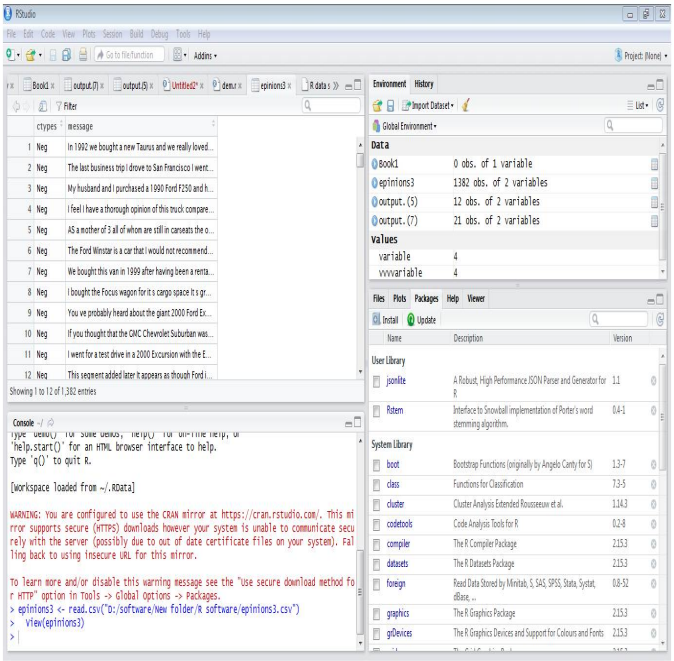

Fig.2 Data Extraction in R

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

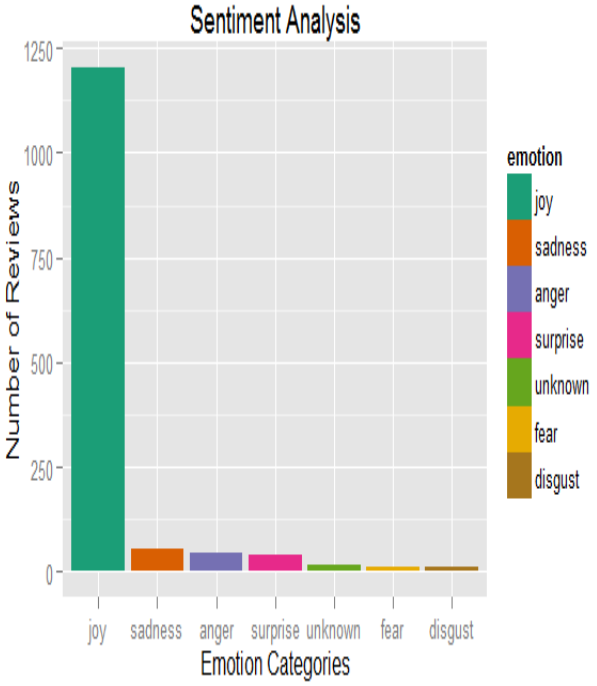

Fig.3. Sentiment Analysis

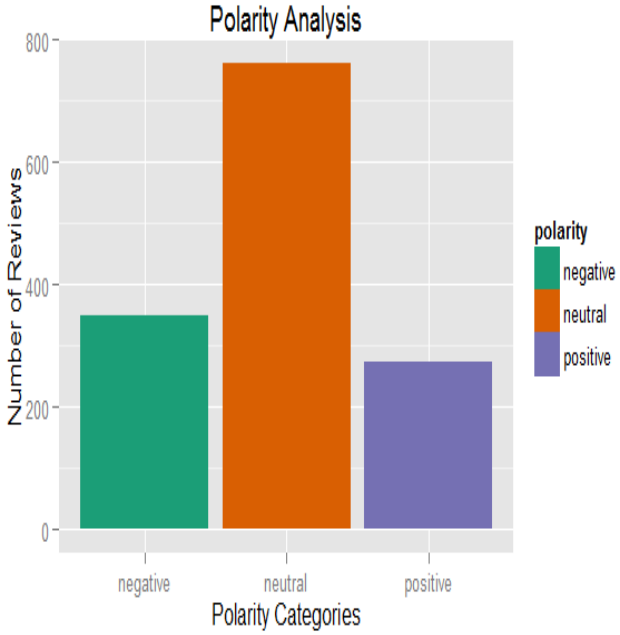

Fig.4. Polarity Analysis

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

X. CONCLUSION AND FUTURE SCOPE

In this paper we have identified whether the sentence is positive or negative opinion and also the number of positive and negative opinion of each extracted. We have determined the polarity of the customer reviews of car products. Product reviews are stored to Cassandra and mongoDB databases and analyzed. Cassandra provides easy scalability. MongoDB, owing to its rich query language and variable schema is also a good choice for storing the product data.

MongoDB, being a document oriented database, has a rich query language and a flexible schema design. A future scope of this work would be the implementation of the third model in MongoDB, directly or indirectly.

REFERENCES

[1] April Reeve, Big Data and NoSQL: The Problem with Relational Databases, infocus.emc.com, September 2012.
[2] Brewer, E. (2012). CAP twelve years later: How the" rules" have changed. Computer, 45(2), 23-29..
[3] Alexandra Cernian,Valentin Garcia, Bogdan Martin." Sentiment analysis from product reviews using SentiWordNet as lexical resource" ECAI 2015-International Conference-7thEdition Electronics, Computers and Artificial Intelligence 25 June ,27 June 2015, Bucharest, ROMÂNIA. 2015 IEEE.
[4] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., ... & Vogels, W. (2007, October). Dynamo: amazon's highly available key-value store. In ACM SIGOPS Operating Systems Review (Vol. 41, No. 6, pp. 205-220). ACM.
[5] MongoDB Documentation, docs.mongodb.org
[6] A.Jeyapriya,C.S.Kanimozhi Selvi" Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm"IEEE sponsored 2nd international conference on electronics and communication systems(ICECS '2015)
[7] Rajni Singh, Rajdeep Kaur "Sentiment Analysis on Social Media and Online Review ".International Journal of Computer Applications(0975 – 8887) Volume 121 – No.20, July 2015.
[8] Leavitt, N. (2010). Will NoSQL databases live up to their promise?. Computer, 43(2), 12-14.
[9] Han, J., Haihong, E., Le, G., & Du, J. (2011, October). Survey on NoSQL[1] database. In Pervasive computing and applications (ICPCA), 2011 6th international conference on (pp. 363-366). IEEE.
[10] Cattell, R. (2011). Scalable SQL and NoSQL data stores. ACM SIGMOD Record, 39(4), 12-27.
[11] Wang, G., & Tang, J. (2012, August). The NoSQL principles and basic application of cassandra model. In Computer Science & Service System (CSSS), 2012 International Conference on (pp. 1332-1335). IEEE.
[12] Chodorow, K. (2013). MongoDB: the definitive guide. " O'Reilly Media, Inc."
[13] Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. International Journal of Web Information Systems, 9(1), 69-82.
[14] Schwinger, W., Retschitzegger, W., Schauerhuber, A., Kappel, G., Wimmer, M., Pröll, B., ... & Zhang, G. (2008). A survey on web modeling approaches for ubiquitous web applications. International Journal of Web Information Systems, 4(3), 234-305.
[15] Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. Journal of Cloud Computing: Advances, Systems and Applications, 2(1), 22.
[16] Brewer, E. A. (2000, July). Towards robust distributed systems. In PODC (Vol. 7).
[17] Moniruzzaman, A. B. M., & Hossain, S. A. (2013). NoSQL database: New era of databases for big data analytics-classification, characteristics and comparison. arXiv preprint arXiv:1307.0191.
[18] Angles, R., & Gutierrez, C. (2008). Survey of graph database models. ACM Computing Surveys (CSUR), 40(1), 1.
[19] Reutter, J. L., & Tan, T. (2011). A Formalism for Graph Databases and its Model of Computation. In AMW.
[20] Hecht, R., & Jablonski, S. (2011). NoSQL[1] evaluation: A use case oriented survey, pp. 336-341.

BIOGRAPHY

Ms. Rajalakshmi. G. Studying M.E. Computer Science and Engineering in Mount Zion College of Engineering and Technology, which is located in Lena Vilakku, Pudukkottai, TamilNadu, India.