IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

# Improved Analysis of Unstructured Datasets using Thesaurus Model

**Simy Mary Kurian[1*], Neema George[2], Jinu P Sainudeen[3], Neethu Maria John[4]**

[1,2,3,4]Department of Computer Science& Engineering, Mangalam College of Engineering, Kerala, India

*Corresponding Author: simy.kurian@mangalam.in, Tel.:+91 9656294800

DOI: https://doi.org/10.26438/ijcse/v7i2.10331037 | Available online at: www.ijcseonline.org

Accepted: 19/Feb/2019, Published: 28/Feb/2019

***Abstract***— Humankind has put away in excess of 295 billion gigabytes (or 295 Exabyte) of information beginning around 1986, according to a report by the University of Southern California. Putting away and checking this information in generally disseminated conditions for all day, every day is an enormous errand for worldwide assistance associations. These datasets require high handling power which can't be presented by conventional information bases as they are put away in an unstructured arrangement. Although one can utilize Map Reduce worldview to take care of this issue utilizing java-based Hadoop, it can't give us with most extreme usefulness. Downsides can be defeated utilizing Hadoop-streaming methods that permit clients to characterize non-java executable for handling this dataset. This paper proposes a THESAURUS model which permits a quicker and more straightforward form of business examination.

***Keywords*—**Hadoop, MapReduce, HDFS, NoSQL

## I. Introduction

Information has never been more essential to the business world as it has turned into a fundamental resource as significant as oil and similarly as hard to mine, model and make due. The volume and veracity of the datasets that are being put away and dissected by the business are unforeseeable and the customary advances for information the board, for example, social data sets can't meet the ongoing business needs. Bigdata advancements assume an indispensable part to resolve this issue. Early thoughts of huge information came in 1999 and at present it turns into an unavoidable peculiarity device through which we oversee business and administration. For a layman the possibility of Bigdata might connect with pictures of turbulent monster distribution centers stuffed office space with various staffs managing immense number of pages and accompanied exhausting proper records under oversight of some old civil servant. In actuality working of Bigdata is straightforward and all around organized, yet interesting to the point of presenting new difficulties and open doors even to specialists of industry. It gives equal handling of information in many machines that are circulated geologically.

In the present information focused world Hadoop is considered as the primary specialist of enormous information innovation because of its open source nature. Anyway as it is a java based environment, it made obstacle for developer from non-java foundation. To resolve this issue it has worked with an instrument, 'Hadoop-Streaming' by empowering adaptability in programming with successful equal computability.

Data has never been more important to the business world as it has become a vital asset as valuable as oil and just as difficult to mine, model and manage. The volume and veracity of the datasets that are being stored and analyzed by the business are unforeseeable and the traditional technologies for data management such as relational databases cannot meet the current industry needs. Bigdata technologies play a vital role to address this issue. Early ideas of big data came in 1999 and at present it becomes an unavoidable phenomenon tool through which we manage business and governance. For a layman the idea of Bigdata may relate to images of chaotic giant warehouses over crowded office space with numerous staffs working through huge number of pages and come with boring formal documents under supervision of some old bureaucrat. On the contrary working of Bigdata is simple and well structured, yet exciting enough to pose new challenges and opportunities even to experts of industry. It provides parallel processing of data in hundreds of machines that are distributed geographically. Necessity of Bigdata arises under the obligation of the following:

1. When existing technology is inadequate to perform data analysis.
2. In the case of handling more than 10TB of dataset.
3. Relevant data for an analysis present across multiple data stores which are filed in multiple formats.
4. When steaming data have to be captured, stored andprocessed for the purpose of analysis.
5. When SQL is inefficient for high level querying.

In today's data centered world Hadoop is considered as the main agent of big data technology due to its open source nature. However as it is a java based ecosystem, it created hurdle for programmer from non-java background. To address this issue it has facilitated a tool, 'Hadoop-Streaming' by enabling flexibility in programming with effective parallelcomputability

Well, it helps the organization to harness their transactional data and use it to identify new opportunities in a cost effective and efficient manner. Primary aim of data analysis is to glean actionable logic that helps the business to tackle the competitive environment.

This will alert the business for their inevitable future by introducing new products and services in favor of the customers. Unfortunately for the matter of convenience 80% of the business oriented data are stored in an unstructured format.

Structured data usually resides in a relational database with predefined structures so converting the data to different models and analyzing them seems mundane. Here the role of Hadoop- Streaming arises which works on a Map and Reduce paradigm by analyzing the unstructured data and presents viable business logic. enabling flexibility in programming with effective parallel computability.

## II. RELATED WORK

The inquiry that experiences a youngster is that why one purposes unstructured dataset when there is generally a chance of utilizing organized information. At the start of figuring, the term stockpiling related just plain texts. Presently client requirements to store more extravagant substance than plain message. Rich information type incorporates pictures, films, music, x-beams ,etc.It gives prevalent client experience to the detriment of extra room. Apache Hadoop [1] is open source programming for dependable, versatile and dispersed processing. Hadoop system permits conveyed handling of enormous datasets across low level ware equipment utilizing straightforward programming models.

This system is motivated by Google's MapReduce structure in which application is separated into various little parts and each part can be run in any hub in the group. Hadoop contains two significant parts - a particular record framework called Hadoop Distributed File System (HDFS) and a Map Reduce structure. Hadoop chips away at partition and vanquish guideline by carrying out Mapper and Reducer in the structure. Mapper work divides the information into records and converts it into (key,value) matches.

The question that encounters a rookie is that why one uses unstructured dataset when there is always a possibility of using structured data. At the outset of computing, the term storage corresponded only plain texts. Now user needs to store richer content than plain text. Rich data type includes pictures, movies, music, x-rays ,etc.It provides superior user experience at the expense of storage space. Such data sets are called unstructured because they contain data that do not fit neatly in a relational database. Industry came up with a third category called semi structured data which resides in a relational database, similar to structured data. However it does not have some organizational property necessary to make them easy to be analyzed.(Eg.XML doc).

A NOSQL database [4] provides mechanism for storage and retrieval of data which is modeled in contrast to the tabular relations used in relational databases. It become common in the early twenty first century when the industrial requirements triggered a need of database structures that support query languages other than SQL.(called “Not only SQL”, non SQL).This is mostly used in big data and real-time applications as it provides simpler design, horizontal scalability and high availability.

Apache Hadoop is open source software for reliable, scalable and distributed computing. Hadoop framework allows distributed processing of large datasets across low level commodity hardware using simple programming models. This framework is inspired by Google’s MapReduce structure in which application is broken down into numerous small parts and each part can be run in any node in the cluster.

Hadoop contains two major components - a specific file system called Hadoop Distributed File System (HDFS) and a Map Reduce framework. Hadoop works on divide and conquer principle by implementing Mapper and Reducer in the framework. Mapper function splits the data into records and converts it into (key,value) pairs. Before feeding the output of the Mappers to Reducer an intermediate Sort and Shuffle phase is implemented in the MapReduce framework to reduce the work load at Reducer machine.

The sorted (key,value)pair is given into Reducer phase. The Reducer function does the analysis of the given input and the result will be loaded to HDFS(eg.The maximum temperature recorded in a year, positive and negative ratings in a business etc.).The analyst has to develop Mapper and Reducer functions as per the demand of the business logic.

A NOSQL information base [4] gives system to capacity and recovery of information which is demonstrated rather than the even relations utilized in social data sets. It become normal in the mid twenty first century when the modern prerequisites set off a need of data set structures that help inquiry dialects other than SQL.(called "Not only SQL", non SQL).This is generally utilized in enormous information and ongoing applications as it gives less complex plan, flat versatility and high accessibility. The most famous NOSQL data sets are MongoDB, Apache Cassandra [3], Datastax, Redis.

Hadoop Streaming (see Figure 1) is an API given by Hadoop which permits client to compose MapReduce capacities in dialects other than java[2]. Hadoop Streaming purposes Unix standard streams as the point of interaction among Hadoop and our MapReduce programs, so the client has the opportunity to utilize any dialects (Eg. Python, Ruby, Perl and so forth) that can peruse standard info and keep in touch with standard result.

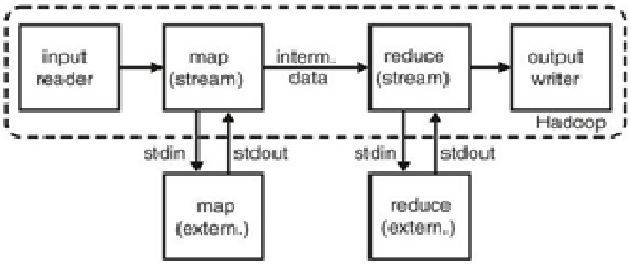

Figure 1 : Hadoop Streaming

Because of the challenges in dissecting the unstructured information associations have gone to various different programming answers for search and concentrate essential data. No matter what the stage utilized, the examination should embrace three significant advances information assortment, information decrease, information investigation [5][6][7]

In Data Collection stage the datasets to be examined can be gathered through two strategies. Initially, information can be downloaded from various hubs containing the predefined records to HDFS.Data reduction mainly deals with the unstructured dataset got accessible, examination cycle can be sent off. It includes cleaning the information, separating significant elements from information, eliminating copy things from the datasets, changing over information designs, and some more. In data analysis stage the preprocessed information is considered to recognize the secret example. Hadoop gives a Mahout apparatus that carries out adaptable AI calculations which can be utilized for cooperative separating, grouping and characterization .The examined information then, at that point, can be envisioned by the necessity of the business utilizing Tableau, Silk, CartoDB, Datawrapper.

## III. METHODOLOGY

Why Big information examination? Indeed, it assists the association with tackling their value-based information and use it to recognize new open doors in a financially savvy and effective way. Essential point of information examination is to gather significant rationale that assists the business with handling the serious climate.

In this stage the datasets to be analyzed can be collected through two methods. Firstly, data can be downloaded from different nodes containing the specified records to HDFS. Alternatively it can be done by connecting to the local servers containing the records. The former can be achieved by tools such as Sqoop, Flume and the latter using Apache Spark[6]. In a real time environment the streaming datasets can be accessed using standard public key encryption technique to ensure authenticity.

Once the unstructured dataset got available, analysis process can be launched. It involves cleaning the data, extracting important features from data, removing duplicate items from the datasets, converting data formats, and many more. Huge datasets are minimized into structural and more usable format using series of Mapper and Reducer functions. This is done by projecting the columns of interest and thus converting it in a format which will be adaptable for final processing. Cleaning text is extremely easy using R language, whereas Pig and Hive supports high level abstraction of data preprocessing.

Before the inception of Bigdata technologies collecting, preprocessing and analyzing terabytes of data was considered impossible. But due to the evolution of Hadoop and its supporting framework the data handling and data mining process seems not so tedious. Programmer with the help of Hadoop Streaming API can write the code in any language and work according to the domain of user. In this stage the pre processed data is studied to identify the hidden pattern.

The underline motivation behind this model is the lack of knowledge base in the existing analysis framework which in turn causes the system to follow some unnecessary repetition. Consider an analysis problem to find the maximum recorded temperature in last 5 years. So the analysis is done by

1. Collecting the data from National Climatic DataCenter [5] and store in HDFS.
2. Project the field which contains the temperaturedata i.e. the column of interest.
3. Store the preprocessed result in HDFS.
4. Find the maximum temperature reported by analyzing the (key, value) pair.
5. Store the final result in HDFS.

This will alarm the business for their inescapable future by presenting new items and administrations for the clients. Tragically for the question of comfort 80% of the business arranged information are put away in an unstructured organization. Organized information as a rule lives in a social data set with predefined structures so changing the information over completely to various models and dissecting them appears to be unremarkable. Here the job of Hadoop-Streaming emerges which deals with a Map and Reduce worldview by dissecting the unstructured information and presents reasonable business rationale.

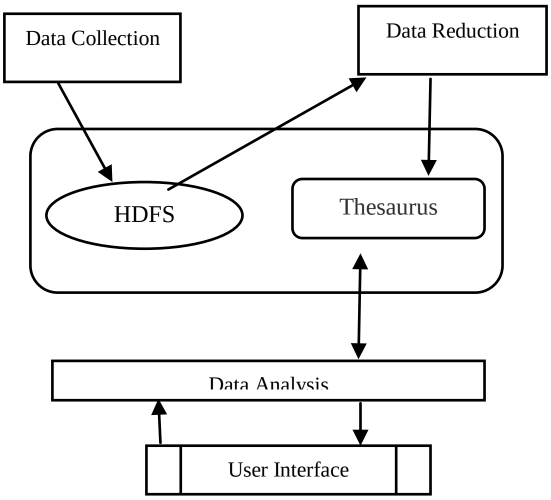

Figure 2. Proposed Architecture

Hadoop provides a Mahout tool that implements scalable machine learning algorithms which can be used for collaborative filtering, clustering and classification .The analyzed data then can be visualized according to the requirement of the business using Tableau, Silk, CartoDB, Datawrapper

Thus the whole process of analysis can be explained in a five step workflow:

1. Collecting the data from alien environment and keep it inside the Hadoop Distributed File System.
2. Apply set of MapReduce tasks to the step one collected data and project the columns of interest based on the user query.
3. Keep the preprocessed data in HDFS for further analysis.
4. Use the preprocessed data for analyzing the pattern of interest.
5. Store the result in HDFS so that with the help of visualization tools user can selectively adopt the method of presentation.

The underline inspiration driving this model is the absence of information base in the current examination structure which thus makes the framework follow some pointless reiteration. Consider an examination issue to track down the most extreme kept temperature in most recent 5 years.

When the informational collection is changed over into a primary configuration the pattern of the dataset ought to be indicated by the pre-processing developer so the examiner need not run over the difficulty of understanding the recently made informational index. This pre-processed dataset can supplant the old datasets with the goal that the pointless stockpiling issue is dealt with by the model.

## IV. RESULT ANALYSIS

The working of the framework is determined in two stages, one for assortment and pre-processing, and second for investigation. In the principal stage the vital information which can be examined are gathered and pre-processed. This information is then put away in the thesaurus module in HDFS and made it accessible for the client to dissect in light of the business needs. Thesaurus contains the organized information as well as the construction of the information stockpiling. In stage two, the expected question can be tended to by alluding the construction. Thus, examiner need not consider the issues of unstructured information gathered by the framework.

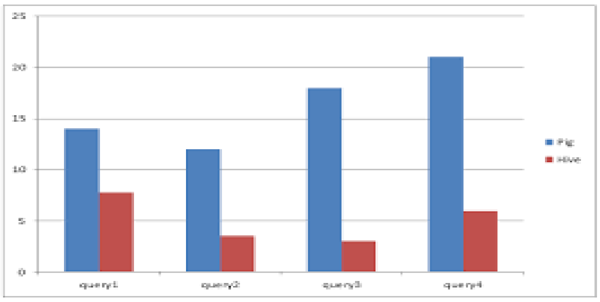

Figure 3 Comparitive analysis

Mining the inner pattern of business invokes the related trends and interests of the customers. This can be achieved by analysing the streaming datasets generated by the customers in each point of time.Hadoop provides flexible architecture which enables industrialist and even starters to learn and analyse this social changes.Hadoop-Streaming is widely used for sentimental analysis using non-java executables.Also proposed a THESARUS model which works in a time and cost effective manner for analysing these humongous data. Future scope is to enable the efficiency of the system by developing a THESARUS model which is suitable to analyse terabytes of data and returns with the relative experimental results.

## V. CONCLUSION

Mining the internal example of business summons the connected patterns and interests of the clients. This can be accomplished by investigating the streaming datasets created by the clients in each place of time.Hadoop gives adaptable design which empowers industrialist and even starters to learn and break down this social changes.Hadoop-Streaming is generally utilized for wistful examination utilizing non-java executables.Also proposed a THESARUS model which works in a period and savvy way for dissecting these humongous information. Future extension is to empower the effectiveness of the framework by fostering a THESARUS model which is appropriate to investigate terabytes of information and gets back with the overall exploratory outcomes

## REFERENCES

[1] Apache Hadoop.[Online].Available:http://hadoop.apache.org

[2] Apache Hadoop-Streaming.[Online].:http://hadoop-streaming.apache.org

[3] Cassandra wiki, operations. [Online]. Available: http://wiki.apache.org/cassandra/Operations

[4] NOSQL data storage [online]: http://nosql-database.org

[5] E. Dede, B. Sendir, P. Kuzlu, J. Weachock, M. Govindaraju, and L. Ramakrishnan, "*A processing pipeline for cassandra datasets based on Hadoop streaming*," in Proc. IEEE Big Data Conf., Res. Track, Anchorage, AL, USA, pp. **168–175,2014.**

[6] E. Dede, B. Sendir, P. Kuzlu, J. Weachock, M. Govindaraju, L. Ramakrishnan, *"Processing Cassandra Datasets with Hadoop-Streaming Based Approaches"*,IEEE Transactions on Services Computing, Vol. **9**,Issue **1**,pp **46-58.**

[7] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae,J. Qiu, and G. Fox, "*Twister: A runtime for iterative mapreduce*," in Proc. 19th ACMInt. Symp. High Perform. Distrib. Comput., pp. **810–818**,**2010**

## AUTHORS PROFILE

*Ms.Simy Mary Kurian* Assistant Professor , Department of Computer Science and Engineering, Mangalam College of Engineering, Kerala, India since 2011.She has completed B.Tech in Computer Science and Engineering from Mahatma Gandhi University and M.Tech in Software Engineering from Karunya Institute of Technology and Science. Her research interest include Image Processing, Data Science, Artificial Intelligence and Bio-inspired Computing .She has associated with many number of undergraduate and research projects.

.*Ms.Neema George* Assistant Professor , Department of Computer Science and Engineering, Mangalam College of Engineering, Kerala, India since 2008.Her research interest include Image Processing, Data Science, Artificial Intelligence and Cloud Computing .She has associated with many number of undergraduate and research projects.

*Ms.Jinu P Sainudeen* Assistant Professor , Department of Computer Science and Engineering, Mangalam College of Engineering, Kerala, India since 2006 . Her research interest include Artificial Intelligence, Machine Learning, Deep Learning. She has associated with many number of undergraduate and research projects.

.

*Ms.Neethu Maria John* Assistant Professor , Department of Computer Science and Engineering, Mangalam College of Engineering, Kerala, India since 2010.

.