International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

Weather Data Analysis Using Big-Data

Dr. D.Thyagarajah, M.E., Ph.D.,[1] S Nivedha,[2]

Assistant Professor, Department of Computer Science and Engineering, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India[1]

B.E Student, Department of Computer Science and Engineering, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India [2]

ABSTRACT: Big Data is a term refers to a collection of large amounts of data which requires new technologies to make potential to get value from it by analysis and capturing method. In every aspect of human life, weather has a lot of importance. It has direct impact on each a part of human society or citizenry. Accurate analytics of weather collecting, storing and processing a large amount of weather data is necessary. So, a scalable data storage platform and efficient or effective change detection algorithms are required to monitor the changes in the environment. An existing or traditional data storage techniques and algorithms are not applicable to process the large amount of weather data. In the proposed system, a scalable data processing framework that is Map-Reduce is used with a climate change detection algorithm which is Spatial Cumulative Sum algorithm and Bootstrap Analysis algorithm called (FWRUT-Frequent weather record Ultra metric tree). This project presents, the large volume of weather data is stored on Hadoop Distributed File System (HDFS) and Map-Reduce algorithm is applied to calculate the minimum and maximum of climate parameters. Spatial Autocorrelation based climate change detection algorithm is proposed to monitor the changes in the climate of a particular city of India.

I. INTRODUCTION

1.1 BIG DATA

Big data could also be an outsized data that it becomes difficult to process the normal database systems. If the data is extremely large, moves in no time, or doesn't fit the structures of the database architectures. to realize value from this data, choose differently to process the data. Big Data generally is defined as high volume, velocity and variety information assets that demand cost-effective, innovative sorts of informatics for enhanced insight and deciding. Big Data is that the frontier of the firm's ability to store, process and access large volume of knowledge it must operate effectively, make decisions, reduce risks, and serve customers.

1.2 CHARACTERISTICS OF BIG DATA

Big data is that the info characterized by four
key attributes volume, variety, velocity and veracity. • Big data implies enormous volumes of knowledge. Data is generated by machines, networks and human interaction on systems like social media that contains terabytes and even petabytes of data.

• Big data extends beyond structured and unstructured data. Unstructured data that has all varieties audio, video, image, text, log files and more.

• Velocity refers to hurry at which the info is generated and moves around. It includes real time working systems like online banking.

• Veracity is that the massive amount of knowledge collected for giant data purposes can cause statistical errors and misinterpretation of the collected information. Purity of the knowledge is critical for value.

1.3 HADOOP

Hadoop is that the foundation for many massive information design.Apache hadoop is Associate in Nursing open supply java programming framework for quick storing and quick process massive information sets with cluster of artifact hardware.

Cluster could be a set of machines in single computer network (Local space Network). The Hadoop is especially well-grooved by the underlying distributed filing system HDFS (Hadoop Distributed File System) and MapReduce layer of parallel programming model engine. Hadoop is employed by varied universities and firms like Google, eBay, Facebook, IBM, LinkedIn and Twitter.

1.4 HDFS AND MAPREDUCE

HDFS might be a reliable distributed classification system that has high-throughput and scalable access to data. MapReduce might be a distributed framework for execution the add parallel. Hadoop has the master/slave style for every method and storage. HDFS might be a specially designed classification system for storing lots of abundance of data sets with cluster of artifact hardware with steaming access pattern.Steaming access pattern means write once and skim any vary of times but don't modification content of files in classification system. HDFS differ from various classification system by its vital. HDFS might be a very big distributed classification system that's very fault-tolerant, provides high output access to the large data and deployed on cheap hardware. HDFS is very used for storing data, and simply adding the number of servers area unit ready to do growth in storage capability and computing power. MapReduce can fill use of the computing resources of each server's constituent, that with efficiency handles with the keep data and calculations. to affect the upper than issues, Google developed the Google classification system (GFS), that might be a distributed classification system style model for method lot of data and created the HDFS and MapReduce. The MapReduce programming model is for method the big amount of data in parallel. Hadoop is degree open provide computer code that disease MapReduce framework, written in Java, originally developed by Yahoo. A MapReduce consists of two tasks notably the Map and shrink task.each Map task takes key-value mix as input associate degreed end up key-value mix as AN output. The input data unit split into various input splits. supported the number of input splits plotter area unit progressing to be assign. Record Reader is degree interface between input split and plotter that's used to convert record into key price mix. plotter will browse key price mix as degree input associate degreed end up key price mix as AN output.

II. EXISTING SYSTEM

In Existing System Rather than considering Apriori and FP-growth, we incorporate the frequent Climate information ultrametric tree (FIU-tree) in the design of our parallel FIM technique. We focus on FIU-tree because of its four salient advantages, which include reducing I/O overhead, offering a natural way of partitioning a dataset, compressed storage, and averting recursively traverse.

Existing parallel mining algorithms for frequent weather record lack a mechanism that enables automatic parallelization, load balancing, data distribution, and fault tolerance on large clusters. As a solution to this problem, we design a parallel frequent weather record mining algorithm called K-NN using the MapReduce programming model. To achieve compressed storage and avoid building conditional pattern bases, K-NN incorporates the frequent Climate information ultrametric tree, rather than conventional FP trees. In K-NN, three MapReduce jobs are implemented to complete the mining task. In the crucial third MapReduce job, the mappers independently decompose weather record, the reducers perform combination operations by constructing small ultrametric trees, and the actual mining of these trees separately. We implement K-NN on our in-house Hadoop cluster. We show that K-NN on the cluster is sensitive to data distribution and dimensions, because weather record with different lengths has different decomposition and construction costs. To improve K-NN's performance, we develop a workload balance metric to measure load balance across the cluster's computing nodes. We develop K-NN-HD, an extension of K-NN, to speed up the mining performance for high-dimensional data analysis. Extensive experiments using real-world celestial spectral data demonstrate that our proposed solution is efficient and scalable.

2.1 DISADVANTAGES OF EXISTING SYSTEM

- Parallel algorithms lack a mechanism that enables automatic parallelization, load balancing, data distribution, and fault tolerance on large computing clusters.
- Not efficient, require more time for mining.

III. PROPOSED SYSTEM

The projected framework for detection of natural action and inclemency station info unit shown on our large environmental condition info is reduced with Hadoop MapReduce framework.

- Proposed spatial accumulative add rule is utilized to observe the day wise changes inside the climate from a number of years. MapReduce rule is utilized to make a table to root.
- Cumulative add technique is use to go looking out forceful changes inside the common of quantity of interest.Here accumulative add technique is utilized to observe the changes inside the climate.K suggests that rule is utilized to observe the day wise changes inside the climate from a number of years.
- They can merely resolve the data by characteristic and analysing the files.

The driver that originated job submits it, and sit up for method to finish. it's taken from a configuration file to specify the input or output directories.

additionally, it will settle for script supported plotter and reducer while not re-compilation

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

3.1 MAPPER

The mapping could be a easy method therein the variables that matched bound are going to be sent to the reducer. It contemplates mappers is acts sort of a distributed search capability and pull (key, value) combines of a file. The computer file format scaner of hadoop opens files and that starts to read file for (key, value) pairs. Once it verifies (key, value) pair, it reads each key and values that passes to the clerk and mapping operator that is employed to filtrate (key, value) pairs that Since clerk isn't a vicinity of Hadoop that scan knowledge. The clerk is collection knowledge from computer file format scaner and computer file format reader is changed to read sequenced files. This routine opens a file and performs an easy loop to scan every (key, value) among file. the required key if filter matched, then values that ar scan into memory and passed to the clerk. If filter isn't match, then values were skipped.

The ensuing (key, value) pairs that matched the standards is analyzed and forward to reducer with sequencing and complete mapping method, once a (key, value) object has created, a comparator If knowledge is combined, a gaggle comparator is additionally required. during a partitioner should be created so as to handle partitioning knowledge into teams of sorted keys.With of these parts in Hadoop takes the (key, value) pairs that is made by exploitation mappers and cluster and type them as mere method. Hadoop assumes that every one values share a key can sent to same reducer and one operation over an oversized knowledge set can use on one reducer, this offers United States of America end in variety of output files

RULE

The Map-Reduce method that is dead for minimum and most operation on the National climatical knowledge is given as follows:

1.Weather knowledge set files is inserted into sequence files on head node of Hadoop Distributed file

2.The sequence files that ar loaded into Hadoop classification system with duplicate issue

3.The job provides Map-Reduce operation that is submitted to move node to run. the pinnacle node schedules with job huntsman and on cluster jobs that is to be run. Hadoop distributes mappers to any or all knowledge nodes that contains knowledge to investigate.

4.For reading the input format reader parades every sequence file that passes all the (key, value) pairs to map operate on every node.

5.The clerk determines, if key matches the standards for question. If clerk keeps (key, value) pairs for delivery to the reducer. The keys, values in a very file ar browse and conjointly analyze by victimisation the clerk.

6.The (key, value) pairs can match question ar to sent to reducer and reducer activity the typical operate on the sorted pairs to form final (key, value) combine results.

7.In a Hadoop Distributed classification system as a sequence file, the ultimate output is hold on.

IV. MODULES DESCRIPTION

4.1 PREPROCESSION

As a volume of database increases day by day traditional frequent weather record mining algorithms becomes inefficient. As a solution to this problem parallel mining of frequent weather record s using FWRUT algorithm is implemented on MapReduce framework. Here we using FWRUT algorithm rather than traditional FP-Tree algorithm because to avoid building conditional patterns and to achieve compressed storage. We build this using Hadoop framework. The working flow of FWRUT algorithm on MapReduce framework consists of three MapReduce job. Synthetic datasets are used for the experimental analysis.

4.2 FREQUENT ONE WEATHER RECORD S GENERATION

The first MapReduce job is accountable for mining all frequent one-weather record s. A group action info is partitioned off into multiple input split files keep by the HDFS across multiple knowledge nodes of a Hadoop cluster. range of clerk are going to be dead supported range of input split. every clerk consecutive reads every group action from its native input split, wherever every group action is kept within the format of key worth pair by the record reader. Then, mappers work out the frequencies of Climate info and generate native one-weather record.

Next, these one-weather record s with a similar key emitted by totally different mappers area unit sorted and incorporate in a very specific reducer, that any produces international oneFinally, rare Climate info area unit cropped by applying the minsupport; and consequently, international frequent one-weather record s area unit generated and written within the kind of pair because

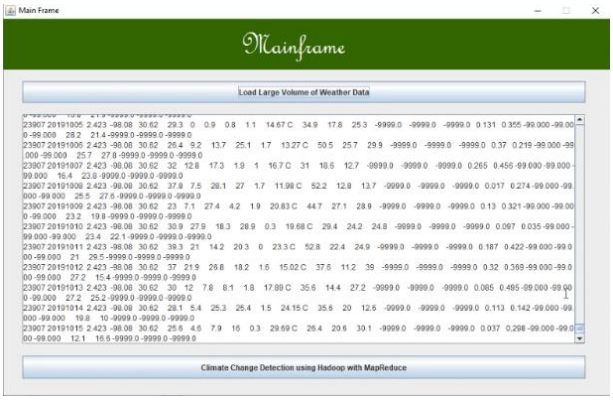

Fig. 1. Loaded dataset

4.3 ALL K WEATHER RECORD GENERATION

Given frequent one-weather record s generated by the primary MapReduce job, the second MapReduce job applies a second spherical of scanning on the information to prune occasional Climate data from

The second job marks associate weather record as a k-weather record if it contains k frequent Climate data ($2 \leq k \leq M$, wherever M is that the largest every plotter of the second job takes transactions as input.

Then, the plotter emits a group of try, within which weather record s consists of the quantity of the Climate data created by pruning and also the set These pairs obtained by the second MapReduce job's mappers area unit combined and shuffled for the second job's reducers. A lot of formally, the output of the second MapReduce job is pair

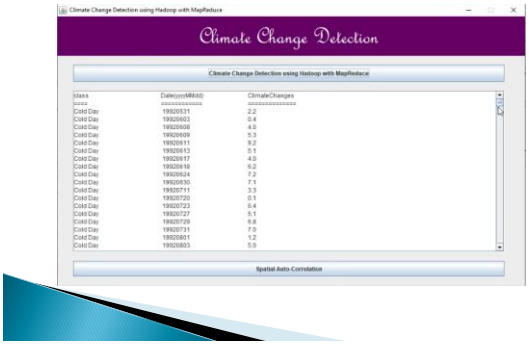

Fig. 2. Weather detection

4.4 FREQUENT K WEATHER RECORD GENERATION

The third MapReduce job a computationally costly section is devoted to: 1) moldering weather record s; 2) constructing k-FIU trees; and 3) mining frequent weather record s. the most goal every|of every} clerk is twofold: 1) to decompose each k-weather record obtained by the second MapReduce job into an inventory of small-sized sets, wherever the quantity

of every set is anyplace between a pair of to k a pair of one ANd 2) to construct an FIU-tree by merging native decomposition results with a similar length.

The third MapReduce job is very climbable, as a result of the decomposition procedure of every clerk is freelance of the opposite mappers. In alternative words, the multiple mappers will perform the decomposition method in parallel. Such AN FIU-tree construction improves information storage potency and I/O performance; the advance is created doable because of merging a similar weather record s prior to victimisation tiny FIU trees. The Map operate of the third job generates a collection of key/value pairs, within which the secret's the quantity of Climate info in AN weather record and also the price is AN FIU-tree that's comprised of non leaf and leaf nodes. Non leaf nodes embody item-name and node-link; leaf nodes embody item-name and its support. In doing therefore, weather record s with a similar range of Climate info square measure delivered to one reducer. By parsing the key-value combine (k2, v2), the reducer is to blame for constructing k2-FIU-tree and mining all frequent weather record s solely by checking the count price of every leaf within the k2-FIU-tree while not repeatedly traversing the tree. Here, the main points on the operate of t-FIU-tree generation (t-weather record) are often found. the decompose () operate could be a algorithmic one, moldering

AN h-weather record into an inventory of k-weather record s, wherever k is AN number between a Categories of Free Tools

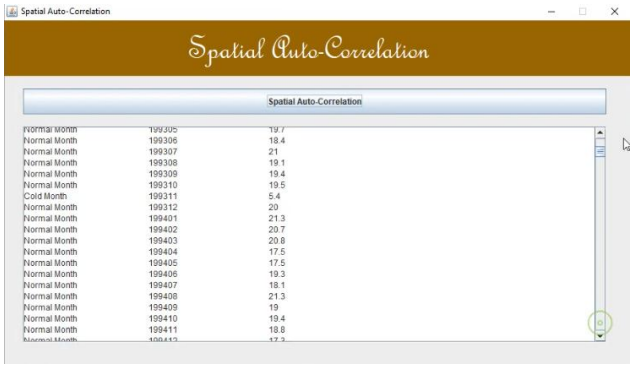

Fig. 3. Weather Regeneration

V. CONCLUSION

- The traditional or existing systems which processes millions of records is a time-consuming process. So here Hadoop with Map-reduce, weather data can be analyzed effectively. Map reduce is a framework which is parallel and distributed systems across large dataset. Using Map-Reduce with Hadoop helps in removing scalability issues.
- This technology which is used to find huge datasets has the potential for significant enhancement to analyze weather.
- The major advantage of Map-Reduce with Hadoop framework speeds up the processing of data, where the volume of data is increasing every day. we intend to use the climate change values to predict the diseases is the future work of the proposed system.

REFERENCES

1. Abdi.H and Williams.L.H , “Principal component analysis”,(2018) Wiley Inter-discipl. Rev., Compute. Statist., vol. 2, no. 4, pp. 433-459.
2. Anusha.K , UshaI RaniI.K , “Big Data Techniques for Efficient Storage and Processing of Weather Data”(2017), International Journal for Research inApplied Science&EngineeringTechnology(IJRASET), Volume 5 Issue VII
3. Basvanth Reddy and Prof Patil.B.A , “Weather Prediction on Big Data Using Hadoop Map Reduce Technique”(2016), IJARCCE, ISSN: 2278-1021Volume-05, Issue-06, Page No (643-647)
4. Bendre.M.R and Thool.V.R , “Analytics, challenges and applications in big data environment: A survey”,(2018) J. Manage. Anal., vol. 3, no. 3,pp. 206-239.
5. BenjellounF.Z ,. Lahcen.A, and Belfkih.S, “An overview of big data opportunities, applications and tools”,(2017) in Proc. IEEE ISCV Conf., Fez,Morocco,pp 16.
6. Ding.C and He.X, “Cluster structure of K-means clustering via principalcomponent analysis”,(2016) Lect. Notes Comput. Sci., vol. 46, no. 4, pp. 414-418.
7. Ding.C and He.X, ”K-means clustering via principal component analysis”, (2018)in Proc. 21st Int. Conf. Mach. Learn., p. 29.
8. Delchambre.L.A, “Weighted principal component analysis: A weighted covariance eigendecomposition approach”(2016) Monthly Notices Roy.Astronom. Soc., vol. 446, no. 4, pp. 3545-3555.
9. Ghinita.G, Kalnis.P, Khoshgozaran.A. Shahabi.A, and Tan.K.L, “Private queries in location based services: Anonymizers are not necessary”,(2018) in Proceedings of the ACM on Management of Data pg.556.
10. Guo.P, Wang.K, LuoA.L , and Xue.L, “Computational intelligence for big data analysis: Current status and future prospect”,(2018)J. Softw., vol. 26,no. 11, pp. 3010-3025.
11. Hammer.P.C,” Adaptive Control Processes” ,(2017) A Guided Tour, R. Bellman,Ed. , pp. 315-316.
12. Hoffmann.H, “Kernel PCA for novelty detection”,(2017) Pattern Recognit.,vol. 40, no. 3, pp. 863-874.
13. Khalid Adam Ismail, Mazlina Abdul Majid,“Big Data Prediction Framework for Weather Temperature Based on MapReduce Algorithm” (2016)IEEEConference on Open Systems (ICOS), Vol.7, Issue 1,PP 244-304
14. MeenaAgrawal, .Pandey A.K, Agrawal C.P, “A Hadoop based Weather Prediction Model for Classification of Weather Data”(2017),IEEE Vol-24,
15. Riyaz P.A., SurekhaI MariamI Varghese, “Leveraging Map Reduce With Hadoop for Weather Data Analytics(2016)enkatesan, “A Survey Project on Climate Changes Prediction Using Data Mining”, IJARCCE, ISSN:2278-1021 Volume-05, Issue-02, Page No (294296),
16. Mr. Sunil Navadia, “Weather Prediction: A novel approach for measuring and analyzing weather data”, (2017)IEEE International conference on ISMAC
17. Veershetty Dagade, Mahesh Lagali , Supriya Avadhani and PriyaKalekarI, “Big Data Weather Analytics Using Hadoop”,(2015) International Journal ofEmerging Technology in Computer Science & Electronics, Volume-14, Issue-02.
18. Ye Ding, Yanhua Li, “Detecting and Analyzing Urban Regions with High Impact of Weather Change on Transport”,(2016) IEEE Transactions on BigData .
19. RuI Zhang ,”A principal component Analysis Algorithm Based on Dimension Reduction Windows”,(2018) vol-28,IEEE
20. G. Zhong, I. Goldberg, and U. Hengartner, “Louis, lester and pierre:Three protocols for location privacy”, (2017)in Proceedings of the Privacy Enhancing Technologies Symposium, vol-34 p.543.