International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Environmental Data Analytics for Empirical Values on Environmental Issues

M. Naveen Babu, A.V. Krishna Prasad

Abstract - In the contemporary era, big data is highly regarded as the driver to promote productivity, efficiency and innovation. Emergence of big data and data science paved way for comprehensive analysis of data for obtaining business intelligence. Big data analytics has become crucial for enterprises to garner accurate knowhow for making well informed decisions. The cloud-big data ecosystem has been realized and thus it became easier to deal with big data and it's processing as the storage and processing are outsourced to cloud. Different cloud computing platforms like Amazon AWS, Google cloud and Microsoft Azure made it a reality to work with big data which provides comprehensive understanding of data. With big data, environmental issues especially air pollution measurement and prediction can add value to existing infrastructure so as to improve the quality of prediction and also help in making strategic decisions. This paper represents the present state of the art on usage of big data analytics for adding value to different industries focusing more on environmental issues. It also provides the empirical values made with Apache Flink and Apache Spark for handling environment data. The preliminary results revealed that these frameworks play crucial role in processing big data.

Keywords – Big data, big data analytics, environmental issues, Apache Flink, Apache Spark

I. INTRODUCTION

Big data is the voluminous data with other features like exponential growth and having different formats of data [1]. Data became big data when it meets the characteristics of big data. In this information era, organizations are not willing to lose data. Therefore, they wanted to maintain data as much as possible. This ambition of organizations is fulfilled with the emergence of cloud computing technology where scalable computing resources are provided on demand. Moreover cloud resources can be used without investment and without time and geographical restrictions. However, the resources are to be used in pay per use fashion. Cloud computing became affordable with technologies like virtualization. Therefore enterprises are aligning their Information Technology (IT) wings to cloud computing and big data analytics. This has become an essential move to stay focused and competent in the scenario of globalization and acute competition in the real world.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
M.Naveen Babu, Research Scholar, Department of CSE, KL (Deemed to be) University, Vijayawada, A.P, India.
A.V. Krishna Prasad, Research supervisor, Department of CSE, KL (Deemed to be) University, Vijayawada, A.P, India..

Having understood the importance of big data, it is important to know that environmental pollution is causing many problems to humans. Health issues are increasing in an unprecedented way due to pollution in urban areas.

The sources of pollution data provide huge volumes of data that needs to be maintained and processed. Thanks to big data and cloud computing, for providing platforms to enable organizations to deal with complex problems. In this context, this research paper focused on big data analytics and the present state of the art on big data processing frameworks and the applications of big data for air pollution monitoring and prediction.

From the literature [5], [7], [10], [14] it is understood that big data is crucial for solving many real time problems. With respect to environment issues big data is widely used as explored in [4], [13], [18], [24], [47] and [48]. It is understood that air pollution monitoring needs big data analysis. The rationale behind this is that voluminous data is being produced by different sources or sensors that work on gathering data related to environment. Various approaches on statistics that need big data analysis are found in the literature. Since environmental issues are linked to health issues to humans, it assumes significance and big data plays crucial role in solving such problems.

The remainder of the paper is structured as follows. Section 2 focuses on big data and the need for big data analytics. Section 3 presents an overview of cloud computing as big data eco system. Section 4 presents distributed programming frameworks like Apache Flink and Apache Spark for processing big data. Section 5 throws light on big data for solving environmental issues. Section 6 presents the results of empirical study made on environmental big data using Apache Flink and Apache Spark. Section 7 concludes the paper and provides directions for future work.

II. BIG DATA AND BIG DATA ANALYTICS

Big data refers to huge amount of data in the form of collection of datasets of different types such as structured, unstructured and semi-structured data [1]. There are plenty of sources of big data such as Online Social Networks (OSNs) like Facebook and Twitter, sensor networks, satellites and so on. As enterprises are willing to maintain complete data for mining business intelligence, there is need for big data.

Moreover, the presence of cloud computing resources enables Organizations to work with big data and big data analytics. Big data has three important characteristics as shown in Figure 1. They are known as Volume, Velocity and Variety.

Volume refers to data which is voluminous (measured in peta bytes). Variety refers to the availability of big data in structured, unstructured and semi-structured formats. Velocity indicates that the data is streaming continuously with exponential growth.

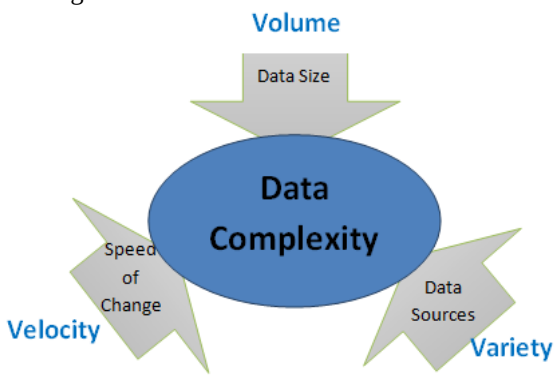

Figure 1: Characteristics of big data

As the characteristics of big data reveal, it is available in different formats and it is growing continuously and it is available in large volumes. Of late two more characteristics have been associated with big data. They are known as veracity and value. Veracity is the feature that needs to be associated with big data. It refers to making the data reliable by cleaning it or by removing noise. Value refers to the advantage of processing big data to an organization. In other words, how the big data is adding value to the organization is given importance. Big data has potential to gain comprehensive business intelligence. If big data is not considered, it may lead to inaccurate BI which may lead to flawed decision making. This phenomenon is illustrated in Figure 2.

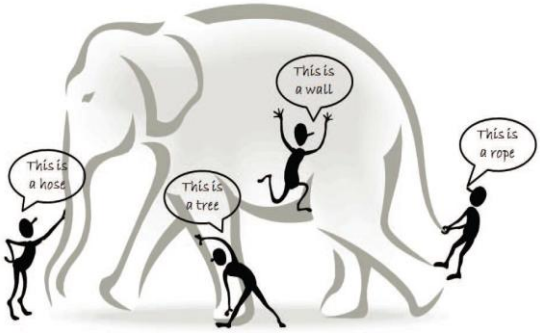

Figure 2: Shows the possibility of biased conclusions if big data is not considered

When big data is not considered, it may lead to biased conclusions. The decisions will not be accurate. The whole picture here is an elephant. However, people understood it like a tree or rope or hose etc. They have not considered the data as a whole. It emphasizes the need for considering big data for processing. Therefore big data has potential to enable highly accurate decisions in industries like economics, road traffic, environmental studies, finance, healthcare, manufacturing and retail to mention few. In all these industries enormous volumes of data is being produced and processing such data can provide required BI for organizations. However, to realize big data analytics, it is essential to have support of cloud computing platforms.

III. CLOUD COMPUTING

According to National Institute of Standards and Technology (NIST) cloud computing is a novel model of computing with certain characteristics such as on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. It has services models pertaining to software, infrastructure and platform and deployment models in terms of private cloud, public cloud, community cloud and hybrid cloud [1]. As mentioned earlier, cloud computing is service oriented and different kinds of services are provided. In the same fashion different deployment models are made available to meet the needs of organizations of different size. Before looking into the service layers and deployment models, Figure 3 here shows the advantages of cloud computing. The benefits include low cost, flexibility, collaboration, software update, mobility, data loss prevention, security and productivity.

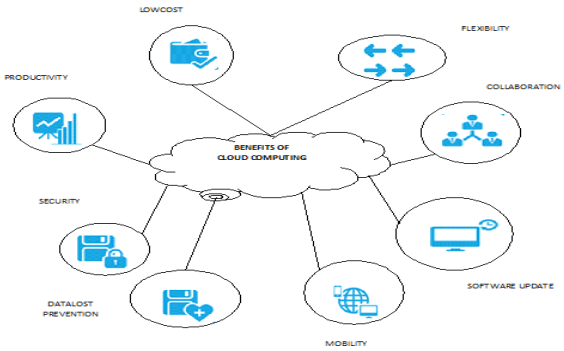

Figure 3: Shows benefits of cloud computing

Computing through Internet reduces cost as organizations need not invest in the computing resources or infrastructure. It is flexible as you just need a PC with Internet connection to get connected to cloud without geographical and time restrictions. Its 365 days availability adds to be flexibility to consumers of cloud based services. With cloud, collaborative computing and collaboration among different parties geographically located is made a reality. This has brought skill sharing and exploiting expertise of people across the globe. With Software as a Service (SaaS), described later in this chapter, cloud computing provides benefits of managing software in a central place and giving access to entire world without the need for installing software at client premises. Moreover, the software updates can be done with ease at one place and render enhanced software to users intuitively. Mobility is the great advantages people and organizations can have with cloud. Wherever you are, you can gain access to cloud and use software, data, platforms and hardware infrastructure. This paves way for seamless access from multiple devices and platforms. The cloud computing technology has provision for storage and data loss prevention with its redundant storage mechanisms that are fail-proof in nature. There is no data loss problem due to this service. Security with different levels is provided by CSPs to safeguard applications and data that are associated with public cloud. With cloud computing in place productivity is more. Often it is very easily realized.

For instance, it is possible to set development environment with cloud in few minutes saving lot of time and investment. Therefore productivity is one of the best known advantages of cloud computing. There is another important benefit of using cloud computing that is the elasticity nature of cloud that scales up to the needs of the world from time to time. This is the single most advantage that let organizations prefer using cloud resources and increases chances of success in business by avoiding major investments and using cloud computing resources. Cloud computing also provides different service models to suit the needs of organizations. Figure 4 illustrates possible basic service models to give better choices to consumer fraternity.

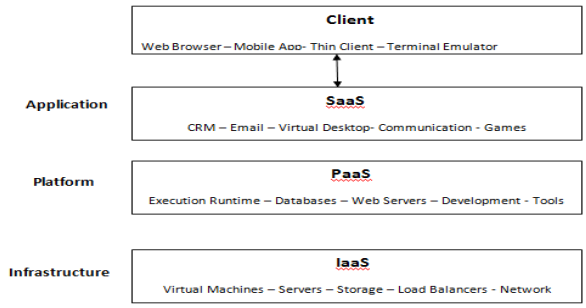

Figure 4: Different service layers of cloud computing

Since cloud is service oriented distributed computing phenomenon, it has three basic kinds of built in services. Other services are also emerging over a period of time. Software as a Service (SaaS) provides different kinds of applications or software to public in pay per use fashion. For instance, Google Cloud renders different every day-use applications in the form of Google Docs. This will help users to use those applications without installing any software in their system. It is wonderful as the users can have mobility and use software and files from anywhere in the world. SaaS has benefits to consumers and also service providers. It is possible to distribute software with ease. Traditional approaches to install update software in client machines are no longer required. It saves operational costs to service providers and service costs to consumers. Platform as a Service (PaaS) is another service layer which is for providing different development platforms. It provides environment required to build various kinds of applications on top of Infrastructure as a Service (IaaS). The IaaS is more useful to the entire world as people and organizations use its services like storage, server machines, Virtual Machines (VMs), networks, security primitives, load balancers and clusters of commodity computers. All these are available under IaaS layer. All these layers can be accessed from users remotely using thin and thick clients. In addition to services aforementioned, cloud computing technology also makes different kinds of deployment models, shown in Figure 5, that cater to the needs of organizational and individual users.

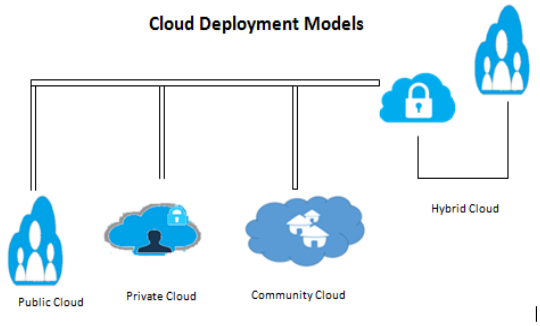

Figure 5: Different deployment models provided by cloud computing

It is essential to meet different needs of organizations. Towards this end, there are four kinds of deployment models. They are known as private cloud, public cloud, community cloud and hybrid cloud. Private cloud is the cloud that is owned by an organization and it is either built or taken as out of the box solution. It runs in the local network and users of organization only gains access to this. On the other hand, public cloud provides access to entire world. Community cloud is similar to private cloud but it is owned by multiple organizations that form a community. Users of those organizations only can access community cloud. Hybrid cloud is the combination of private and public clouds. This will pave way for extended cloud to many organizations.

IV. FRAMEWORKS FOR BIG DATA ANALYTICS

Distributed programming frameworks or distributed computing frameworks that play crucial role in big data processing are covered here. There are many such frameworks. However, in this study, Apache Flink and Apache Spark are considered. They are used with environmental data (big data) for observing their utility in big data analytics. Distributed programming framework provides reusable building blocks to developers and makes their application development and analytics easier. The framework provides pre-defined execution flow, fault tolerance, handling a failed task and a host of such features.

i Apache Flink

Apache Flink is an open source stream processing framework. It is used to perform big data analytics. It executes data streams related applications in data-parallel and pipelined fashion. It can process data as bounded and unbounded streams. Unbounded streams contain starting point but there is no precisely defined end. They need to be processed continuously. It is not feasible to wait till all the data is accumulated. Bounded streams on the other hand can have start and end. Therefore, they can be processed once the complete data is arrived. Apache Flink is very good at processing both bounded and unbounded streams.

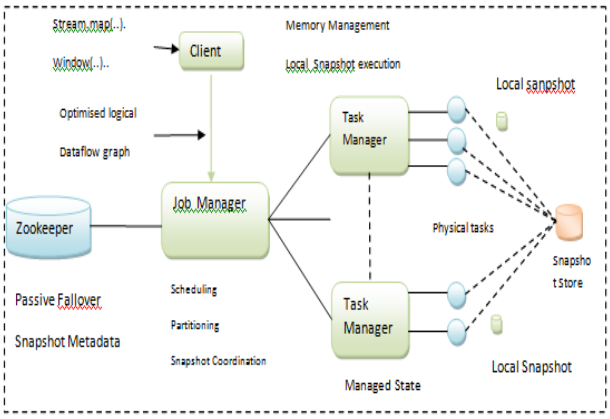

Figure 6: Overview of Apache Flink framework

As presented in Figure 6, Apache Flink has its architecture with pre-defined workflow. It has managed environment that will provide functionalities to applications. The framework supports passive fall over, snapshot Meta data, scheduling, partitioning and snapshot coordination. It has memory management and structured to hold and process data. It can integrate cluster resource managers and thus applications can be deployed anywhere in distributed environments. Moreover applications can run at any scale. It also can leverage in-memory performance. Stateful Flink applications can have access to in-memory data structures

Apache Flink has APIs for stateful event-driven applications, stream or batch data processing and high level data analytics.

ii. Apache Spark

Apache Spark [5] is one of the distributed programming frameworks to handle big data. It is suitable for iterative workloads in distributed environments. It helps in providing in-memory computations. For such workloads, Spark is better than other frameworks like Hadoop. It is so for two reasons. The first reason is that it supports a new type of data structure known as Resilient Distributed Dataset [6]. It plays crucial role in in-memory computations. The second reason is that Spark has rich software architecture that makes distributed programming easier.

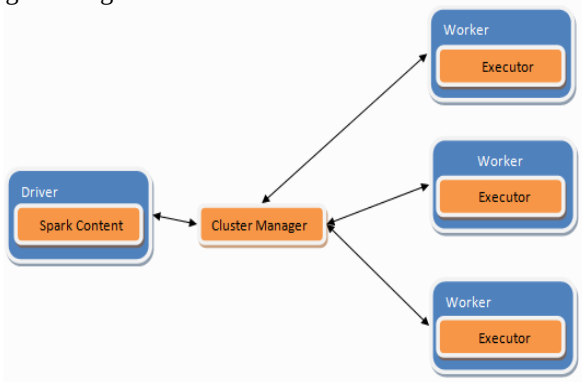

Figure 7: Shows overview of the Spark framework

Cluster manager and worker nodes are involved in processing big data. Immutable collection of data is stored in RDDs. From distributed file systems, data can be loaded into RDDs. With RDDs, refined version of MapReduce programming paradigm is achieved. It provides rich set of user defined functions to have computation pipeline. Lazy computation of RDDs can speed up processing. Each RDD is subjected to sequence of operations. All tasks in Spark are carried out on RDDs constructed. The tasks may have two kinds of relationships known as concurrency and dependency. There are two types of dependencies known as wide and narrow. When two consecutive map() functions work on RDD, it results in narrow dependency. On the other hand a wide dependency causes shuffle and then reduce() function.

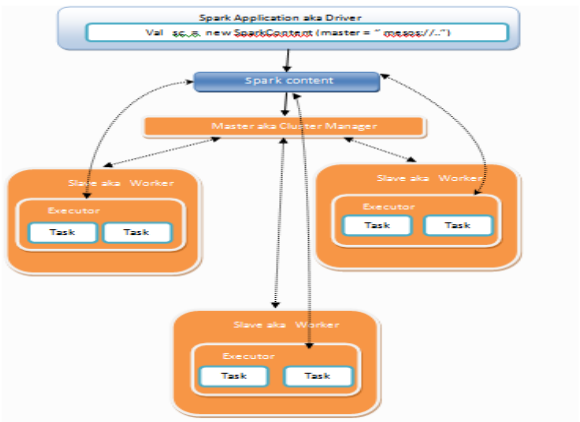

Figure 8: More detailed view of Spark framework

As presented in Figure 8, more details of the framework are visualized. The operations that are to be applied on RDDs are analyzed by Spark at compile time. Once it is done, Spark generates Directed Acyclic Graph (DAG). It tries to keep many narrow dependencies in one stage and wide dependencies into different stages. Preferably, tasks are assigned to worker nodes. This is achieved using delay scheduling technique. It ensures that worker is assigned only when it holds the data to be processed. There are different locality levels that can be considered. They include cluster locality,

rack locality and node locality. Spark will acquire tolerance to non locality gradually.

V. BIG DATA ANALYTICS ON ENVIRONMENTAL ISSUES

There are many environmental issues as studied in [4]. The issues are related to pollution, climate, degradation, waste and resources. Major source of pollution is combustion of fossil fuels [45] Pollution is of many types. It includes air, water, soil, oil and electromagnetic (EM). Air pollution refers to impurities in air and reduction of air quality. Air pollution in cities may be caused by number of factoring including traffic volumes, constructions, industries and vehicles. Oil spills also cause many environmental issues like spoiling coastlines, water bodies and beaches. Water pollution refers to the impurities added to water. Noise pollution on the other hand is the amount of noise in the urban areas due to vehicles and traffic. Climate issues are also there due to increase in sea levels. Air monitoring can help in providing information related to air pollution.

Work	Scope	Objective	Data Processing Method	Data Source
[51]	Campus	Monitoring and real time observation	Web display and databases storage	Potentially deployed ZigBee sensors
[52]	City	Air pollution information access	Spatial data mining	Government database
[53]	City	Air pollution detection	Remotely sensed image fusion	LandSat Thematic Mapper data
[54]	City data	Air pollution level monitoring	Data fusion and neural network	Environmental monitoring network
[55]	City	Air pollution monitoring and control	Data collection and visualization	Sensing data from participants
[56]	City vehicles	Air and fog pollution recognition	Light detection and recognition	Front camera and LIDAR sensor on

Table 1: Shows different air quality monitoring systems found in literature

As shown in Table 1 there are many air quality monitoring systems. They have common or related objectives but the processing method may difference and data sources on which they operate may differ. The process of air pollution detection provided may change from time to time due to technological innovations. In our future work, we proposed a framework that exploits the cloud and big data ecosystem as much as possible.

It is also studied in [4] that big data analytics can help in identifying degradation of forests, taking steps to improve it besides monitoring resources that are renewable and non-renewable. There are many applications of big data such as assessing sustainability of supply chains [2], achieving environmental sustainability [3], for building green applications [4] and realization of smart cities [5]. Air quality monitoring [6], disease prevention [7], air pollution monitoring and control systems [8], Construction waste management [9] and mapping social media and big data [10] are other applications of big data. Big data is also used in material sciences [11], risk assessment applications [12]; predicting life expectancy in the wake of increase in pollution [13], big data based air quality prediction [14] and building environmental applications [15].

With big data it is possible to have renewable energy sources and their utilization effectively to realize health and climate benefits [16]. Smart city applications can be constructed and sustainable urban development is possible by considering big data analytics [17]. Environment protection and monitoring of air pollution is possible with effect analysis of big data [18]. In the smart cities, forecasting of environment can help reduce certain risks. Deep learning technique with big data using CNN-LSTM is explored in [19] for big data analytics. Content analysis is explored in [20] with big data.

Occupational and environment health is explored in [21] with big data analytics. It is understood that big data analytics can help improve quality of services in healthcare industry. Big data analysis can provide intelligence need to development of society as it can improve resource utilization and reduce waste in human activities [22]. Overall quality of life gets increased when big data is used for acquiring business intelligence [23]. Big data analysis can help in understanding air pollution and its effect on public health [24]. Big data is exploited in [25] with Google Street View Cars for high resolution pollution mapping. Air quality maps are produced in order to help in making decisions.

Air quality enhancement and driving restrictions are studied with big data processing [26] and found that big data analytics can provide solutions to complex environmental issues. Big data can also help to have sustainable usage of renewable energy resources effectively [27]. Enforcing policies on environment is made possible with BI acquired with big data analytics [28]. The big data related to odd and even license plate model is used to ascertain the severity of air solution and its reduction [29]. In the age of big data, environmental study can be made and environmental justice can be achieved [30].

In [31] an air quality prediction model is proposed using heterogeneous urban big data. It is built with the help of a model known as spatio-temporal granger causality model. Smart environment monitoring is explored using Internet of Things (IoT) and big data analytics [33]. For environmental sustainability sensor based big data is analyzed in [34] for building smart sustainable cities. Environmental analysis with big data can also help in determining routes to ships in sea as investigated in [35]. Real time risk assessment in many applications like Indian Railways can be done using big data [36]. Big data concept is used in [37] in order to analyze volumes of published data. Pollution analysis and control is investigated with big data support in [38]. Different spatiotemporal approaches are studied in [39] for big data analysis to have environmental justice. Statistical issues with big data analytics are explored in [40].

Big data and mobile crowd sensing are used for air quality estimation in [41]. IoT and virtualization are investigated to have a system for big data processing to analyze environmental data [42]. A pollution forecasting framework is defined in [43] with big data using fuzzy logic and classification technique. Highly polluting firms, their data, and the effect of public health is investigated in [44]. Different departments of government are using big data analytics for making good decisions [46]. Environmental policies and the related health impacts are investigated in [47]. A real time air pollution model based on CNN is proposed in [48] while [49] explores emergent pollutants that threw challenges in some part of the world. Enhancement of air pollution prediction is investigated using big data in [50]. From the literature, it is understood that big data plays vital role and big data analytics is inevitable in the present world for monitoring and improving air quality.

VI. EXPERIMENTAL RESULTS

Big data analytics need distributed programming frameworks for storage and handling of big data. There are many such frameworks available. In this paper, two important frameworks are considered and comparative study of the same is made. The frameworks are known as Apache Spark and Apache Flink. Experiments are made with TeraSort benchmark which works on big data. The observations are made to know the performance difference between the two frameworks in terms of execution time, network usage and throughput.

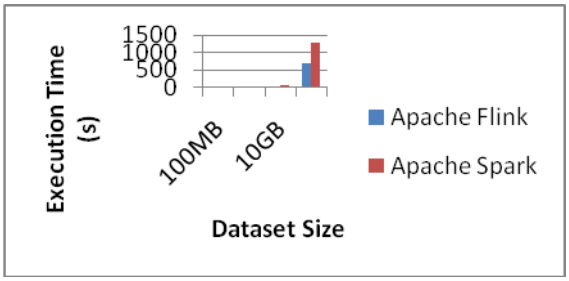

Figure 9: The execution time taken by frameworks with TeraSort benchmark

As presented in Figure 9, it is evident that the execution time required by the two framework to execute TeraSort benchmark for given environment dataset is observed. The size of data is provided in horizontal axis while the vertical axis presents the execution time taken. The results revealed that Apache Flink has taken less time when compared with that of Apache Spark. An important observation is that when data is increased in size, the time taken is also increased linearly.

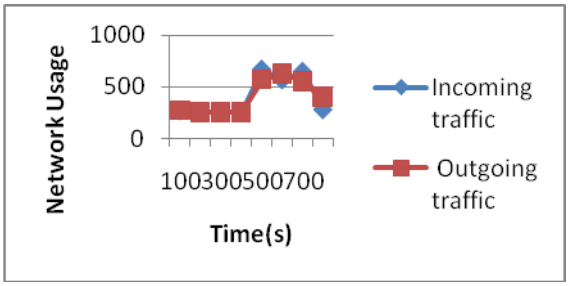

Figure 10: Network usage of Apache Flink with TeraSort benchmark

As shown in Figure 10, the network usage with incoming traffic and outgoing traffic is analyzed when Apache Flink is used for experiments. The results revealed that over a period of time, the network usage is changed for both the traffic flows. The usage is very high when elapsed time reaches 600 seconds. Network usage is again reduced when elapsed time reached 800 seconds.

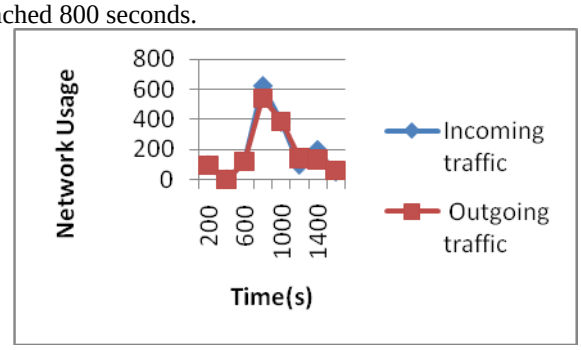

Figure 11: Network usage of Apache Spark with TeraSort benchmark

As shown in Figure 11, the network usage exhibited by Spark is different from that of Apache Flink both incoming and outgoing traffic flows. The network usage increased when elapsed time is increased. After some time, again it is decreased. When elapsed time is 400 through 800, the network usage is increased. Afterwards, from 800 through 1200, it is decreased.

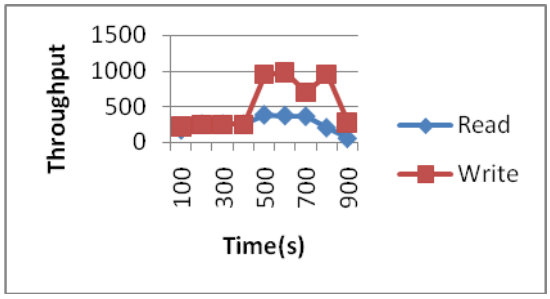

Figure 12: Throughput exhibited by Apache Flink with TeraSort benchmark

As presented in Figure 12, the read and write operations are examined to observe throughput exhibited by the Flink framework over time. As the elapsed time is increased, there is increase in the write operation's throughput and reduction in the read operation's throughput. When elapsed time is 900, the observation is stopped. The read and write operations showed comparatively less throughput at the end of the elapsed time. In the middle that is between the elapsed time 400 to 700, the throughput of both operations increased and then decreased. Overall, the write operation's throughput is more than that of read.

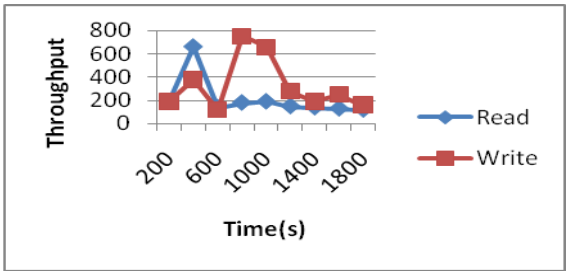

Figure 13: Throughput exhibited by Apache Spark with TeraSort benchmark

As presented in Figure 13, the read and write operations are examined to observe throughput exhibited by the Apache Spark framework over time. The read operation showed higher throughput initially but later on it is reduced. At elapsed time 800 seconds, the write operation throughput is shown highest. Afterwards the write throughput is gradually decreased. The results revealed that the two frameworks are capable of processing big data (environmental data) effectively. However, there are certain differences in the performance as shown in this section.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we explored the big-data eco system and distributed programming frameworks such as Apache Flink and Apache Spark. It throws light into introduction to big data, the need for big data processing and how big data can help adding value to businesses. Especially it covers the utility of big data for adding value to environmental study. It is understood that big data can help in the research of climate changes and environment including aim pollution in terms of prediction and making strategic decisions for well being of society. Our investigation into big data revealed that with the

emergence of cloud computing and many distributed programming frameworks paved way for unprecedented possibilities. With pollution data (big data) we have made an empirical study on the usage of Apache Flink and Apache Spark frameworks. In future we continue this research to develop a framework that can help in measuring and prediction of air pollution by exploiting big data and cloud eco-system. Such framework can be used in Decision Support System (DSS) pertaining Pollution Control Boards (PCBs).

REFERENCES

1. Rubén Casado and Muhammad Younas. (2014). Emerging trends and technologies in big data processing, p1-15.
2. Taliva Badiezadeh, Reza Farzipoor Saen and Tahmoures Samavati. (2017). Assessing sustainability of supply chains by double frontier networkDEA A big data approach. Elsevier, p1-7.
3. M, cen, L, zheng, Z, fisher, R, Liang, X, wang, Y & Huisingh, D 2016, 'How would big data support societal development and environmental sustainability Insights and practices' Journal of Cleaner Production, vol. 142, pp. 489-500.
4. Jinsong Wu, Song Guo, Jie Li and Deze Zeng . (2016). Big Data Meet Green Challenges Big Data Toward Green Applications. IEEE Systems Journal. 10 (3), p1-14.
5. Rob Kitchin. (2014). The real-time city Big data and smart urbanism, p1–14.
6. Yu Zheng, Furui Liu and Hsun-Ping Hsieh. (2013). U-Air When Urban Air Quality Inference Meets Big Data. ACM, p1-9.
7. Meredith A. Barrett, Olivier Humblet, Robert A. Hiatt and Nancy E. Adler. (2013). BIG DATA AND DISEASE PREVENTION, p1-10.
8. Diego Mendez, Alfredo J. Perez, Miguel A. Labrador and Juan Jose Marron. (2011). P-Sense A Participatory Sensing System for Air Pollution Monitoring and Control. IEEE, p1-4.
9. Lu, W, Chen, X, Ho, DCW and Wang. (2016). Analysis of the construction waste management performance in Hong Kong: the public and private sectors compared using big data. Journal of Cleaner Production. 112 (1), p521-531.
10. Ming-Hsiang Tsou. (2015). Research challenges and opportunities in mapping social media and Big Data. Cartography and Geographic Information Science, p1-7.
11. Xiaogang li. (2015). Materials science Share corrosion data. 5 2 7, p1-3.
12. Peter P. Egeghy, Linda S. Sheldon, Kristin K. Isaacs, Halûk Özkaynak, Michael-Rock Goldsmith, John F. Wambaugh, Richard S. Judson and Timothy J. Buckley. (2016). Computational Exposure Science An Emerging Discipline to Support 21st-Century Risk Assessment. Environmental Health Perspectives. 124 (6), p1-6.
13. Avraham EBENSTEIN, Maoyong FAN, Michael Greenstone, Guojun HE, Peng YIN and Maigeng ZHOU. (2012). Growth, Pollution, and Life Expectancy China from 1991-2012, p1-20.
14. Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang and Tianrui Li. (2015). Forecasting Fine-Grained Air Quality Based on Big Data. ACM, p1-10.
15. Carlos Granell, Denis Havlik, Sven Schade, Zoheir Sabeur, Conor Delaney, Jasmin Pielorz, Thomas Usländer, Paolo Mazzetti, Katharina Schleidt, Mike Kobernus and Fuada Havlik . (2016). Future Internet technologies for environmental applications. Elsevier, p1-15.
16. Jonathan J. Buonocore, Patrick Luckow, Gregory Norris, John D. Spengler, Bruce Biewald, Jeremy Fisher and Jonathan I. Levy. (2015). Health and climate benefits of dierent energy-eciency and renewable energy choices. Nature Climate Change, p1-9.
17. Margarita Angelidou, Artemis Psaltoglou and Nicos Komninos. (2017). Enhancing sustainable urban development through smart city applications. Journal of Science and Technology Policy Management, p1-26.
18. Muhammad Imran Qureshi, Amran Md. Rasli, Usama Awan, Jian Ma, Ghulam Ali, Faridullah, Arif Alam, Faiza Sajjad and Khalid Zaman. (2014). Environment and air pollution health services bequeath to grotesque menace. Springer, p1-11.
19. Chiou-Jye Huang ID and Ping-Huan Kuo. (2018). A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities, p1-22.
20. Isti Surjandari, Arian Dhini, Esther Widya Impola Lumbantobing, Anita Titiani Widari and Irfan Prawiradinata. (2015). Big Data Analysis Of Indonesian Scholars' Publications A Research Theme Mapping. International Journal of Technology, p650-658.
21. David M. Stieb, Cécile R. Boot and Michelle C. Turner. (2017). Promise and pitfalls in the application of big data to occupational and environmental health, p1-4.
22. M, cen, L, zheng, Z, fisher, R, Liang, X, wang, Y & Huisingh, D 2016, 'How would big data support societal development and environmental sustainability? Insights and practices' Journal of Cleaner Production, vol. 142, pp. 489-500.
23. Shivam Gupta, Jorge Mateu, Auriol Degbelo and Edzer Pebesma. (2018). Quality of life, big data and the power of statistics. Elsevier, p1-4.
24. Xiaoyu Chen, Shuai Shao, Zhihua Tian, Zhen Xie and Peng Yin. (2016). Impacts of air pollution and its spatial spillover effect on public health based on China's big data sample. Journal of Cleaner Production, p1-36.
25. Joshua S. Apte, Kyle P. Messier, Shahzad Gani, Michael Brauer, Thomas W. Kirchstetter, Melissa M. Lunden, Julian D. Marshall, Christopher J. Portier, Roel C.H. Vermeulen and Steven P. Ha. (2017). High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. American Chemical Society, p1-10.
26. HengjunHuang ,Deyin Fu, Wei Qi . (2017). Effect of driving restrictions on air quality in Lanzhou, China: Analysis integrated with internet data source. Journal of Cleaner Production. 142, p1013-1020.
27. PouyaIfaeia ,AbdolrezaKarbassib , SeungchulLeea , ChangKyooYooa. (2017). A renewable energies-assisted sustainable development plan for Iran using techno-econo-socio-environmental multivariate analysis and big data. Energy Conversion and Management. 153 , p257-277.
28. Robert L. Glicksman, David L. Markell, Claire Monteleoni. (2017). Technological Innovation, Data Analytics, and Environmental Enforcement. Ecology Law Quarterly. 44 (1), p1-49.
29. XiaoyaoXie ,XiaodongTou, Li Zhang . (2016). Effect analysis of air pollution control in Beijing based on an odd-andeven license plate model. Journal of Cleaner Production. 1 (10), p1-10.
30. Alice Mah. (2016). Environmental justice in the age of big data: challenging toxic blind spots of voice, speed, and expertise. Environmental Sociology, p1-13.
31. Julie Yixuan Zhu, ChenxiSun,Victor O.K. Li. (2017). An Extended Spatio-temporal Granger Causality Model for Air Quality Estimation with Heterogeneous Urban Big Data. IEEE Transactions On Journal Name, p1-13.
32. Chiou-Jye Huang and Ping-HuanKuo . (2018). A Deep Cnn-LstmModel for Particulate Matter (PM2.5) Forecasting in Smart Cities. IEEE, p1-22.
33. Riyadh Arridha,SritrustaSukaridhoto,DadetPramadihanto,NobuoFunabiki,. (2016). Classification Extension based on IoT-Big Data Analytic for Smart Environment Monitoring and Analytic in Realtime System. IEEE, p1-16.
34. Simon Elias Bibri. (2017). TheIoT for Smart Sustainable Cities of the Future: An Analytical Framework for Sensor–Based Big Data Applications for Environmental Sustainability. IEEE, p1-47.
35. Kai Wang ,Xinping Yan ,YupengYuan,Xiaoli Jiang ,Gabriel Lodewijks ,Rudy R. Negenborn . (2017). Study on Route Division for Ship Energy Efficiency Optimization Based on Big Environment Data . IEEE, p1-7.
36. Ali Jamshidi,1 Shahrzad Faghih-Roohi,2 Siamak Hajizadeh,1 Alfredo N´unez, ˜ 1,∗ Robert Babuska,2 Rolf Dollevoet,1 Zili Li,1 and Bart De Schutter2. (2017). A Big Data Analysis Approach for Rail Failure Risk Assessment. IEEE. 37 (8), p1-13.
37. Mr. Akhilesh Kumar Varma,Dr. Kunwar Singh. (2017). Scientometric Analysis of Literature on Big Data Research Output in India Based on Scopus Database. International Journal of Library and Information Studies. 7 (4), p1-7.
38. LyudmylaDymytrova. (2010). Epiphytic lichens and bryophytes as indicators of air pollution in Kyiv city (Ukraine). Folia Cryptog. Estonica. 46, p33-44.
39. Yoo Min Park and Mei-Po Kwan. (2017). Multi-Contextual Segregation and Environmental Justice Research: Toward Fine-Scale Spatiotemporal Approaches. International Journal of Environmental Research and Public Health, p1-19.
40. D. R. Cox,Ruth H. Keogh. (2017). Big Data: Some Statistical Issues. IEEE, p1-11.

41. Cheng Feng, Wendong Wang, Ye Tian, XirongQue and Xiangyang Gong. (2017). Estimate Air Quality Based on Mobile Crowd Sensing and Big Data. IEEE, p1-9.
42. Roberto Morabito, Vittorio Cozzolino, Aaron Yi Ding, NicklasBeijar, and JörgOtt. (2018). Consolidate IoT Edge Computing with Lightweight Virtualization. IEEE NETWORK, p1-10.
43. IliasBougoudis1 ,Konstantinos Demertzis2 , Lazaros Iliadis3 , Vardis-Dimitris Anezakis2 , Antonios Papaleonidas2. (2018). FuSSFFra, a fuzzy semi-supervised forecasting framework: the case of the air pollution in Athens. Neural Comput&Appli. 29, p375-388.
44. Jun Cheng, Yuyu Liu. (2018). The effects of public attention on the environmental performance of high-polluting firms: Based on big data from web search in China. Journal of Cleaner Production. 186, p335-341.
45. PallavPurohit. (2018). Small and bad. IEEE. 1, p17-18.
46. Ching-Yen Kuo ,Ren-Hao Pan , Chin-Kan Chan , Chiung-Yi Wu , Dinh-Van Phan and Chien-Lung Chan . (2018). Application of a Time-Stratified Case-Crossover Design to Explore the Effects of Air Pollution and Season on Childhood Asthma Hospitalization in Cities of Differing Urban Patterns: Big Data Analytics . International Journal of Environmental Research and Public Health, p1-15.
47. Julia Kravchenko, H. Kim Lyerly, William Ross. (2018). The Health Impacts of Environmental Policy: The North Carolina Clean Smokestacks Act. IEEE. 79 (5), p1-5.
48. .Duc Le, Sang Kyun Cha. (2018). Real-time Air Pollution prediction model based on Spatiotemporal Big data. IEEE, p1-6.
49. Valeria Dulio, Bert van Bavel, Eva Brorström-Lundén, JoopHarmsen ,JulianeHollender, Martin Schlabach, JaroslavSlobodnik, Kevin Thomas and Jan Koschorreck. (2018). Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. IEEE. 30 (5), p1-13.
50. TalatZaree, Ali Reza Honarvar. (2018). Improvement of air pollution prediction in a smart city and its correlation with weather conditions using metrological big data. Turkish Journal of Electrical Engineering & Computer Sciences. 26, p1302-1313.
51. Y. Ma, S. Yang, Z. Huang, Y. Hou, L. Cui, and D. Yang, "Hierarchicalairqualitymonitoringsystemdesign,"in*Proc.Int.Symp.Integr.Circuits*, Dec. 2014, pp.284–287.
52. [52] V. Ng, S. C. F. Chan, and S. Au, "Web agents for spatial mining on air pollution meteorology," in *Proc. 6th Int. Conf. Comput. Supported Cooperative Work Design*, Jul. 2001, pp.293–298.
53. A. Zia, V. DeBrunner, A. Chinnaswamy, and L. DeBrunner, "Multi-resolution and multi-sensor data fusion for remote sensing in detecting air pollution," in *Proc. IEEE 5th Southwest Symp. Image Anal. Interpretation*, Apr. 2002, pp.9–13.
54. J.M.Barron-Adame,M.G.Cortina-Januchs,andA.Vega-Corona,"Datafusionandneuralnetworkcombinationmethodforairpollutionlevelmonitoring,"in*Proc.IEEEInt.Conf.Ind.Informat.*,June2009,pp.522– 527.
55. D. Mendez, A. J. Perez, M. A. Labrador, and J. J. Marrony, "P-Sense: A participatory sensing system for air pollution monitoring andcontrol," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, Mar. 2011, pp.344–347.
56. P. Sallis, C. Dannheim, C. Icking, and M. Mder, "Air pollution and fogdetectionthroughvehicularsensors,"in*Proc.IEEEAsiaModellingSymp.*, Sep. 2014, pp.181–186.

AUTHORS PROFILE

M. NAVEEN BABU, completed M.Tech in computer science and engineering and Research scholar from KL (Deemed to be) University.

A.V.KRISHNA PRASAD , Research Supervior of KL (Deemed to be) University and professor of MVSR Engineering College.He is CSI (computer society of India) Hyderabad chapter in charge. He published many papers in different international journals.