IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-12 E-ISSN: 2347-2693

A Big Data Analysis: Weather Forecasting Data Analaysis With Fixed Width Clustering Algorithm

A. Patil[1*], A. Palve[2]

[1*]Dept. Of Computer Engineering, SITRC (Savitribai Phule Pune University), Nashik, India
[2]Dept. Of Computer Engineering, SITRC (Savitribai Phule Pune University), Nashik, India

[*]*Corresponding Author: ajit.patil1091@gmail.com,*

Available online at: www.ijcseonline.org

Received: 28/Nov/2017, Revised: 10/Dec/2017, Accepted: 21/Dec/2017, Published: 31/Dec/2017

***Abstract*—** In today's digital world remote senses daily generate large amount of real time data know as Big-Data, wherever insight data encompasses a potential significance if collected and aggregative effectively. There is a great deal added to real-time remote sensing Big Data than it seems at first, and extracting the useful information in an efficient manner leads a system toward a main computational disputes, such as to analyze, aggregate, & store, where data are remotely gathered. Keeping in view the above mentioned factors, there's a requirement for planning a system design that welcomes each real-time, additionally as offline processing. In this paper we propose efficient and scalable solution to analyze pent bytes of data across an extremely wide increasing wealth of weather variables. In this research we are working on data analysis using Apache Hadoop and Java . Extensive experiments are carried out to find out the best tools among Distributed computing using Pig and Hive Queries. The proposed architecture has the potential of dividing, load balancing, & parallel processing of only utile data. Thus, it results in effectively analyzing real-time remote sensing Big Data using earth observatory system. Furthermore, the proposed architecture has the capability of storing incoming raw data to perform offline analysis on largely stored dumps, when required. Fixed width clustering algorithm is used to improve the accuracy of results.

***Keywords*—** Big Data, data analysis decision unit (DADU), data processing unit (DPU), land and sea area, offline, real-time, remote senses, remote sensing Big Data acquisition unit (RSDU).

I. INTRODUCTION

Big Data & its anatomy is gaining a huge interest by the researchers, mainly focusing on research disputes robustly corresponded to authentic applications, like processing, modelling, querying, mining, & distributing large-scale repositories[1]. "Big Data" assorts the particular kinds of data sets containing formless data, which lie in data layer of skilled computing applications [2] & the Web [3]. The data ordered within the fundamental layer of of these technical computing application assumptions have some exact identities in common, like 1) large scale data, that refers to the scale & the data warehouse; 2) scalability problems, that check with the application's doubtless to be running on giant scale (e.g., big Data); 3) (ETL) technique from low, raw data to well thought-out data up to positive magnitude; & 4) development of easy explicable analytical over big data warehouses with a read to deliver an intelligent & important data for them. Big data are sometimes generated by on-line dealings, video/audio, email, variety of clicks, logs, posts, scientific data, remote access sensory knowledge, mobile phones, & their applications [4], [5]. Advancement in big data sensing & engineering revolutionizes the method remote data collected, processed, analyzed, & managed [6-]–[9]. Particularly, last designed sensors utilized in the earth & planetary observatory system are generating continuous stream of data. Moreover, majority of labor are drained the assorted fields of remote sensory satellite image data, like change detection [10], gradient-based edge detection [11], region similarity primarily based edge detection [12], & intensity gradient technique for economical intraprediction [13].

Big data analysis is someway a hard job than finding, characteristic, understanding, & citing data [14]. Having a large-scale data, all of this should happen in an exceedingly mechanized manner since it needs various arrangement yet as linguistics to be articulated in ways of computer-readable format. However, by analyzing easy data having one data set, a mechanism is needed of the way to style a database.

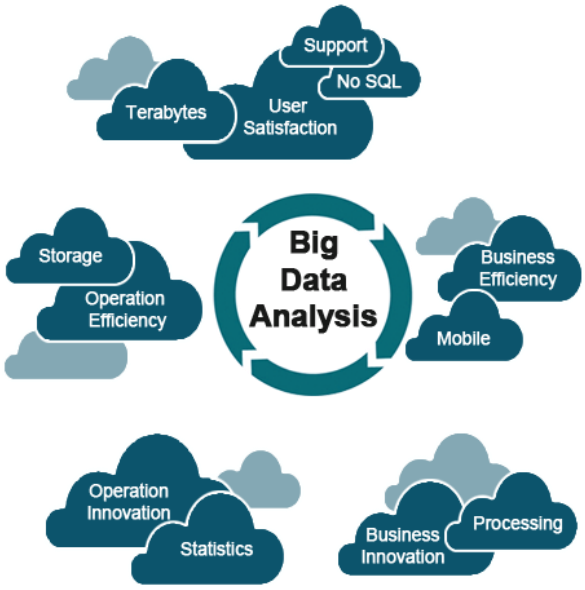

Fig 1. Big Data Analytics

It must have secondary routes to store all of identical data. In such conditions, the mentioned style might need a plus over differents sure method & attainable drawbacks for a few other functions. In order to address these needs, various analytical platforms have been provided by relational databases vendors [15]. These decks come in various shapes from software only to analytical services that consort in third-party feasted environment. In remote access networks can develop an covering quantity of raw data. We called it to the first step, i.e., data acquisition, where the unnecessary data is filtered compacted by orders of magnitude. With a perspective to use such filters, useful information is not disposed. For instance, in consideration of new reports, is it adequate to keep that information that is mentioned with the company name? Alternatively, is it necessary that we may need the entire report, or simply a small piece around the mentioned name? The second challenge is by default generation of accurate metadata that describe the composition of data & the way it was collected & analyzed. Such quite metadata is tough to research since we tend to might have to understand the source for every information in remote access. Usually, the data gathered from remote areas aren't in a format prepared for analysis. So, the next step cites us to data extraction, that drags out the helpful information from the underlying sources & delivers it in an exceedingly structured formation appropriate for analysis. However, this is often distant from reality; generally we've to touch upon inaccurate data too, or a number of the data may well be inexact.

A great deal of interest within the area of big data & its analysis has up, primarily driven from in depth variety of analysis challenges strappingly associated with bonafide applications, like prototyping, processing, querying, mining, & distributing large-scale repositories. The term massive data classifies specific forms of data sets comprising formless data, that dwell in information layer of technical computing applications & the net. the data keep within the fundamental layer of of these. technical computing application eventualities have some precise individualities in common, like giant scale information, that refers to the scale & the info warehouse; measurability problems, that consult with the applications doubtless to be running on giant scale (e.g., massive Data) (ETL) methodology from low, {raw data|} to well thought-out data up to sure extent; & development of uncomplicated explainable associate analytical over big data warehouses with a read to deliver an intelligent & significant data for them. big data are sometimes generated by on-line dealings, video/audio, email, variety of clicks, logs, posts, remote access sensory information, mobile phones, & their applications. These data are accumulated in databases that grow additional normally & become sophisticated to confirm, form, store, manage, share, process, analyze, & visualize via typical info computer code tools.

The rest of the paper is organized as follow, Section II contains the deep literature survey on the existing system on big data. Section III states the motivation for behind the development of the project, whereas the proposed approach is explained in Section IV. Results and observations are discuss in the section V and the paper is concluded in the last section i.e. Section VI.

II. RELATED WORK

In The peer research big data analytics survey is presented in the below pie chart fig 1.

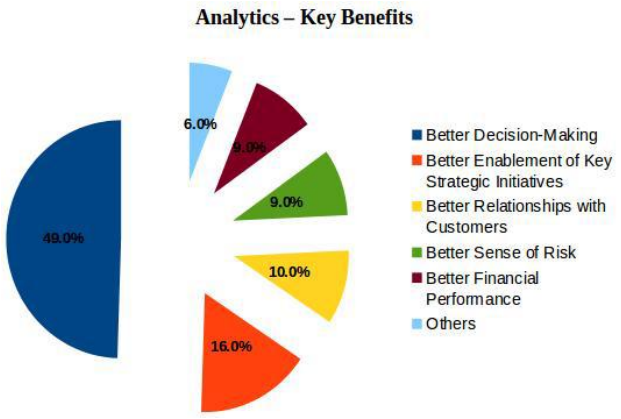

Fig 2. Peer-Research Big Data Analytics Survey

Researcher have gain interest in Big Data so many systems are being proposed & deep study is being done on Big Data.

Authors in [16] describes a short explore the Arduino microcontroller & a number of its applications & however it are often employed in learning. Arduino is associate open-source microcontroller employed in electronic prototyping. Arduino hardware & its parts shall be checked out. software system & the surroundings that Arduino works on are each faced at too. Few usages are assumed as cases which will facilitate create learning Arduino additional fascinating. this could be used as a serious thanks to encourage students & others to be told additional regarding physical science & programming. Authors delineate why a cloud-based result's required, describe our model implementation, & report on some example applications we've got disbursed that prove personal information possession, control, & analytics. He address these problems by planning & implementing a cloud-based design that gives shoppers with quick access & fine-grained management over their usage information, additionally because the ability to analyse this information with algorithms of their selecting, as well as third party applications that analyse that information in a very privacy conserving fashion[17].

Panagiotis D. Diamantoulakis implements the large information Analytics for Dynamic Energy Management in sensible Grids. The sensible electricity grid alters a two-way flow of power & information among suppliers & shoppers so as to ease the ability flow improvement in terms of economic potency, responsibility & property. This infrastructure permits the consumers & the micro-energy producers to require a additional active role within the electricity market & the dynamic energy management (DEM). the foremost vital challenge in a very smart grid (SG) is the way to trespass of the users' participation so as to cut back the value of power.[9] L. Aniello explore the thought of a framework investing multiple information sources to boost protection capabilities of CIs. Challenges & opportunities square measure hashed out on 3 main analysis directions: i) use of distinct & heterogeneous information sources, ii) watching with reconciling roughness, & iii) attack modelling & runtime combination of multiple information analysis techniques.[12]

III. MOTIVATION

As per the statics studied on[16], there's exponential amendment within the data rates generated on digital universe. With a read in using current tools & technologies to investigate & store, a vast volume of information don't seem to be up to the mark , since they're unable to extract needed sample knowledge sets. Therefore, we have a tendency to should style an field of study platform for analyzing each remote access real-time & offline data. once a commercial enterprise will pull-out all the helpful data obtainable within the big data instead of a sample of its data set, in this case, it's an authoritative profit over the market competitors. big data analytics helps us to achieve insight & create higher selections. Therefore, with the intentions of exploitation big data, modifications in paradigms are at utmost. To support our motivations, we've represented some areas wherever big data will play a crucial role. Understanding surroundings needs huge quantity of data collected from varied sources, like remote access satellite observant earth characteristics [measurement data set (MDS) of satellite data like images], sensors watching air & water quality, scientific discipline circumstances, & proportion of co2 & alternative gases in air, & so on. Through linking all the info drifting like co2 emission, increase or decrease on greenhouse effects & temperature, will be realized. In tending eventualities, medical practitioners gather huge volume of information regarding patients, case history, medications, & alternative details. The preceding data are accumulated in drug-manufacturing firms. the character of those data is extremely complicated, & generally the practitioners are unable to indicate a relationship with alternative data, which ends up in missing of vital data. With a read in using advance analytic techniques for organizing & extracting helpful data from big data ends up in customized medication, the advance big data analytic techniques provide insight into hereditarily causes of the illness.

IV. REMOTE SENSISNG BIG DATA ANALYTICS ARCHITECTURE

The planned Remote Sensing big data Analytics design is pictured in fig 1. The design is enforced exploitation fixed width clustering algorithm which is able to facilitate in improvising the results. The input to the system comes from numerous dataset i.e. Austine, Bronse, Cape, Charlotte. in spite of terribly recent emergence of big data design in scientific applications, varied efforts toward big data analytics design will already be found within the literature.

.

Fig 3. Remote sensing earth observatory image.

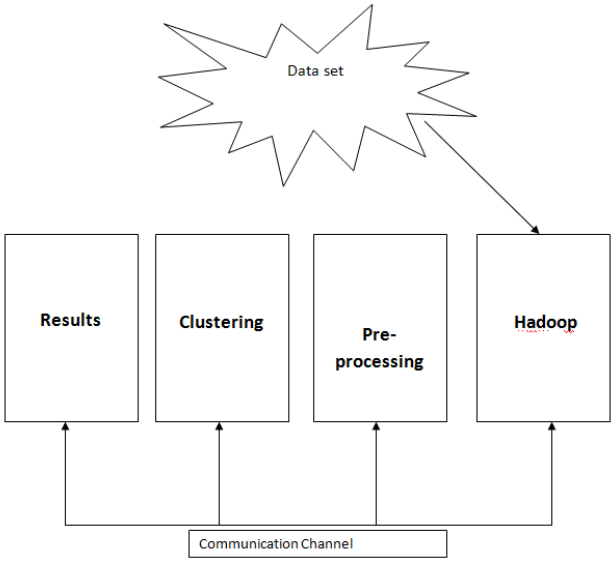

Fig 4. Remote Sensing Big Data Architecture

The design is divided into three elements i.e. Remote sensing data acquisition unit(RSDU), data processing unit (DPU), data analysis & decision unit (DADU).

A. Remote Sensing Big Data Acquisition Unit

Remote sensing elevates the elaboration of earth observatory system as efficient parallel data acquisition system to fulfil specific machine demands. the earth & space Science Society originally approved this answer because the normal for parallel processing during this explicit context. As satellite tools for Earth observation incorporated additional advanced qualifications for mended huge knowledge acquisition, shortly it absolutely was discerned that ancient process processing} technologies couldn't give sufficient power for processing such reasonably data. Therefore, the requirement for data processing of the large volume of information was needed, that may expeditiously analyze the big data. For that reason, the advised RSDU is brought within the remote sensing big data design that collects the info from numerous satellites round the globe.

It is potential that the received data are distorted by scattering & absorption by numerous part gasses & mud particles. we tend to assume that the satellite will correct the incorrect data. physicist or SPECAN algorithms are employed by the remote sensing satellite uses to create the data into image format. For economical knowledge analysis, remote sensing satellite preprocesses knowledge below several things to integrate the info from numerous sources, that not solely decreases storage value, however conjointly improves analysis accuracy. Some relative knowledge preprocessing techniques are data integration, data cleansing, & redundancy elimination. when preprocessing part, the collected knowledge are transmitted to a ground station exploitation downlink channel. This transmission is directly or via relay satellite with associate degree acceptable following antenna & communication link in an exceedingly wireless atmosphere.

B. Data Processing Unit

In data processing unit (DPU), the filtration & load balancer server have 2 main obligations, like filtration of information & load reconciliation of process power. Filtration identifies the helpful data for analysis since it solely permits helpful info, whereas the remainder of the info square measure blocked & square measure discarded. Hence, it leads in enhancing the functioning of the total planned system. Apparently, the load-balancing a part of the server provides the power of dividing the total filtered knowledge into elements & assign them to varied process servers. The filtration & load-balancing formula differs from analysis to analysis; e.g., if there's solely a requirement for analysis of ocean wave & temperature knowledge, the measure of those represented knowledge is filtered out, & is mesmeric into elements. every process server has its formula implementation for process incoming section of information from FLBS. every process server makes applied math calculations, any measurements, & performs different mathematical or logical tasks to come up with intermediate results against every section of information. Since these servers perform tasks severally & in parallel, the performance planned system is dramatically increased, & the results against every section square measure generated in real time.

C. Data Analysis & Decision Unit

DADU contains 3 major parts, like aggregation & compilation server, results storage server(s), & decision making server. once the outcomes are prepared for compilation, the process servers in DPU send the partial results to the aggregation & compilation server, since the combined results aren't in organized & compiled kind. Therefore, there's a requirement to combination the connected results & organized them into a correct kind for additional process & to store them. within the projected design, aggregation & compilation server is supported by totally different algorithms that compile, organize, store, & transmit the results. Again, the formula varies from demand to demand & depends on the analysis wants. Aggregation server stores the compiled & organized results into the results storage with the intention that any server will use it because it will method at any time. The aggregation server conjointly sends a similar copy of that result to the call-making server to method that result for creating decision. The decision-making server is supported by the choice formula, that surprise in contrast to things from the result, &

then create numerous selections (e.g., in our analysis, we tend to analyze land, sea, & ice, whereas different finding like storms, Tsunami, earthquake can even be detected). the choice formula should be sturdy & correct enough that expeditiously turn out results to get hidden things & create selections. the choice a part of the design is critical since any tiny error in decision-making will degrade the potency of the total analysis. DADU finally exhibits or broadcasts the conclusion, in order that any application will use those selections at real time to form their growth.

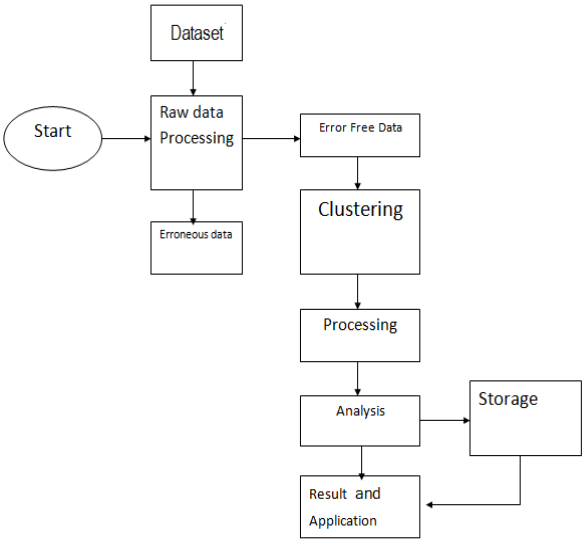

Fig 5 Flowchart of the remote sensing Big Data architecture.

V. RESULTS & OBSERVATION

The architecture is developed using Hadoop & java. The graph shown in result analysis shows that , For small size data, i.e., less than 200 MB, the Hadoop execution takes more average processing time to process 1 MB data of the product, while a simple Java implementation is efficient in this case. However, when the product size is increasing, the average process time starts decreasing in Map Reduce implementation. Moreover, when the product size exceeds 200 MB, it produces better results as compared with simple Java implementation. Table 1 shows the results comparison of the between the CPO K-means and our proposed method of fixed width, the results depicts the performance of the proposed system.

Dataset	CPO K-mean	Fixed Width
Austine	45.15	11.63
Brigham_city	42.49	12.87
Cape_Charles	41.68	10.24
Charlottesville	45.69	15.15
Sundance	43.72	11.55

Table 1. Result Comparison

The developed system is verified on the above listed 4 dataset & the results graphs generated depicts that our suggested architecture has enhance the results.

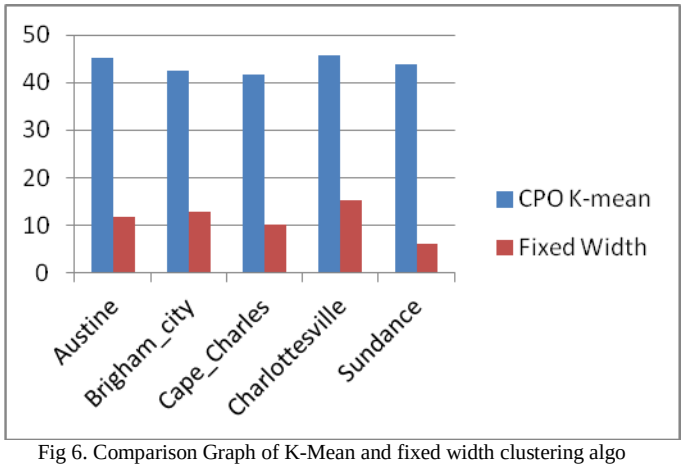

Fig 6. Comparison Graph of K-Mean and fixed width clustering algo

VI. CONCLUSION

The proposed architecture is developed using fixed width clustering algorithm to improve the accuracy of results. Details literature study is done which state the disadvantages of the existing system & it also allows researchers for any type of remote sensory Big Data analysis by formulating algorithms for each degree of the architecture relying on their analysis requirement. Comparison shown in the results section depicts the results for 4 dataset(11.63, 12.87, 10.24, 15.15, 11.55). The improved analysis of the big volumes of data that are becoming usable, there is the potential for making quicker advances in many scientific disciplines & improving the profitability & success of many enterprises.

REFERENCES

[1] M. M. U. Rathore, A. Paul, A. Ahmad, B. W. Chen, B. Huang and W. Ji, "Real-Time Big Data Analytical Architecture for Remote Sensing Application," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4610-4621, Oct. 2015.

[2] H. Herodotou et al., "Starfish: A self-tuning system for Big Data analytics," in Proc. 5th Int. Conf. Innovative Data Syst. Res. (CIDR), 2011, pp. 261–272.

[3] K. Michael and K. W. Miller, "Big Data: New opportunities and new challenges [guest editors' introduction]," IEEE Comput., vol. 46, no. 6, pp. 22–24, Jun. 2013.

[4] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.

[5] R. D. Schneider, Hadoop for Dummies Special Edition. Hoboken, NJ, USA: Wiley, 2012.

[6] R. A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, 2nd ed. New York, NY, USA: Academic Press, 1997.

[7] D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ, USA: Wiley, 2003.

[8] C.-I. Chang, Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Norwell, MA, USA: Kluwer, 2003.

[9] J. A. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction. New York, NY, USA: Springer, 2006.

[10] J. Shi, J. Wu, A. Paul, L. Jiao, and M. Gong, "Change detection in synthetic aperture radar image based on fuzzy active contour models and genetic algorithms," Math. Prob. Eng., vol. 2014, 15 pp., Apr. 2014.

[11] A. Paul, J. Wu, J.-F. Yang, and J. Jeong, "Gradient-based edge detection for motion estimation in H.264/AVC," IET Image Process., vol. 5, no. 4, pp. 323–327, Jun. 2011.

[12] A. Paul, K. Bharanitharan, and J.-F.Wang, "Region similarity based edge detection for motion estimation in H.264/AVC," IEICE Electron. Express, vol. 7, no. 2, pp. 47–52, Jan. 2010.

[13] A.-C. Tsai, A. Paul, J.-C. Wang, and J.-F. Wang, "Intensity gradient technique for efficient intra prediction in H.264/AVC," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 5, pp. 694–698, May 2008.

[14] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with Big Data," in Proc. 38th Int. Conf. Very Large Data Bases Endowment, Istanbul, Turkey, Aug. 27–31, 2012, vol. 5, no. 12, pp. 2032–2033.

[15] [20] P. Chandarana and M. Vijayalakshmi, "Big Data analytics frameworks," in Proc. Int. Conf. Circuits Syst. Commun. Inf. Technol. Appl. (CSCITA), 2014, pp. 430–434.

[16] Adamu Galadima, D. Sacc, and J. D. Ullman, \Big Data: A research agenda," in Proc. Int. Database Eng. Appl. Symp. (IDEAS13), Barcelona, Spain, Oct. 0911, 2013.

[17] C.-I. Chang, "Hyperspectral Imaging: Techniques for Spectral Detection and Classi_cation", Norwell, MA, USA: Kluwer, 2003.

