An Approach to Extract the Data using Hadoop ETL in Disaster Event

M.Saranya[1]
Department of computer science, Research, and development center
Raja doraisingam Government Arts College, Sivagangai
Email id: varunila15@gmail.com
A.Prema[2]
Department of computer science, Research, and development center
Raja doraisingam Government Arts College, Sivagangai
Email id: jeyaprem2010@gmail.com
K.Chelladurai[3]
Department of computer science, Research, and development center
Raja doraisingam Government Arts College, Sivagangai.
Email id: kchelladurai05@gmail.com

ABSTRACT

A key role of Computer Scientists in has been devising ways to manage and analyses the data produced management situation. The area of disaster Management receives increasing attention from multiple disciplines of research. In this paper, we make an effort to organize the current Knowledge in analyzing the Disaster situation Hadoop Map reduce framework the output of the result consist of information extraction, load, Transformation retrieval, information filtering Map Reduce framework and produce output Hadoop tools in Big data analytics
Key Words: **Big Data, Disaster Management, ETL, Hadoop, Map reduce**

I. INTRODUCTION

Big Data is a gathering of the large dataset that cannot be processed using conventional Computing technique. Big Data is not merely a data rather it has become a complete subject Which involves various tools techniques as well as framework The necessitate of big data generated from the large companies like facebook, yahoo, Google, youtube etc for the purpose of massive amount of data also Google contain the large amount of in sequence[1]

The ETL method consists of scheming a target, transforms statistics for the target, scheduling and monitor processes. The reason of using ETL tools is to save time and make the whole process more steady The ETL tools are adapted to provide the functionality to meet the venture necessity. Hence many of them choose to make their own dataware house themselves [2, 3, and 4]

The technology and growing quantity of data(Big Data), need is felt towards implement effective analytics techniques(Big Data analytics) to analyses this big degree of data for unknown and useful facts, pattern, Big data is the term for data sets so large and difficult that it becomes complicated to process using fixed data management tools or processing application. This paper reveals most recent progress on big data networking and big data [5]

Dhole Poonam et al presented that Hadoop is open-source software that enables reliable, scalable, distributed computing on clusters of low-priced servers [6, 10]

It is calculated to scale up from single servers to thousands of machines, each contribution local computation and storage Hadoop consist of two components Hadoop distributed file system (HDFS) and Map Reduce framework. [7]

Wang, F. Described Map Reduce is a software framework for distributed processing of large data sets on computer clusters.map reduce is intended to make possible and simplify the processing of vast amounts of data in parallel on a large cluster of commodity hardware in a reliable, fault-tolerant manner [18]

II. RELATED WORKS

They presented the design and assessment of a data conscious cache framework that requires a minimum change to the original Map Reduce encoding model for provisioning incremental meting out for big data application using the Map Reduce mock-up [9].

The author stated the meaning of the technology that grip data like Hadoop, HDFS, and Map Reduce. The author not compulsory about various schedulers used in Hadoop and about the scientific aspects of Hadoop. The author also focuses on the weight of YARN which overcomes the boundaries of Map reduce [10]

The author gave some vital emerging frame for big data Analytics and a 3-tier structural design model for Big Data in Data Mining. In the future, 3-tire design model is more scalable in operational with different situation and also profit to trounce with the main matter in Big Data Analytics for the store, analyze, and hallucination. The framework model given for Hadoop HDFS spread data storage, real-time NoSQL databases, and Map Reduce distributed data giving out over a cluster of product servers [11]

Disaster management refers to the inclusive approach in all phase of disaster successfully dropping the collision of disaster. Disaster management series

consists of the next different phases .such as Felling, Relief &Rescue, healing, and mitigation phases. Over the past decade, innovative use of statement and meteorological means of INSAT system is begin operationally used towards the track, monitor, and forecast of the cyclone. The recent achievement includes deluge mapping of all the major floods in the country in near real time mode, drought severity appraisal using settlement data on tnightly/journal time scales, the landslide zonation pilgrimage route in the Himalayas, monitor of the cyclone and harm estimation. The potential of the geographic positioning system(GPS) to precisely determine the position of a location is being used to measure ground movements associated with plate tectonics ISRO has well-known a decision support center (DSC) at National Remote Sensing Agency, Hyderabad. To provide sensible in order meeting the user needs in provisos of in sequence content, turn-around-time and set-up

Many regions in India are highly susceptible to usual and another disaster on the explanation of physical Situation. About 60% of the total area of the country is susceptible to seismic injure of buildings in undependable degrees. The most susceptible areas, according to the present seismic zone map of India, area situated in the Himalayan regions. Kutch and the Andaman and Nicobar Islands, which are chiefly earthquake risk prone. Over 8% Indian area of 40 million hectares is flat to floods, and the regular area overstated by floods annually

The author conquered the week points of established Extract, Transform and Load tool's architecture and proposed a three layers architecture base on metadata. That built ETL process more bendable, multipurpose and efficient and finally they designed and implemented a new ETL tool for drill dataware house. A systematic review method was projected to identify, extract and analyze the main proposal were branded and compared based on the skin, activities, and document of ETL processes and finished the study by reflecting on the approach being considering and providing an update frame for prospect study[12]

Each year, a number of natural disasters beat across the sphere, killing hundreds and causing billions of dollars in property and infrastructure damage. Minimizing the impact of disasters is very important in today's society. As the capabilities of software and hardware evolve, so does the role of information and communication technology in disaster alleviation, grounding, answer, and revival. A large measure of disaster-related data is accessible, together with reply plans, records of earlier incident, imitation data, social media data, and websites. Though existing data management solutions offer little or no mixing capabilities. Moreover, the recent advance in cloud compute, big data, and NoSQL opens the door for the new solution in disaster data management. In this paper, a knowledge as a service (KaaS) skeleton is future for disaster cloud data management (Disaster-CDM), with the objectives of

1) Storing large amount of disaster-related data from various source
2) Facilitate search, and
3) Behind their interoperability and addition.

Data are stored in a cloud setting using a mixture of relational and NoSQL databases. The case study accessible in this paper illustrates the use of disaster CDM on a case in point of imitation models [13].

Hristidisr et al. surveyed data management and examination in the disaster area. The main focus of their survey was on data therapy techniques lacking the storage facet. In disparity, in disaster-CDM, storage and therapy are considered as vital parts. Hristidis et al. acknowledged the following data therapy technology as pertinent in disaster data management: in order extraction, information retrieval, information filtering, data mining, and conclusion support. Similarly, disaster-CDM uses a number of technologies from information extraction and retrieval. The survey reveals that the greater part of the research has listened carefully on a very narrow area of disaster management, for example, a precise disaster event such as an earthquake or flood, or specific disaster-related behavior such as statement among actors, estimating disaster damage, and use of mobile devices. They also predictable the need for elastic and customizable disaster-management solution that could be practical in a different disaster situation. Disaster-CDM Aims to supply such solution using cloud and NoSQL approach.[14]

Vibhavari has given the design and estimate of a data alert cache framework that require lowest amount change to the original map reduce encoding model for provisioning incremental giving out for Big Data application using the Map Reduce mock-up[20] the author declared the meaning of some of the technologies that switch Big Data Like Hadoop, HDFS and Map reduce. The author optional about a variety of schedules used in Hadoop and About the technical aspect of Hadoop the author also focus on the importance of YARN which overcomes the confines of map reduce[10]

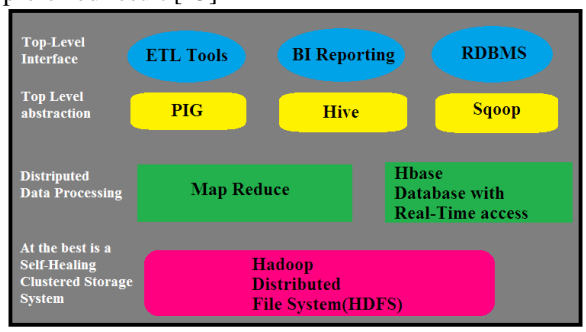
The author carry on with the Big Data meaning and improve the meaning given in that includes the 5V Big Data property: Volume, Variety, Velocity, Value, Veracity, and recommended other size for Big Data Analysis and classification in particular compare and different Big Data technology in e-science industry, trade, social media, health care[27]

S.Vikram Phaneendra et al. illustrate that in historic days the data was less and simply Handled by RDBMS but in recent times it is difficult to handle vast data through RDBMS tools, Which is favored as "Big Data "in this they tell that big data differ from other data in five dimensions such as volume, velocity, variety, value, and complexity. They illustrate the Hadoop structural design consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handles great data sets, the scalable algorithm does log management application of Big Data

can originate out in economic, trade industry, Healthcare, mobility, insurance. The authors also listening carefully on the challenge that needs to be faced to enterprises while handling Big Data: -Data privacy, search analysis, etc [29]

III. HADOOP

Hadoop was formed by Doug cutting and Mike Cafarella in 2005.Doug cutting, which was operational at Yahoo! At the time, named it behind his son's toy elephant. It was initially residential to support sharing for the Nutch search engine project Hadoop shows MAD individuality. The 'M' Stands for magnetic i.e. it can store all type of data source and attract them towards itself. The 'A' refer to the quickness as various operations on big data simply can be easily performed on it. The 'D' stands for deep and it is able of performing Ad-hoc and multipart analytics. To execute depth analytics over the big Data, Hadoop provide preferred result [15]



IV. EXTRACT-TRANSFORM-LOAD

Improvement of data warehouse involves the ETL process. It is a multipart grouping of process and technology. This system consists of three practical entities: Extract, Transforms, and Load. Extract function extracts related information from basis data for decision making which then desired to be transformed into the dissimilar scheme to match the dataware house schema. The last function loads the data into data warehouse [16] the dataware house platform is categorized on the basis of their utilize system, hardware server, storage system. By in view of the above cloud and Hadoop based data ware housing stage are the most reliable to meet today's necessity.ETL process also wants to be developed on the same stage to gain the maximum earnings [17] organizer have listed out their load to process peta byte of data every day. They have demanded the requirement develop the well-organized ETL, competent of behavior all kinds of data [18]

V. HADOOP ETL

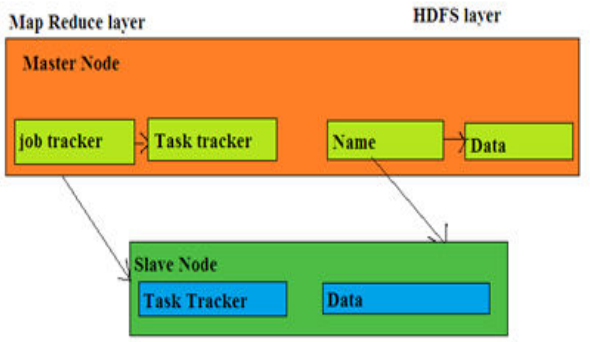
An ETL workflow by way of Hadoop process comprises a range of tasks as follows:

- Initial, input data starting various data source, which contain data in different format to HDFS
- Map the consolidate data into a bench to make it query table
- The target data is misshapen within a finalize format and is mapped to reason source
- Adapt all input data sources information into target format and make it obtainable at central

Use the finalize data obtainable at central for reporting or analytics

VI. MAP REDUCE

Map reduce is a parallel programming that aims at proficiently processing large data sets This programming model is based on three concepts: (i) On behalf of data as key-value pairs,(ii)significant map function, and(iii) significantly reduce function. The map function takes key value pairs as an input, and produce zero or more key price pair. Outputs with the same key are then gathered to collect (shuffled) so that key-{list of values} pairs are given to reducers. The reduce function to process all the values related to a given key. The most famous execution of this model is the Hadoop framework which provides a distributed platform for executing map reduce jobs [19]



VII. EXISTING WORK

Franklin et al., in this piece we there the challenge on administration and analysis of disaster data by means of the lately proposed example of data spaces. We select to there the problem from this viewpoint given that a disaster management system shares many challenges with the organization of mixed data sources in the data spaces pattern. A data space is a freely included set of data sources, where combination occurs in an indolent way after that a pay-as-you-go come near in the case of disaster management, the source include information from government agency, business entities and non-government organization (NGOs), media announcement and web blogs [21]

It has been shown that an active disaster management system has the possible of .u; - manage best by a loosely coupled, flexible, dynamic data space system as here is a table arrival of in sequence, consisting of structured and unstructured data of different types and from autonomous sources which need to be professionally stored, indexed, retrieved and optimized. They argue that there is a wish for a disaster

management data space support platform (DM-DSSP), with basics optional by Franklin et al [21 22]

Most of the time, the world is at risk. Capacity structure and vigilance at the global level are in enough. Two third of the earth is enclosed by the ocean. The calculation about the mega push below the ocean bottom is a phenomenal task for the people living in the one-third of the earth. For a safe and happy endurance, people in the earth need to take on the task in an orderly and technical manner at the earliest likely time for this task, the arrangement of a global disaster Authority (GPA) is required [23]

Space knowledge has its own possible role in the relief, healing, alleviation and forecasting phase of disaster management. The earth surveillance satellites give complete and multi sequential reporting of large areas in real time and at common intervals. Thus they have become valuable for incessant monitoring of impressive as well as exterior parameters related to the natural disaster. The deliverable that can be expected from space data sets at the present state includes high temporal revert to, high spatial decision, stereo mapping capability, interferometric SAR and onboard giving out. With more advanced in the space skill stored in prospect, with complicated sensors and more capability, it is possible for superior management of natural disasters [8]

VIII. PROPOSED WORK

This paper integrates the concept of Hadoop ETL and Map Reduce to get a better decision in disaster management and reduce the time in data processing. The association of this Hadoop ETL was examined through sample sales records and found out that the transformation time was less than the existing transformation time, in order to provide an optimal solution during disaster situation.

We are using the term of methods in having data and through the processing scalability of disaster. The importance of the timely, accurate and effective use of available information in disaster management scenarios has been extensively discussed in literature the data will be produced over flood affected areas. In this paper, disaster and disaster management are defined, the infrastructures that are damaged in a different disaster are mentioned and then a comparison is performed between 2003 Bam earthquake and tsunami in Japan that shows disaster management in developing countries significantly can degrease the losses in case of different disaster. Based on the result, we can say that natural disaster can be controlled if they correctly manage. Suitable infrastructure can control the disaster before it becomes catastrophe and preparedness before disaster can significantly decrease losses.

IX. METHODOLOGY FOR PROPOSED WORK

The extracted sample data of disaster event is given in the table1 and is represented in the form graph in figure 3.

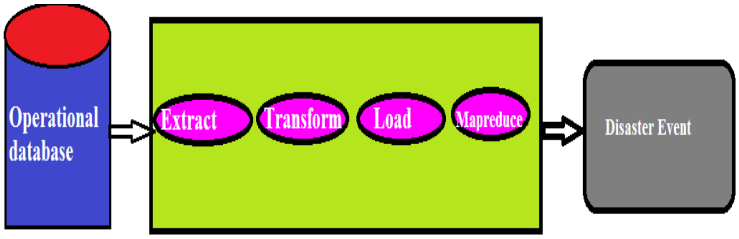
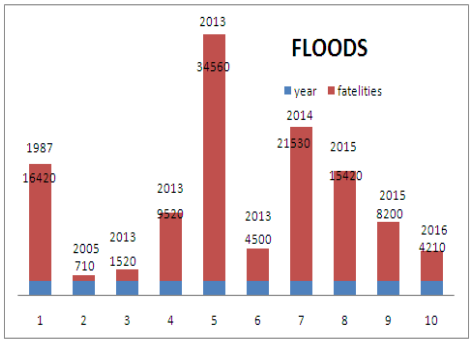


Fig: 3 Methodology for proposed work

Year	Event	Area	Fatalities	Action
1987	Floods	Assam	16420	severe
2005	Floods	Maharashtra	710	Normal
2013	Floods	Kenya	1520	Normal
2013	Floods	Zimbabwe	9520	Normal
2013	Floods	Mozambique	34560	Severe
2013	Floods	Niger	4500	Normal
2014	Floods	South Africa	21530	Severe
2015	Floods	Malavi	15420	Severe
2015	Floods	Gujarat	8200	Normal
2016	Floods	Assam	4210	Normal



X. CONCLUSION

The highly developed sketch of Hadoop ETL is projected in order to achieve better performance of ETL process. In this paper we in attendance for the first time an inclusive survey of the pains on utilize and advancing the organization and analysis of data to handle disaster management situation .we planned our findings across the subsequent computer Science disciplines: data integration and ingestion, information Extraction, information retrieval, information filtering and decision support. This proposed concept was demonstrated through sample records and it suggested that this approach reduces the time. The goal of this paper is to find out an optimal solution to take right decision during disaster situation

We presented concrete future study directions for computer scientists to advance the knowledge of

useful data management and analysis in disaster management.

XI. REFERENCES

[1]Varsha B.Bobad, "International Research Journal of Engineering and Technology (IRJET)", Volume: 03 Issue: 01

[2]Inmon, William "Data Mart Does Not Equal Data Warehouse".DM Review.com. (2000-07-18)

[3]Jeffrey R. Bocarsly, "The Data Warehouse Toolkit." Complex ETL Testing-A Strategic Approach

[4]R. Kimball and M. Ross. WileyPublishing, Inc., 2002.

[5]"Survey of Recent Research Progress and Issues in Big Data" www.cse.wustl.edu/~jain/cse570-13/ftp/bigdata2/index.html 1/13

[6]Dhole Poonam B, Gunjal Baisa L, "Survey Paper on Traditional Hadoop and Pipelined Map Reduce" International Journal of Computational Engineering Research Vol, 03 Issue, 12

[7]Varsha B.Bobade" Survey Paper on Big Data and Hadoop" International Research Journal of Engineering and Technology

[8]V. Bhanumurthy*, G Behera "Deliverable from space Data sets for Disaster Management-present and true trends"

[9]Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, —"Survey Paper on Big Data"International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[10]Amogh Pramod Kulkarni, Mahesh Khandewal, —"Survey on Hadoop and Introduction to YARN", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

[11]Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), —"Big Data Framework" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499

[12]Inmon, William (2000-07-18). "Data Mart Does Not Equal Data Warehouse". DMReview.com

[13]Katarina Grolinger, Miriam A.M. Capretz." Knowledge as a Service Framework for Disaster Data Management"

[14]V. Hristidis, S. Chen, T. Li, S. Luis, and Y. Deng, "Survey of Data Management and Analysis in Disaster Situations," Journal of Systems and Software, vol. 83, no. 10, pp. 1701-1714, 2010.

[15]Song .Y, Davis Karen C," Analytics over large scale Multidimensional Data: The Big Data Revolution, Communications of ACM," 2011

[16]Merinela Mircea," Business Intelligence--Solution for Business Development", Intech Publisher, 2012

[17]Kuldeep deshpande, and dr. Bhimappa desai,"limitations of dataware house platforms and Assessment of hadoop as an alternative," Volume 5, Issue 2, pp. 51-58, IJITMIS, 2014

[18]Yongqiang He et al RCFile: "A Fast and Space-efficient Data Placement Structure in Map Reduce-based Warehouse Systems," ICDE, 2011

[19]Vibhavari Chavan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014, 7932-7939

[20]Franklin, M.,Halevy, A., Maier, D., 2005. From databases to data spaces:"A new Abstraction for Information Management. ACM SIGMOD Record" 34 (4), 27–33.

[21]Saleem, K., Luis, S., Deng, Y., Chen, S.-C., Hristidis, V., Li, T., 2008. "Towards a business Continuity information network for rapid disaster recovery." In: Proceedings of the 9th Annual International Conference on Digital Government Research, Montreal, Canada, May 18–21, pp. 107–116

[22]Senthi Vadivel Bhupatthi Rav" Disaster Management: A Global Issue" International journal of civil and structural engineering Volume 1, No 1, 2010

[23]Sagiroglu, S.Sinanc, D.,Big Data: A Review ,2013, 20-24.

[24]Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), —"Big Data Framework" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499

[25]S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).