IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Integrated Tamil News Storage and Analysis Using Big Data Analytics

B. Anandakumar[1*], R. Manimegalai[2]

[1]Dept of Computer Science, Rathinam College of Arts and Science, Coimbatore, Tamilnadu,India.
[2]Dept of Computer Science, Rathinam College of Arts and Science, Coimbatore, Tamilnadu,India.

Available online at: www.ijcseonline.org

Accepted: 24/Sept/2018, Published: 30/Sept/2018

***Abstract*—** This is a Application of Big data analytic. In this concept all Tamil news are stored into Big data analytical tools such as R Tool, MangoDB, Hadoop etc. The News is stored in the form of text, Photos, videos, audio etc. The News is gathered from all daily Tamil news paper, Epaper, Ebook and Social Media Content. Then this News will be analysed in various Angle. Moreover Government Policies, Plan, Advertisement etc. will be stored into the big data Tool. This information is stored into Date wise, month wise and year wise. The news will be classified into various types such as political, Sports, International, national, Local news etc. This is stored into the current details as well as past details. So, we can analysis the every news which is arising in the current problem. Now days Social Networks impact all areas of society. So, the Social Media information's are stored into big data tools. The news is collected over 100 years. So, all the information's are digitized. This proposed project is stored information's of Old literatures of Tamil. Moreover the proposed project stored information's of archaeological information like stone inscription, Palm leaf etc. So, this project consists of all details of Tamil. So, we can store and analyse information's. Moreover it will be stored into Cloud Computing Environment. So, we can access anywhere else. The proposed project is stored all type of Tamil news come from various sources. Now days all the daily are converted into pdf format. So, that will be stored into Big Data Tool. The Proposed Project Consist of old literature of Tamil, Tamil Stone Age arts, palm leaf letters etc.

***Keywords*—** Bigdata Analytic, Cloud Computing, MangoDB, Hadoop, R Tool

I. INTRODUCTION

Big Data Technology[3] is used to store large volume of Data which come from various sources Especially Social network such as WhatsApp, Facebook, Twitter, Instragram etc. Moreover Big Data Tools used to analyse the stored information. It will give hidden truth into visible one. So, it will improve the Business. It is used to store variety of information. For example in Social media information is belongs to video, audio, photos, text, animated file etc. All information are stored into big data tools. The Big data consist of three Concepts. These are volume, velocity and variety. There are three types of structures available in big Data. These are 1. Structured information 2. SemiStructured Information 3. Unstructured Information's. The structured data specify the length and its types. Semi structured data has no separation between data and Schema. Unstructured data has no fixed length that means it has no fixed row and column. For example Social media messages, videos, audio are all belongs to unstructured data types. IOT (Internet of Things) is a kind of technology which is used to store information into Big Data Tools.

After the Social media impact there is bulk amount of information arise from all over the world. In 2021 Social media users will be increase into 3.02 billion[1].In India every year 31% of the growth[2] in social media users. So, this information is stored into digital form. Moreover Social media reflect people mindset. So, we want to store various information and various formats.

There is a proverb available in all over world "today news is tomorrow History". So, we want to store this news in digital format. After the Social networking Growth tamil information's are increased Rapidly. So, we need to digitize all tamil information. Moreover all data created from social networking site are unstructured nature. It store Tera byte of storage. So we need to convert from unstructured data to structured data. Then only we can get the analyse report.

The Proposed System is the combination of big data Analytic and IOT technologies. The Proposed System Used the Following Tools to summarize the details. Moreover this information are Digitized. The proposed System Store large amount of information.

II. METHODOLOGY

1. Apache Hadoop : This is a cluster based Tool which is developed from Java software framework. In this tool we can

parallel run various nodes. It allows data travel from one node to another node. It replicate data in a cluster which provides high availability. HDFS (Hadoop Distributed File System) is the storage system of file system which split data into various node.

2. NoSQL (Not Only SQL): SQL handle structured based data. But in the case of NoSQL can handle unstructured data. It is used to analysis large volume of unstructured Data. There Many NoSQL available for analysing a unstructured data.

3. Hive : This tool consist of SQL like query option HIVESQL. It is run on Top of the Hadoop. It is working in the form of Distributed data Management. This tool also analysing Bulk amount of data storage.

4. Sqoop: This tool is used to transfer structured data to Hadoop. Moreover it transfers relational databases to Hadoop or Hive.

5. Presto : Presto is open source tool which is developed by facebook. It handles petabyte of data and rapidly retrieves data from unstructured data ware houses. It doesn't depend on Mapreduce. But in the case of Hive it depends on Map reduce technique.

6. Knime : Knime is integrated Development Tool which analyse hidden data. It consists of 1000 modules, ready run examples.
7.OpenRefine: This is a powerful tool which execute in a complex and Congested data. Moreover it transfers data from one format to another data format.

8. R-Programing: R language is used to develop statistical software and graphical software.

9. Orange : Orange is the open source software which is used to visualize the data and create interactive work flows for analyze. Orange is the combination of various visualization, scatter plots, bar charts, trees, networks and Maps.

10. Google Fusion Table: this is kind of tool which is used to analyse data, visualize data and Mapping.
The above Tools are various Data analysing, visualization, fusion, extraction and statistical tools. These tools are used to analyse large unstructured data. It is used to convert unstructured format to structured format.

III. RESULTS AND DISCUSSION

The proposed system consists of the following Components. These are attached with data warehouse as well as big data analytical tools. The data warehouse consists of the following components. These are 1. Old Tamil Literature 2.Sports news 3.Political news 4.Local News 5. Tamil archaeology Details 6.International News 7. National news 8. Social media Data. This Component may be expanded based on the classification of the information and news. Moreover all the news which are stored into the big data is nearly 100 years news are digitalized. So this news will be the history in future.

In the proposed system consist of old Tamil literature module itself. This module is store the Tamil literature. It will be useful for given the quote. Moreover we can give the Excerpt to all information. For example Union government Budget is submitted many times. But Before budget submission the minister gives Excerpt to all people. The second Component is consisting of sports news. This component Consist of all sport information. The third Component of the proposed system is political news. This will be classified into national, international, state, local political news etc. This news is stored into last 100 years Tamil news. This will be referenced in future.

The fourth components of the proposed system are Local Tamil news. This is classified into district news, zone level news, local functions, leaser speech etc all are stored into local news. So, this will be reference in future. The next Component is the Tamil archaeological details. In this component archaeological details are digitized and stored into data warehouses. The details belong to palm leaf, stone inscription, Copper plate, Pot letters all are store into data warehouse. The sixth Component is International News. The international news is take bulk amount of memory. So, this will be classified into Continent level news. Further it will be sub divided into Country level. In this Components consist of political, conference, meetings, etc.

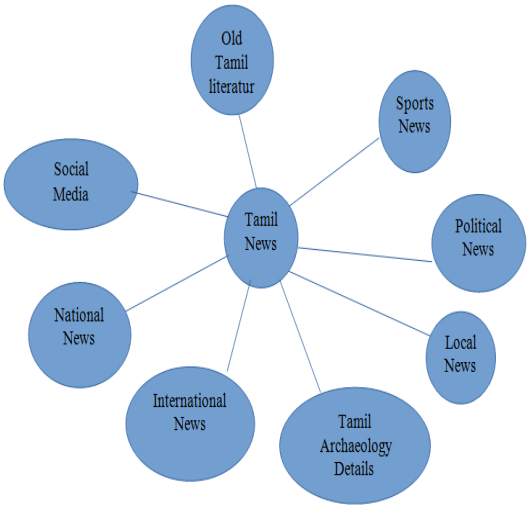
The next component is national level news. This will be sub divided into state level news and Zonal level news. The zonal level is subdivided into south, north, east, west and central zone. The final Component is social media news. This information are capture large amount of Memory. In this Component consist of various data types such as text, image, video, audio, voice etc. This will be analyzed in future. In the proposed system store a bulk amount of information. The system is working in the form of Commands. All commands are working in Tamil letters. Moreover we can give the command in the form of voice. So, the voice will be converted into Tamil font. This will be act as command. So, in the proposed system any user can easily handle the proposed System. Moreover all command are converting into icon that will be display in mobile app. So, the proposed system is user friendly in nature and it will be working into Tamil and English language.

The second Major part of the proposed system is analytical area. The analytical area consists of six components itself. These are 1.Integration and management 2.Methods & techniques 3.Service Analytics 4.Ancient Tamil Literature Comparison 5.Social Network Analytics 6.Information Management. The first Component integrates the all information which are store into data warehouse. It will be managed into various sub components. The second Component is the methods and techniques which retrieve various information from data warehouse. This will be displayed into various techniques such as visual reports, Graph, Maps etc. So, the user can easily understand the statistic technique and analysed report. The third component is service analytics. This will be analysed in various service. The fourth component stores all Tamil literatures from Sangam age to till date. So, we can give the Excerpt to any information. The next component is the social network analytics. This is vast area which is analyzed into unstructured data. So, it's a Complex activity for the analyzer. The last Component is Information Management. There is Bulk amount of data store into data warehouse. So, we should differentiate from one information to information. So, we should Divide the component into subdivide. Then only we can easily retrieve the information as soon as possible.

IV. CONCLUSION AND FUTURE SCOPE

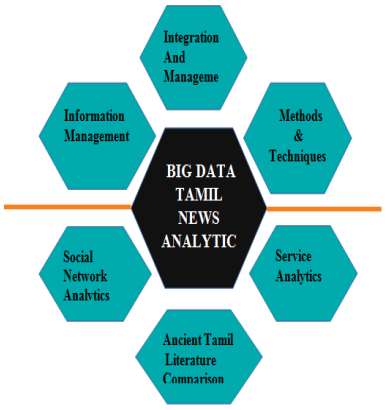
The proposed project will be converted into other Indian language such Telugu, Kannda, Malayalam etc. Moreover it will reduce the text. It displayed into visual form.

V. STRUCTURE OF THE PROPOSED SYSTEM



Structure of Integrated Tamil news Data warehouse

Fig.1



Structure of Big Data Analytics Component

Fig.2

ACKNOWLEDGMENT

I thank to Rathinam College Management for supporting me to publish journal. I thank to Our Principal Dr.R.Muralidaran encourage to publish the Journal work. I thank to my HOD Mr.S.Raja and my department Colleague to help my work. I would like to thank my wife Mrs.Kavitha for help to publish the work.

REFERENCES

[1]. Facebook 2017 Statistics Report.
[2]. Social Media growth statistical report.
[3]. Big Data Analytics, Tools and Technology for effective Planning by Arun k.Somani, Ganesh Chandra Deka.
[4]. Internet of Things, A Hands on Approach by Arshdeep Bahga,vijay Madisetti

Authors Profile

Mr. B.Anandakumar pursed Master of Computer Applications from Bharathidasan University, Tamilnadu India in 2001 and Master of Philosopy from bharathiar University, Tamilnadu in the year 2011. He is currently working as Assistant Professor in Department of Computer Science, Rathinam College of Arts and Science,Coimbatore,Tamilnadu,India. He has published more than 4 research papers in reputed international journals and including conferences. His main research work focuses on Big Data Analytics, Data Mining, IoT and Cloud Computing. He has 12 years of teaching experience and 7 years of Research Experience

Mrs.R.Manimegalai is an Assistant professor in Department of Computer Science, Rathinam College of Arts and Science, Eachanari, Coimbatore. She is having more than seven years of teaching experience and two years of research experience. Her core area of Research is Data mining, Image processing, Soft Computing.

