Drought Prediction using Geo-Spatial Big Data

A Abisha, R Beulah Jayakumari, D Doreen Hephzibah Miriam

Abstract*: *The digital world with digital processing, requires large storage space. The continuous explosion of the data such as text, image, audio, video, data centers and backup data lead to several problem in both storage and retrieval process. In this paper drought analysis and prediction is done using big data processing tools such as Hadoop and hive which can increase high. Previously to analyze and predict drought, traditional techniques such as AVISO model is used which is complex to process, requires more processing time, cannot process huge data and also has more security issues like malware in the database, abuse of privileges, etc. The system proposed in this paper can process huge data and has more processing speed. Here, drought analysis and prediction is carried out. To analyze drought dataset with more than ten lakhs are processed and drought type is found using map-reduce algorithm which maps and reduces the data using numerical summarization. Drought types such as D0, D1, D2,D3 D4 are analyzed to obtain reduced output. The obtained drought type are clustered using hive. To predict drought, random forest algorithm acts as an predictor which creates multiple decision trees and finds the best split among them. Finally, the predicted output is visualized using the time series model. The tools used in this paper include Hadoop and hive which can process huge data and it is the solution of Big Data. Hadoop is an open-source software framework for storing data and processing them efficiently, even if the data size is very huge. Hadoop uses Hadoop Distributed File System(HDFS) for storage and MapReduce for processing the data. Hive is a query processing tool which is built on top of Hadoop. It is a Structured query language(SQL)-like language called HiveQL (HQL). In this paper hive is used to cluster the data obtained from MapReduce. Thus using Big Data improves performance more than 50% compared to traditional system.

Index Terms*: *Big Data, Hadoop, Hive, Random Forest.

I. INTRODUCTION

The term "big data" has first appeared in the mid-1990s and gradually became popular at 2008 and began to be recognized in 2010. Big data is huge amount of data that cannot be processed using traditional systems. Traditionally, geospatial data refers to geo-referenced data that correlates to Earth's environmental components and processes and further to the interaction between humans and Earth by using spatial technology assisted with ground station systems. Drought is among the most disastrous natural hazards and occurs in virtually all geographical areas. Severe drought events in recent decades, including 2010–2011 East Africa drought, 2011 Texas drought, and 2012–2015 California drought. It has become the threat to the world. Hence, analyzing and predicting drought is important for the safety of environment.

Revised Manuscript Received on July 22, 2019.

A Abisha, PG Student, Department of computer science and engineering, Jerusalem college of Engineering, Chennai, India.

Dr. R Beulah Jayakumari, Associate Profesor, Department of computer science and engineering, Jerusalem college of Engineering, Chennai, India.

Dr. D Doreen Hephzibah Miriam, Director, Computational Intelligence Research Foundation, Chennai, India.

II. RELATED WORKS

In [1], a high resolution meteorological drought forecast model was developed to provide drought forecast information based on several indexes for ungauged areas. The contribution of long-range climate forecast data was not significant under certain conditions used in this study, but further improvement can be done and is expected if forecast skill is improved. It is complex to implement and uses SPI based traditional processing techniques.

In [2], normalized difference vegetation index (NDVI) to predict drought is used. By calculating the correlation between LST and NDVI, it can be clearly noticed that they show a high negative correlation. The correlation between LST and NDVI is -0.635 for the year 2002 and -0.586 for the year 2012. The LST when correlated with the vegetation index it can be used to detect the agricultural drought of a region, as demonstrated in this work. It can be used only to analyze agricultural drought and it also uses traditional technique for processing.

In [3], images of past drought were compared with post-drought images of our targeted area and land use maps were developed for spatio-temporal analysis. The results of the study revealed that vegetation in Thar showed an improving trend from 2002 to 2011 and then declined from 2011 to 2014. The rainfall occurred at a below average rate and SPI values for each year were calculated to be negative, indicating below average rainfall. This actual precipitation data was then compared with the data obtained from the Tropical Rainfall Measuring Mission (TRMM) satellite and R2. Further, the average temperature for the five years under study was analyzed by graphical representation and it was revealed that the temperature of Thar has increased by almost 1 °C during the last decade. It has used spatial data images obtained from satellite and is difficult to analyze and is cost effective and it has used traditional techniques for processing.

In [4], it is assumed that due to severe drought events that have been occurred and many disastrous impacts in recent decades, efforts have been made recently to drought monitoring. This study has an integrated R package and explains a wide range of its applications for drought modelling and assessment based on univariate and multivariate drought indices. It is complex to implement and R does not provide flexibility for prediction like other tools.

In [5], potential of the Standardized Precipitation Index (SPI), Markov chain models, and time series modelling to characterize meteorological drought is used. The Markov chain analysis learnt that the probability of having two consecutive drought years

appears to be higher in the southern sub basins. For a return period of 17 years, the SPI is lower than −1.5 (severely dry) in many sub basins. It uses traditional techniques for processing and uses only one drought index, hence result will not be accurate.

In [6], two drought monitoring systems have been developed in the Czech Republic based on the SoilClim and AVISO soil moisture models are studied. SoilClim is based more on real soil properties and aimed primarily at agriculture, while AVISO complements the system with more theoretical presumptions about soil, showing, rather, climatological. Aviso model is one of the traditional technique used for prediction algorithms and it also consumes time.

III. SYSTEM ARCHITECTURE

In the system architecture, drought analysis and prediction using geo-spatial big data is given below in the fig 1. Input data is stored in HDFS and map-reduce algorithm is applied to analyze drought and drought type is obtained using HDFS. After drought type is obtained hive is used to cluster data and random forest algorithm is applied to predict drought and the output is visualized using time series model.

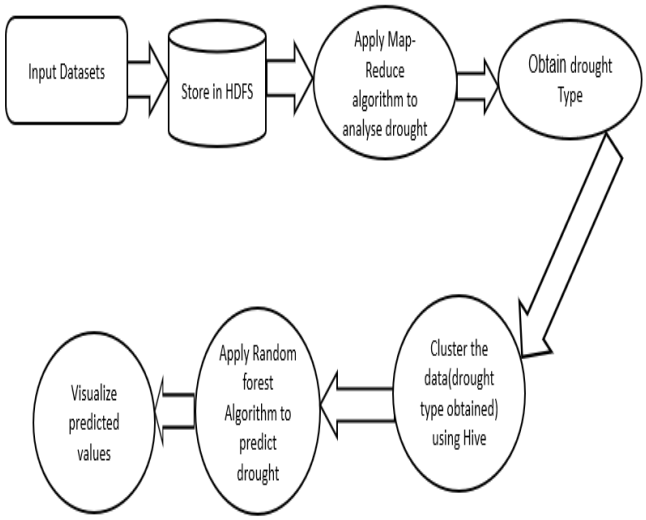

Fig. 1 System Architecture

IV. SYSTEM DESIGN

In the proposed system, Hadoop is used which will give more processing speed and it helps to process 10 lakhs records in few seconds. Here drought type is used to find the drought accuracy. Drought analysis is done using map-reduce algorithm using mapper and reducer functions. The map-reduce algorithm processes n lakhs record to give reduced records by finding the averages as shown in table 1. After drought type is found, hive is used to cluster similar drought types. After drought analysis is done drought prediction is done using Random forest algorithm to predict drought for upcoming years and to manage storage efficiently.

Table I Drought Type And Description

Drought Type	Description
NONE	Percentage of the state that is not in drought
D0	Percentage of the state that is in abnormally dry conditions
D1	Percentage of the state that is in moderate drought
D2	Percentage of the state that is in severe drought
D3	Percentage of the state that is in extreme drought
D4	Percentage of the state that is in exceptional drought

A. Data

Drought datasets are obtained and stored in HDFS. The Hadoop Distributed File System (HDFS) is the primary data storage used by Hadoop applications. It contains one NameNode and one or more DataNode architecture to implement a distributed file system that provides high-performance across Hadoop clusters. After placing the data in HDFS, mean of drought type has to be calculated. Dataset description is shown in table II.

Table II Dataset Description

FIELDS	DESCRIPTION
state	Name of the state.
NONE	Percentage of the state that is not in drought.
D0	Percentage of the state that is in abnormally dry conditions.
D1	Percentage of the state that is in moderate drought.
D2	Percentage of the state that is in severe drought.
D3	Percentage of the state that is in extreme drought.
D4	Percentage of the state that is in exceptional drought.
validStart	The starting date of the week that these observations represent.

validEnd	The ending date of the week that these observations represent.

B. Apply Map-Reduce Algorithm

Map-Reduce is performed to reduce the huge volume of data. The Numerical summarization(average) is applied on the data to group the data by Year and Month. Mapper is used to map the values using key, value pairs. Reducer is used to reduce the values and perform summarization in the data. Map-Reduce algorithm is shown in table III.

Table III Map-Reduce Algorithm

INPUT :	Key : state, month, year Value : drought values from D1-D4
OUTPUT :	Reduced Output using summarization

ALGORITHM :	class Mapper
	method Map(<key:state, month, year>, <value:droughttype D1-D4>)
	for all term t in droughttype D1-D4 do
	Emit(term t, count (c=1))
	class Reducer
	method Reduce(term t, counts [c1, c2,...])
	sum=value
	average = 0
	for all count c in [1….n] do
	average = sum /c
	Emit(term t, average)

C. Clustering the Data using Hive

Hive is a tool which is used to process structured data in Hadoop. It resides on top of Hadoop and is used to process query given by the user. Here clustering is done using hive to group the data with same drought type. Clustering data will be useful to find the similar drought types.

D. Apply Random Forest Algorithm

Random forests is a method that operates by constructing multiple decision trees during training phase. The decision produced by majority of the trees is chosen by the random forest as the final decision. It can work with missing values [1]. As it is a tree based algorithm, the prediction is highly accurate, stable and easy to interpret. Random forest algorithm is shown in table IV. Training datasets are formed from the existing data and testing data is the data which has to be predicted. Prediction is done using python packages and hence it is predicted very efficiently without any complexity. The predicted output is stored as an csv file.

Table IV

Random Forest Algorithm

INPUT:	Training data derived from the drought type.
OUTPUT:	Predicted drought for upcoming years
ALGORITHM:	➢ From the set of inputs create training and testing data sets. ➢ Create n number of decision trees. ➢ Classify the data in decision tree based on certain conditions. ➢ Classify till 100% accuracy is obtained. ➢ Now predict the final value by predicting the output based on certain conditions. ➢ The decision of the majority of the trees is chosen as the final value.

E. Visualize predicted values

Finally visualize the obtained data using time series model. Time series data shows how an indicator performs over a period of time. After the predicted values are obtained by applying random forest algorithm, visualization is done using python packages. The visualized output gives a clear view of the drought type present for the particular state with month and year for the particular state. Thus visualizing the output is done finally for better understanding.

V. RESULT ANALYSIS

The dataset is placed in HDFS and map-reduce is performed twice to obtain

summarization and to find the drought type. The first map-reduce is done for analysis and to reduce 10 lakh records to 10 thousand records. The second map-reduce is done to find drought type. The obtained output with drought type is shown in fig 2.

AK	Alaska	2000	01	100.0	None
AK	Alaska	2000	02	100.0	None
AK	Alaska	2000	03	100.0	None
AK	Alaska	2000	04	100.0	None
AK	Alaska	2000	05	100.0	None
AK	Alaska	2000	06	2.0884259259259257	D1
AK	Alaska	2000	07	29.53185185185185	D1
AK	Alaska	2000	08	100.0	None
AK	Alaska	2000	09	100.0	None
AK	Alaska	2000	10	100.0	None
AK	Alaska	2000	11	100.0	None
AK	Alaska	2000	12	100.0	None
AK	Alaska	2001	01	100.0	None
AK	Alaska	2001	02	100.0	None
AK	Alaska	2001	03	100.0	None
AK	Alaska	2001	04	100.0	None
AK	Alaska	2001	05	100.0	None
AK	Alaska	2001	06	100.0	None
AK	Alaska	2001	07	4.327259259259259	D1
AK	Alaska	2001	08	100.0	None
AK	Alaska	2001	09	100.0	None
AK	Alaska	2001	10	100.0	None
AK	Alaska	2001	11	100.0	None
AK	Alaska	2001	12	100.0	None
AK	Alaska	2002	01	100.0	None
AK	Alaska	2002	02	100.0	None

Fig.2 Map-Reduce Output

CPU time spent for first map-reduce(to find average of 10 lakhs records) is nearly 102710 milliseconds(ms) which is equal to 102 seconds. CPU time spent for second map-reduce (to find drought type of 10 thousand records) is 2710 ms which is equal to 2.7 seconds. Thus huge data is processed within few seconds. Next, Clustering is done using hive by creating table and loading the output obtained from map-reduce into hive. After loading the data to hive, clustering is done. Next, data prediction is done on the data obtained from map-reduce. Random forest algorithm is applied and the output is predicted for upcoming years using training data. After prediction of all the states, visualization is done using python language and the output is shown in fig 3.

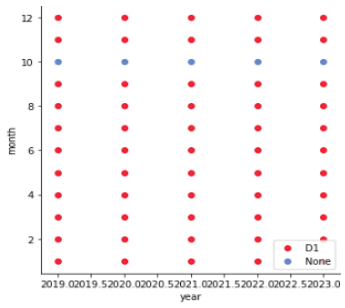

Fig.3 Predicted output of a state

VI. CONCLUSION AND FUTURE WORK

The velocity of data generation and growth is increasing because of the proliferation of mobile devices and other device sensors connected to the Internet. The technique that we have discussed improves storage efficiency and thereby improve the performance by enabling storage resources to transfer and handle more data. In this paper, Drought analysis and prediction is done. And also, Big Data tools such as Hadoop, hive are used which can process huge data within few seconds. Drought prediction is done using random forest algorithm for prediction and visualization is done for visualizing the predicted output using time series model. In future, an web based monitoring system can be developed to monitor drought on regular basis. Thus drought analysis and prediction of drought in particular region using Big Data which can benefit agriculture, economy, human sustainability, etc. of a country in present and in future is explained above.

REFERENCES

[1] Jinyoung Rheea, Jungho Imb, 2017. Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data.

[2] Sruthi.S, M.A.Mohammed Aslam , 2015. Agricultural Drought Analysis Using the NDVI and Land Surface Temperature Data; a Case Study of Raichur District.

[3] Muhammad Bilal ,Muhammad Usman Liaqat , Muhammad Jehanzeb Masud Cheema, Talha Mahmood and Qasim Khan ., 2017. Spatial Drought Monitoring in Thar Desert Using Satellite- Based Drought Indices and Geo-Informatics Techniques.

[4] Zengchao Hao , Fanghua Hao , Vijay P. Singh , Wei Ouyang , Hongguang Cheng 2017. An integrated package for drought monitoring, prediction and analysis to aid drought modeling and assessment.

[5] Brahim Habibia, Mohamed Meddib, Paul J.J.F. Torfsc, Mohamed Remaound, Henny A.J. Van Lanenc ., 2018. Characterisation and prediction of meteorological drought using stochastic models in the semi-arid Chéliff–Zahrez basin (Algeria).

[6] Petr Št ˇepánek, Miroslav Trnka, Filip Chuchma , Pavel Zahradníˇcek, Petr Skalák, Aleš Farda , Rostislav Fiala , Petr Hlavinka , Jan Balek , Daniela Semerádová and Martin Možný , 2018. Drought Prediction System for Central Europe and Its Validation.

[7] Vicente-Serrano, S. M., S. Begueria, and J. I. Lopez-Moreno,2010. A multiscalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index. Journal of Climate, 23, pp. 1696-1718.

[8] D. A. Wilhite, Drought: A Global Assessment, Natural Hazards and Disasters Series, Routledge, London, UK, 2000.

[9] A. K. Mishra and V. R. Desai, "Drought forecasting using stochastic models," Stochastic Environmental Research and Risk Assessment, vol. 19, no. 5, pp. 326–339, 2005.

[10] Mishra AK, Singh VP. A review of drought concepts. Journal of Hydrology 2010; 391(1-2): 202–16. doi:10.1016/j.jhydrol.2010.07.012.

[11] Vernimmen RRE, Hooijer A, Aldrian E, van Dijk AIJM. Evaluation and bias correction of satellite rainfall data for drought monitoring in Indonesia. Hydrology and Earth System Sciences 2012; 16 (1): 133–46. doi:10.5194/hess-16-133-2012.

[12] Alam, A.T.M.J., Rahman, M.S., Saadat, A.H.M., 2013. Monitoring meteorological and agricultural drought dynamics in Barind region Bangladesh using standard precipitation index and Markov chain model. Int. J. Geomech. 3, 511e524.

[13] Condra, G.E. Drought, Its Effects and Measures of Control in Nebraska; Nebraska Conservation Bulletin 25: Lincoln, NE, USA, 1944; 43p. 2.

[14] Wilhite, D.A.; Buchanan, M. Drought as hazard: Understanding the natural and social context. In Drought and Water Crisis: Science, Technology and Management

Issues; Wilhite, D.A., Ed.; CRC Press: New York, NY, USA, 2005; pp. 3–29.
[15] Sholihah, R.I.; Bambang, H.; Shiddiq, D.; Panuju, D.R. Identification of agricultural drought extent based on vegetation health indices of Landsat data: Case of Subang and Karawang, Indonesia. Procedia Environ. Sci. 2016, 33, 14–20.
[16] Huang, C.J.; Zhao, S.Y.; Wang, L.C.; Shakeel, A.A.; Chen, M.; Zhou, H.F. Alteration in chlorophyll fluorescence, lipid peroxidation and antioxidant enzymes activities in hybrid ramie (Boehmeria nivea L.) under drought stress. Aust. J. Crop Sci. 2010, 7, 594–599.
[17] Sruthi, S.; Aslam, M.A.M. Agricultural Drought Analysis Using the NDVI and Land Surface Temperature Data; a Case Study of Raichur District. In Proceedings of the International Conference on Water Resources, Coastal and Ocean Engineering (Icwrcoe 2015), Mangalore, Karnataka, India, 12–16 March 2015.
[18] Bachmair, S., Stahl, K., Collins, K., et al., 2016. Drought indicators revisited: the need for a wider consideration of environment and society. Wiley Interdiscip. Rev. water 3 (4), 516e536.
[19] Dutra, E., Magnusson, L., Wetterhall, F., et al., 2013. The 2010e2011 drought in the Horn of Africa in ECMWF reanalysis and seasonal forecast products. Int. J. Climatol. 33 (7), 1720e1729.
[20] Hao, Z., Hao, F., Singh, V.P., 2016c. A general framework for the multivariate multiindex drought prediction based on multivariate Ensemble Streamflow Predictions (ESP). J. Hydrol. 539, 1e10.

AUTHORS PROFILE

A Abisha completed Bachelor's Degree (B.Tech) in Information Technology in the year 2017 at Jerusalem college of Engineering affiliated with Anna University. Presently she is pursuing Post Graduation (M.E) in Computer Science and Engineering at Jerusalem college of Engineering affiliated with Anna University. Her area of interest include Big Data and Web Development.

Dr. R Beulah Jayakumari, is presently working as Associate professor in Department of Computer Science and Engineering at Jerusalem College of Engineering.she received her B.E in Electrical and Electronic Engineering from Madras university, Chennai, M.E in C Computer Science and Engineering from Anna University, Chennai and Ph.D in Computer Science and Engineering from Anna University. Her research area includes Wireless Sensor Network,Mobile adhoc Network, IOT and Pervasive Computing. She has published more papers in International and National Journals and Conferences. She is a life member of ISTE
.

Dr. D Doreen Hephzibah Miriam is the director of Computational Intelligence Research Foundation. She received her B.Tech in Information Technology from Madras University, Chennai, M.E in Computer Science and Engineering from Anna University, Chennai and Ph.D in Computer Science and Engineering from Anna University. Her research interests include Parallel and Distributed computing, Peer to Peer computing, Grid Computing, Cloud Computing and Big Data Analytics. Previous positions include Professor & Head at Loyola ICAM College of Engineering and Technology in the Department of Information Technology. Assistant professor at the Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Chennai Teaching Research Associate at the Department of Computer Science , Anna University, Chennai and Assistant Professor at the Department of Information Technology, SSN College of Engineering, Chennai. She has published about 30 papers in International and National Journals and Conferences. She is a life member of ISTE. She is a reviewer for Computer and Electrical Engineering Journal and Future Generation Computer Science journal.