Comparative Study of Data Analysis Methods in a Fog

S. N. Khandare [*1] *and S. P. Deshpande* [2]

Abstract

Fog computing is distributed computing consisting of Cloud layer, Fog Layer, and IoT devices/ sensors. Fog computing extends cloud based services closer to end devices in order to provide quick response to real time IoT applications. IoT devices with different sensors generate a huge amount of data which are processed at Fog layer to find insight for business, to reduce workload at Cloud, to reduce use of network bandwidth, to respond in time effective manner or reduce latency and response location awareness. For analysing data effectively at insight for business, Fog layer is a need for intelligent computing systems like data mining and data analytics. In a Fog environment the data analytics can be performed at Fog layer or at Cloud Layer and Fog Layer. The Fog nodes in fog layer are limited in computing resources, so the data analytics must be distributed among the fog nodes to work in a distributed way. Machine learning based data analytics model with various Data analytic methods works better in such situations. In this paper we are comparing various data analytics methods used in Fog computing.

Keywords : **Cloud, data analytics, Fog computing**

NOMENCLATURE

RMSE	Root Mean Square Error.
MLP	Multilayer Perceptron.
KNN	K-Nearest Neighbor.
DT	Decision Tree.
NB	Naïve Bayes.
LPM	Local Position Measurement.
GPS	Global Positioning System.
DCDA	Distributed Cooperative Data Analytics.

I. INTRODUCTION

Large amount of data is generated due to the exponential growth in use of smart devices, IoT devices, and sensors. This huge data needs to be processed in a time effective manner. Cloud computing efficiently works to handle the need of computing resources for processing or computing such huge data but it is not feasible to send such huge amounts of data everytime to Cloud for processing. It creates latency, increases workload on Cloud and creates congestion in networks. It is due to the centralized location of the Cloud and large physical distance between end devices and Cloud data centre. Fog computing technology emerged as an effective solution in such a scenario. The Fog computing environment consists of three layers. The bottom layer consists of data generating

Manuscript Received : August 21, 2022 ; Revised : September 8, 2022 ; Accepted : September 10, 2022. Date of Publication : October 5, 2022.

S. N. Khandare[*1], *Assistant Professor,* P. G. Department of Computer Science and Technology, Degree College of Physical Education, Amravati. Email : sandeepkhandare@gmail.com ; ORCID iD : https://orcid.org/0000-0002-0021-3546

S. P. Deshpande[2], *Professor,* P. G. Department of Computer Science and Technology, Degree College of Physical Education, Amravati. Email : shrinivasdeshpande68@gmail.com ; ORCID iD : https://orcid.org/0000-0002-9984-016X

DOI : https://doi.org/10.17010/ijcs/2022/v7/i5/172580

devices such as IoT devices and smart devices with sensors etc. The middle layer is Fog layer which is composed of fog devices connected with each other to extend Cloud services near IoT device/end devices and at the top there is a Cloud layer having huge resources. The Cloud layer performs complex data computations on summarized data received from fog nodes [1][2][3]. An IoT device with a sensor continuously generates data and many Fog applications required data streams to be processed in a time effective manner which required real-time analysis. Fog computing has come up as a solution in such a situation. Fog layer as it consists of the computing devices near the data generator performs most of the analytical operations so as to reduce data sent to the Cloud. It results in reducing workload at Cloud, avoids network congestion and minimizes latency in responding to the end devices. The Fog Data Analytics (FDA) is combination of Cloud Computing, Fog Computing, and analytics on data at the Cloud layer or at Fog layer. All three layers of the Fog computing environment are involved in Fog data analytics. The end user layer generates data and forwards it to the Fog Layer. Fog Layer compromises Fog Nodes (FNs) connected to each other, handles priority-based requests and performs data analytics individually or in combination. Fog nodes in the fog layer are limited in computing resources, so the data analytic task must be distributed among the fog nodes to work in a distributed way. The data analytics models with Machine Learning work better in such situations. Methods like Feature Selection, Supervised Learning, Distributed Decision Trees, Clustering Methods, and Parallel and Distributed Association Rule Mining are used for data analytics in fog environments [4].

II. METHODS OF DATA ANALYTICS

Data analytics at Fog layer is very important nowadays when use of sensor and IoT devices is increasing which creates large data that needs to be processed in a time effective way. There are many approaches suggested by researchers to make data analytics at fog effective. Take a look at some approaches.

The optimization problem for distributing Deep Neural Network models across different compute hierarchies is used to minimize communication between compute hierarchies [5]. Deep learning is a promising technology to analyse huge amounts of data generated by IoT. Distributing deep learning in resource constraint Fog nodes is a complex task. The model optimally distributes the layers of deep learning models or queries of complex event processing engines among edge, fog, and cloud devices with the help of the particle swarm optimization algorithm. Due to the optimal distribution of computing tasks among different resources, the data is sent to the cloud to reduce and minimize inter partition communication.

The Machine Learning based anomaly detection model [6] is developed for generating insights on smart meter sensor data in a hierarchically distributed fog computing architecture. The intelligence and learning for anomaly detection is distributed in layers of the distributed Fog Computing architecture. The Core Layer performs Model Training and Dynamic Threshold calculation using prediction methods such as Linear Regression, Support Vector Regression, Random Forest Regression, Gradient Boosting Regression, and the real-time anomaly detection is performed at the Edge layer by using Real-time Anomaly detection algorithm. Four models are used for predictions for 16, 24, and 48 hours. The score of RMSE is used for prediction. The Support Vector Regression performs better for predicting 16 hours. The Random Forest Regression and Linear Regression was found better for 24 hours and 48 hours respectively.

The data analytical model based on Fog is used with decomposition of multivariate linear regression using the statistical query model and summation form [7]. This analytics model runs in a distributed manner in the Fog-enabled IoT deployments to make data analytics effective. Use of this model in real-world dataset is evaluated using a fog computing test bed. The result shows it avoids sending raw data to the Cloud and offers balanced computation in the infrastructure. When this model is compared with cloud centric approach, results show an 80% reduction in the amount of data transferred to the cloud with a 98% drop in the time taken for receiving the final result. The results of analytics by using these models are equally effective as the traditional cloud-centric approach.

A distributed analytics framework uses Fog for processing IoT data in a time effective manner [8]. It minimizes latency by processing/analysing data at Fog nodes. A distributed data analytics algorithm for Fog enabled IoT systems to analyze data in a distributed manner and a privacy preserving secure protocol for privacy preservation of data holders are used to reduce

latency. The algorithm and protocol are tested on three different cases. In case of seismic imaging, the algorithm shows an upto one order of magnitude acceleration in minimizing the objective value. In case of diabetes progression prediction problem, the method is upto two times faster in terms of reaching a reasonable training MSE error and in case of Enron spam email classification task it achieves above 0.95 accuracy in 1s wall-clock time with 5 ms communication latency and 10 mbps bandwidth network configuration.

The platform consisting of TensorFlow, Docker, and Kubernetes is implemented [9] to support complex data analytics techniques like Deep Learning which efficiently analyse maximum data at the Fog layer and reduce the amount of data to be sent to data centers for analysis. TensorFlow, Docker, and Kubernetes are the technologies used to implement Fog computing platforms. Result of evaluation of the system shows (i) Use of distributed analytics improves performance up to 54.1%, (ii) Equal complexity of operators perform best in distributed experiments, and (iii) Use of Docker containers reduce overhead by 5% .

A multi-tier Fog computing model [10] is used for smart city applications for large scale data analytics. This model uses dedicated Fog nodes in tier 1 and ad-hoc Fog nodes in tier 2. Each Fog is made up of a cluster of computing devices (Master-Worker node). The users of IoT applications request for analytics services to Fog masters. The Fog masters node analyse service request and select analytics algorithm and computing engine. It assesses service requirements and then accordingly schedules Fog members nodes and computing resources for working collectively with distributed computing engines to provide advanced analytics. Logistic Regression (LR) and Support Vector Machine (SVM) are used in this model to work efficiently.

The layered architecture with distributed Fog computing performs data processing and analysis for air quality monitoring [11]. The model of data analytics performs data analytics, and data filtration near the IoT devices/end devices. Clustering methods like K-means and other statistical methods are used to analyse data at the Fog layer to reduce its amount which is to be sent to the Cloud. The clusters can be used to train different models like MLP, SVM, KNN, DT, and NB for supervised classification. The dataset is split into a 70:30 ratio for training and test set formation. The result of data analysis and processing shows that the SVM performed best with an accuracy of 99% on the test set, MLP, KNN, and NB show satisfactory performance. The DT with an accuracy of 69% performed the worst.

The sports analytics system with distributed Fog computing and IoT processes sport data effectively [12]. The use of sensors and IoT bring revolution in the sports industry. The sports data analysis is performed by using image processing, video analytics, and many other techniques. The sports data analytics need to be performed time effectively. It is found that Fog computing can be the best solution for real-time distributed processing of the data generated by devices used in sports. Video streams, Motion Analysis, Tracking using LPM (Radio Signals), and GPS analysis is performed in sports domain.

An analytics everywhere framework is used for analysing IoT data streams to generate location intelligence which is required by many IoT applications in smart cities to enhance performance of data analytics and to reduce operational cost [13]. This architecture is based on an integrated fabric of compute nodes that are designed to perform many analytical tasks which are triggered by IoT data streams. The framework consists of three elements: Resource Capability, Analytical Capability, and a Data Life-cycle. The edge-fog-cloud supports streaming analytics to maximize the potential insights from IoT data streams and is useful for integrating IoT devices.

A Fog computing middleware is used to support scalable and flexible distributed cooperative data analytics (DCDA) at the edge of a network [14]. Processing, computing, analysing, and sharing information can be carried out at the edge for providing high-level situation awareness. This middleware allows the analysis of the most time-sensitive data at the network edge. The DCDA minimizes latency, conserves network bandwidth, and allows collecting and processing data across a large geographic area. When the framework is evaluated with the help of experiment, it is found that the suggested middleware is scalable, flexible, and energy efficient for real-time distributed applications.

A data processing model with Machine Learning is used for data processing in the Fog environment in an efficient way [15]. In this model, a distributed Machine Learning algorithm is used in classification of streaming data generated by a distributed source. A distributed online multitask learning algorithm is used at the Fog layer to process data. Jointly learning multiple tasks is

more efficient than working in a stand alone mode, it is proved by experiment.

The Fog computing platform is designed and implemented to perform IoT data analytics in a distributed way [16]. The new algorithm is designed to deploy analytical tasks among devices. The new platform has features such as graph processing, the programming model, lightweight virtualization for resource provisioning, and a new algorithm. Evaluation of design framework is done with the help of simulation and result shows that the new algorithm i.e. the SSE algorithm performs better than the ODP algorithm by upto 89.4% and it reduces CPU, RAM and network use by 18.4%, 12.7%, and 898.3%, respectively. SSE is a proposed algorithm referred as SSE algorithm, after the initials of the three intuitions, (i) Scarcest resource first (ii) Shortest path first (iii) Early feature extraction.

A Fog-Engine model [17] provides a data analytics facility near IoT devices so that the quantity of data can be reduced before offloading data to the Cloud. The FOG-engine is composed of three units : (1) An analytics and storage unit responsible for preprocessing and data analytic and storage; (2) Networking and communication unit responsible for the communication (3) An orchestrating unit handling cluster formation and data distribution across a cluster of Fog engines. Fog engine provides lower latency, higher throughput, and less usage of network bandwidth.

The data stream mining algorithms are used to improve data analytics processes at Fog computing. The traditional data mining and data stream mining algorithms are compared in Fog environments [18]. A simulation experiment is used to compare two classification algorithms namely, decision tree algorithm (C4.5) and data mining decision tree algorithm, that is, Hoeffding Tree (HT). The classification rules in C4.5 are applied in the Cloud platform resulting in high accuracy. In a Fog computing environment a large amount of data is generated by IoT/end devices and are sent in the form of stream for computation. This data stream is processed with the help of stream mining model to handle incremental learning where a portion of the data stream is analysed at a time. Each time new data is seen in Fog computing which is processed in a time effective way with accuracy. To improve the performance of Hoeffding Tree, the FS-Harmony search is used to provide accuracy.

An analytical framework is used in which analytics is performed at edge devices by using AI. Analysing data in a distributed way at Fog using AI is another efficient way which can be compared with a centralized system. To perform analytics efficiently in a distributed way there is a need for optimal operating point. The analytical model used provides expression to find and design the optimal configuration of a decentralised Machine Learning system [19]. At this level of optimization the computation resources and communication is reduced and it also improves accuracy.

The Smart-Fog architecture is implemented using a Fog architecture based on unsupervised Machine Learning for big data analysis [20]. The prototype is developed using Intel Edison and Raspberry Pi which is tested on physiological data of patients with Parkinson's disease (PD). The k-means clustering analysis is used by using Python programming language in Fog environment. The Smart-Fog architecture can work efficiently in health problems such as speech disorders and real time clinical speech processing.

The study of data analytics methods in Fog computing shows that decentralized data analytics with AI works promisingly in the IoT environment to process/analyze huge data in real-time and in reducing load of Cloud and network as well.

III. CONCLUSION

The paper contains the comparative study of data analytics methods used in Fog environments. Fog computing is distributed in nature and consists of the fog nodes which are limited in resource. In such a situation, analysing huge data generated by IoT devices or sensors in Fog environments is a complex task. Distributed Data Analytics models based on Fog with Machine Learning work efficiently in such situations. Various Machine Learning methods are used in Fog computing environments for handling different analytical requirements, for example, Deep Learning methods are used to identify significant features to reduce dimensionality of data. Distributed and parallel methods of learning such as multi hyper plane model, Machine classification Model, Divide and Conquer SVM, Neural Network classifiers ,MLP, KNN, DT, and NB are used for classification of data. Decision trees and decision forest are used to support decisions to fulfill computational requirements in a distributed environment. Parallel and Distributed Association Rule Mining is used for handling geographical spread, variety, volume, and velocity of

data. The Parallel and incremental clustering methods are used to handle processing of huge volumes of data.

AUTHORS' CONTRIBUTION

The present paper was developed under the guidance of Dr. S. P. Deshpande who continuously motivated, provided knowledge, and expertise to build and resolve issues. Sandeep Khandare started with the literature survey of various papers, defined the scope, and understood the requirements. He drafted the paper on the study and was also responsible for documentation.

CONFLICT OF INTEREST

The authors certify that they have no affiliation with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in the manuscript.

FUNDING ACKNOWLEDGEMENT

The authors did not receive any financial support for the research, authorship, and/or for the publication of the article.

REFERENCES

[1] M. Antonini, M. Vecchio, and F. Antonelli, “Fog computing architectures: A reference for practitioners,” *IEEE Internet of Things Mag.*, vol. 2, no. 3, pp. 19–25, 2019, doi: 10.1109/IOTM.0001.1900029.

[2] R. Mahmud, R. Kotagiri, and R. Buyya, “Fog computing: A taxonomy, survey and future directions,” In Di Martino, B., Li, KC., Yang, L., Esposito, A. (Eds.) Internet of Everything. Springer, Singapore. doi: 10.1007/978-981-10-5861-5_5.

[3] M. Mukherjee, L. Shu and D. Wang, "Survey of Fog computing: Fundamental, network applications, and research challenges," *IEEE Commun. Surveys Tut.*, vol. 20, no. 3, pp. 1826–1857, 2018, doi: 10.1109/COMST.2018.2814571.

[4] C. S. R. Prabhu, T. Jan, M. Prasad, and V. Varadarajan, “Fog Analytics - A Survey,” *Malaysian J. Comput. Sci.*, pp. 140–151, 2020, doi: 10.22452/mjcs.sp2020no1.10.

[5] M. K. Pandit, R. Naaz and M. A. Chishti, "Distributed IoT Analytics across Edge, Fog and Cloud," in *2018 4th Int. Conf. Res. Comput. Intell. Commun. Netw.*, 2018, pp. 27–32, doi: 10.1109/ICRCICN.2018.8718738.

[6] R. Jaiswal, A. Chakravorty and C. Rong, "Distributed Fog Computing Architecture for real-time anomaly detection in Smart Meter Data," in *2020 IEEE 6th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, 2020, pp. 1–8, doi: 10.1109/BigDataService49289.2020.00009.

[7] M. Taneja, N. Jalodia and A. Davy, "Distributed decomposed data analytics in fog enabled IoT deployments," *IEEE Access*, vol. 7, pp. 40969–40981, 2019, doi: 10.1109/ACCESS.2019.2907808.

[8] L. Zhao, “Privacy-preserving distributed analytics in Fog-enabled IoT systems sensors,” *Sensors 2020*, vol. 20, no. 21, 6153, 2020, doi: 10.3390/s20216153.

[9] P. -H. Tsai, H. -J. Hong, A. -C. Cheng, and C. -H. Hsu, "Distributed analytics in fog computing platforms using tensorflow and kubernetes," in *2017 19th Asia-Pacific Netw. Operations Manage. Symp.*, 2017, pp. 145–150, doi: 10.1109/APNOMS.2017.8094194.

[10] J. He, J. Wei, K. Chen, Z. Tang, Y. Zhou and Y. Zhang, "Multitier fog computing with large-scale IoT data analytics for smart cities," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 677–686, 2018, doi: 10.1109/JIOT.2017.2724845.

[11] M. Ahmed, R. Mumtaz, S. M. H. Zaidi, M. Hafeez, S. A. R. Zaidi, and M. Ahmad, “Distributed fog computing for Internet of Things (IoT) based ambient data processing and analysis,” *Electronics 2020*, vol. 9, no. 11, 1756, doi: 10.3390/electronics9111756.

[12] D. Jha, A. Rauniyar, H. D. Johansen, D. Johansen, M. A. Riegler, P. Halvorsen, and U. Bagchi, "Video analytics in elite soccer: A distributed computing perspective," presented at the *2022 IEEE 12th Sensor Array Multichannel Signal Process. Workshop (SAM),* 2022, pp. 221–225, doi: 10.1109/SAM53842.2022.9827827.

[13] H. Cao and M. Wachowicz, “An edge-fog-cloud architecture of streaming analytics for internet of things applications,” *Sensors 2019*, vol. 19, no. 16, 3594, 2019, doi: 10.3390/s19163594.

[14] J. Clemente, M. Valero, J. Mohammadpour, X. Li, and W. Song, "Fog computing middleware for distributed cooperative data analytics," *2017 IEEE Fog World Congr.*, 2017, pp. 1–6, doi: 10.1109/FWC.2017.8368520.

[15] G. Li, P. Zhao, X. Lu, J. Liu, and Y. Shen, "Data analytics for Fog computing by distributed online learning with asynchronous update," in ICC 2019 - 2019 *IEEE Int. Conf. Commun.*, 2019, pp. 1–6, doi: 10.1109/ICC.2019.8761303.

[16] H. -J. Hong, P. -H. Tsai, A. -C. Cheng, M. Y. S. Uddin, N. Venkatasubramanian, and C. -H. Hsu, "Supporting Internet of Things analytics in a Fog computing platform," in *2017 IEEE Int. Conf. Cloud Comput. Technol. Sci.* (CloudCom), 2017, pp. 138–145, doi: 10.1109/CloudCom.2017.45.

[17] F. Mehdipour, B. Javadi, and A. Mahanti, "FOG-Engine: Towards Big Data Analytics in the Fog," in *2016 IEEE 14th Int. Conf. Dependable, Autonomic Secure Comput.,* in *14th Int. Conf. Pervasive Intell. Comput., 2nd Int. Conf. Big Data Intell. Compu. Cyber Sci. Technol. Congr.* (DASC/PiCom/DataCom/CyberSciTech), 2016, pp. 640–646, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.116.

[18] B. B. Ma, S. Fong, and R. Millham, "Data stream mining in fog computing environment with feature selection using ensemble of swarm search algorithms," in *2018 Conf. Inf. Commun. Technol. Soc.*, 2018, pp. 1–6, doi: 10.1109/ICTAS.2018.8368770.

[19] L. Valerio, A. Passarella, and M. Conti, "Optimising cost vs accuracy of decentralised analytics in Fog computing environments," in *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 1986–2002, 1 July-Aug. 2022, doi: 10.1109/TNSE.2021.3101986.

[20] D. Borthakur, H. Dubey, N. Constant, L. Mahler, and K. Mankodiya, "Smart fog: Fog computing framework for unsupervised clustering analytics in wearable Internet of Things," in *2017 IEEE Global Conf. Signal Inf. Process.* (GlobalSIP), 2017, pp. 472–476, doi: 10.1109/GlobalSIP.2017.8308687.

About the Authors

Sandeep N. Khandare received M.C.A. Degree in Engineering and Technology from Sant Gadge Baba Amravati University. At present he is working as Assistant Professor with P. G. Department of Computer Science and Technology, Degree College of Physical Education, Hanuman Vyayam, Prasarak Mandal, Amravati, Maharashtra, India.

S. P. Deshpande is working as Professor (M.C.A.) at P. G. Department of Computer Science and Technology, Degree College of Physical Education, Hanuman Vyayam, Prasarak Mandal, Amravati, Maharashtra, India. He did M.C.A. and M.Sc.(Physics) from Sant Gadge Baba Amravati University and received Ph.D. in Computer Science and Engineering from the same university. His areas of interest include Software Engineering, Database, Data Warehousing, Data Mining, and Data Analytics.