

The Big Data for WSN Nodes: Leveraging Scalable Architecture

Deepti Shinghal¹, Kshitij Shinghal², Amit Saxena^{2*}, Shuchita Saxena², and Rajul Misra³

¹Regular Member, International Association for Engineering & Technology, Singapore.

²Dept. of E&C Engg., Moradabad Institute of Technology, Moradabad, U.P., India

³Dept. of Electrical Engg., Moradabad Institute of Technology, Moradabad, U.P., India

Abstract. Certain applications requires a scalable cost effective storage and execution system with facility to store data and have feature to analyze data to its finest granularity level in future. This increase the quality and accuracy of result analysis. Wireless sensor Network (WSN) nodes deployed for certain data intensive applications such as surveillance, war zone monitoring etc. generates a massive amount of raw data. There is an essential requirement of storing this data in its native format for analytics purpose in anticipation of future requirements. In present work, a data lake implemented on Amazon AWS is presented for storage of data in original version for future reference. Data Lake implementation service is utilized for storing the data generated in big volumes, high speed and in variety. The data in Data Lake is stored in three zones i.e. raw, reformed and curated. This paper proposes an efficient method of storing structured, unstructured and semi-structured, data in to Data Lake for future retrieval and analytics purpose. The results are comprehensively presented highlighting the advantages of using Data Lake in place of data warehouses.

Keywords. — Big data, Data Lake, Scalable Architecture, WSN nodes

1 Introduction

Data Lakes are becoming popular and appealing choice for data storage. This data is stored in its original version and is generally exceeds in volume and complexity. Data lakes are having inherent advantages and appealing to be employed for storage of data that is generating automatically from digital sensor based IoT devices like Wireless Sensor Network (WSN) nodes. Wireless Sensor Networks (WSNs) are specially designed ubiquitous sensors, which are deployed in a typical environment to monitor or log the changes in certain physical parameters. The aim of WSN network is to generally collect the data and forward it to a central base station. Figure 1 shows the general architecture of a WSN node.

* Corresponding author: amitssaksena@gmail.com

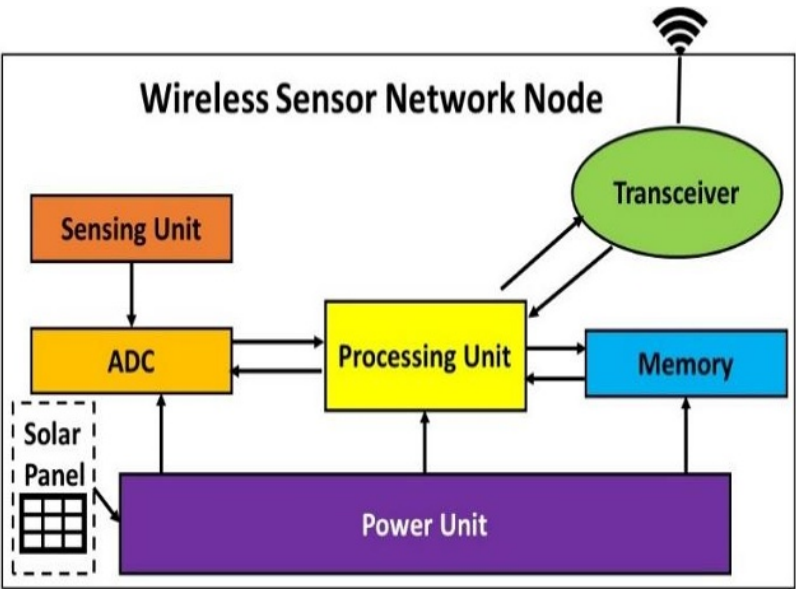


Fig. 1. General Architecture of a WSN node

Modern day WSN nodes are bidirectional in the way that they are capable of transmitting data to a central base station as well as of receiving the data/command to control sensor activity.

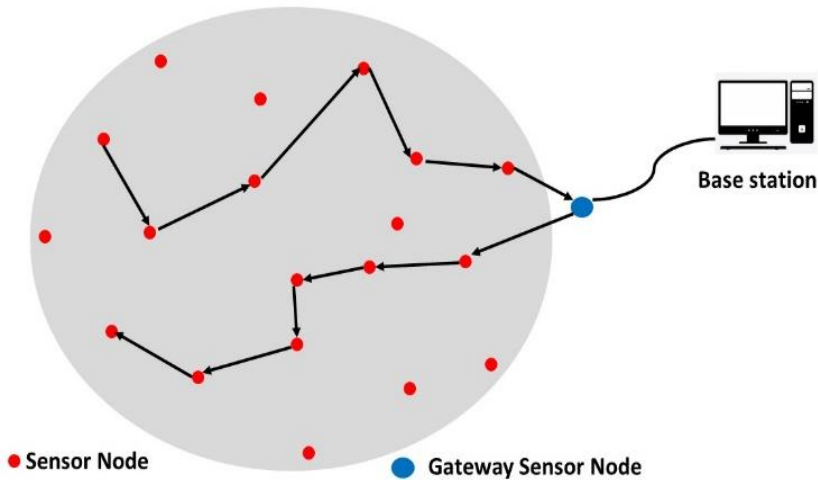


Fig. 2. Modern Day WSN node with bidirectional communication capability

In general, such networks were developed for special applications such as warzone surveillance, industrial hazard monitoring Industrial machine monitoring etc. Figure-3 depicts some such applications of WSN node.

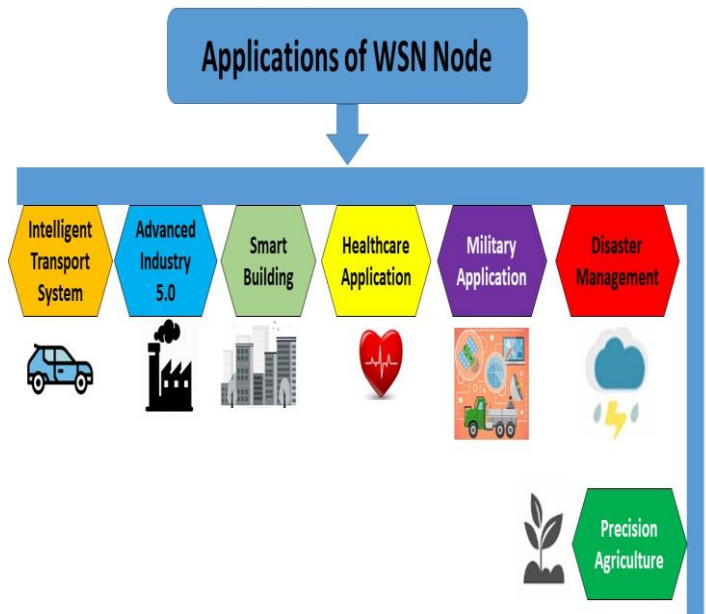


Fig. 3. Typical Applications of WSN nodes that require bidirectional data communication and generates huge amount of data.

The nature of surveillance applications and certain monitoring applications is such that is generates data in huge amount, which is increasing explosively. Generally in WSN networks to save the battery and increase the node lifetime the collected data is aggregated or processed or it is many times event triggered to save the battery life and thereby extend the WSN node life. However, there exist several applications such as war zone monitoring, terrorist surveillance, which requires data for analytics. The surveillance data acquired is collection of data, which may be required for analytics purpose at any reference point in future. Any trivial detail may be of significance. This large collection of data i.e. in structured form, unstructured form and semi-structured form is stored in Data Lake. Figure-4 shout Different types of Data collected in WSN Data Lake.

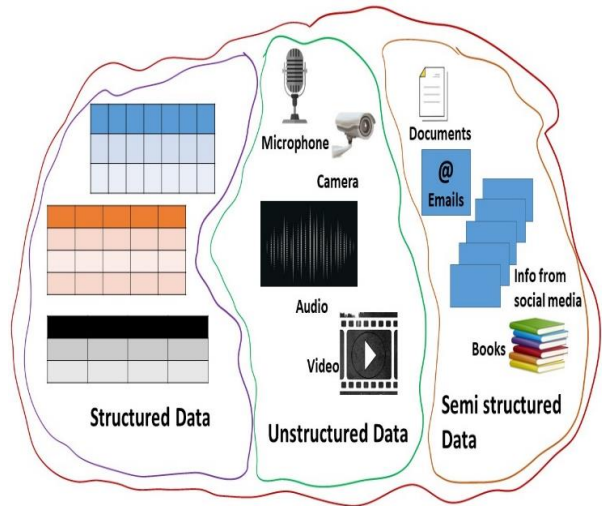


Fig. 4. Various types of Data collected in WSN Data Lake

The figure 4 depicts data gathered from various sources. For example, surveillance audio and video data collected from microphone and camera is grouped as unstructured data. Data gathered from various social media applications, sites etc. for the premises under surveillance such as email documents, chats etc. falls in the group of semi-structured data. Whereas data collected from sensors processed and aggregated by the base node is stored as structured data. In particular example of site surveillance every bit of data, even raw, structured or unstructured may be useful and can be required in future for data analytics purpose. In the WSN data lake, it become very important to have all the three types of data sitting side by side. This is needed because in large number of applications for carrying out the analytics to a greater depth. The requirement is of Data Lake with more than one form of data. For example in recent pandemic outbreak only structured data was not sufficient but unstructured data such as Images (JPEG, PNG etc.) audio (MP3, amr etc.) and videos (MP4, MKV, etc) and semi structured data such as social media posts, email blog posts also becomes very important besides using structured data for analytics.

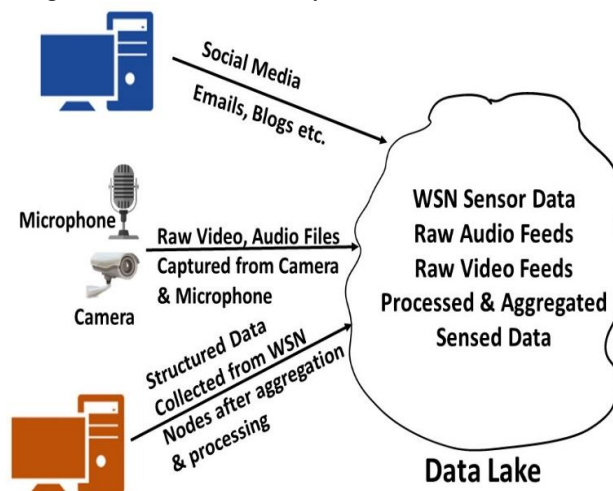


Fig. 5. Sources feeding data in to Data Lake

Figure-5 depicts the various sources of structured, unstructured and semi-structured data, which feeds the data lake.

In this paper, architecture for AWS Data Lake for WSN nodes is discussed. Rest of the paper is organized as follows section 2 depicts literature review followed by Section 3 describing proposed methodology of developing the data lake. Section 4 comprises of results and discussion and finally section-5 covers conclusion and future research direction followed by the references used during the work.

2 Related Work

E. Zagan et al. presents the literature survey describing the different data storage technology, Data Lake architectures along with their advantages and drawbacks [1]. Abdallah Saleh Ali Shatat et al. (2022) have used two models for flood disaster detection. The big data was acquired and processed for flood detection. Further, the Adaptive Billiards-Inspired Optimization (A-BIO) in union with Optimized Ensemble-learning-based detection (OED) was used for optimizing and reducing the complexity of the proposed model. The performance of the detection model was analysed and it was claimed that the model has high accuracy in detecting the flood and hence avoided the huge impacts of flood disasters [2]. R.

Chaudhry et al. proposed a model having Wireless Sensor Network, which provides scalability, configurability and controllability through a software. The system helps in optimizing energy and in controlling delay performance [3]. A. M. Olawoyin et al. proposes a model for collecting the large amount of data from different sources based on the concepts of machine learning. They also reported a baseline metadata for managing huge structured, unstructured and semi-structured data [4]. S. Bagwari et al. suggested a review report on a ground monitoring systems available using Wireless Sensor Nodes in areas where landslides are very frequent. The techniques are classified based on their performance, efficiency and reliability [5]. H. Fang et al. provides a review report of the data lake technology and their advantages in the field of big data [6]. I. Ali et al. conducted a survey and discussed the similarities and differences between the various data collection techniques used in IoT, WSN's and sensor clouds [7]. S. Ramchand et al. provides the literature survey for the data lake concepts and its implementation. The architectures are also discussed in the paper for implementing the concept of data lake [8]. Fei Wang et al. (2022) have reviewed recent advancements in big data analytics and machine learning techniques that could be helpful in IoT applications. The review was carried out in different domains such as platform, frameworks and applications etc. The main points derived from the review was that the big data analytics can be a useful tool for improving the performance of IoT applications and related challenges. The issues like security and latency exists in storing big data on cloud [9]. S. A. Yadav et al. compared the different algorithms using Machine Language for the detection of faults in WSNs. The functions of wireless sensor network changes with the change in environmental conditions like temperature, humidity, speed and wind etc. a comparative study of the effect of all these factors on the WSN are presented in the paper [10]. R. Wrembel et al. discussed the challenges and issues across in implementing the data lake technology for both structured and unstructured data [11]. N. Tripathi et al. proposed the comparative analysis of the different techniques of how to effectively use the available frequency spectrum by using WSN and different algorithms. The paper compares the performance and efficiency of the various methodologies available for routing and communication [12]. C. Liu et al. proposes the virtualization model based analysis system for the collection of big data and thus reducing the cost of data handling [13]. Anne Laurent et al. provides the comparison analysis between the data lake technology and other existing technologies for the collection of big data. The paper also focus on the implementation of the data lake for information system [14]. Houssem Chihoub et al. proposes the architecture of the data lake system for query handling and exploring data in a big data [15]. O. R. Ahutu et al. proposed a technique by using a MAC routing protocol of improving the performance of the Wireless Sensor Network taking energy considerations into account [16]. V. Sadhu et al. proposed a model to meet out all the challenges of Wireless sensor Networks as if scalability, accuracy, cost etc. the model uses the characteristics of MOSFET in realizing the sensors using analog joint source channel coding [17]. M. F. Khan et al. proposes a survey discussing the existing techniques of integrating the multiple wireless sensor network through cloud to improve the efficiency of these sensor networks [18]. R. K. Dwivedi et al. presents a comparative study of the technologies available of connecting the sensor networks with the cloud computing to run many applications simultaneously [19]. D. Tracey et al. propose a scalable architecture for the devices connected through Internet of Things using peer-to-peer connectivity approach [20]. D. Tracey et al. discussed the different techniques through which sensors are able to communicate with each other. The paper describes the algorithm and architecture for transferring data through cloud computing and wireless sensor nodes [21]. G. M. Dias et al. discuss the method of integrating Wireless Sensor Networks using cloud computing and IoT. It increases efficiency of the system as it responds only when the environment changes [22].

3 Proposed Framework

The Literature review emphasizes on the problem data lake and storing the data is a data lake. The data lake improves the performance of analytics, increases reliability, makes the task of making deep analytics very simple specially the tasks involving big data in large volumes and it improves cross-referencing of data directly at the source application its self.

The presented WSN data lake is divided in to four zones for ease of data access resulting in orderly, simple and easy analytics as depicted in figure 6.

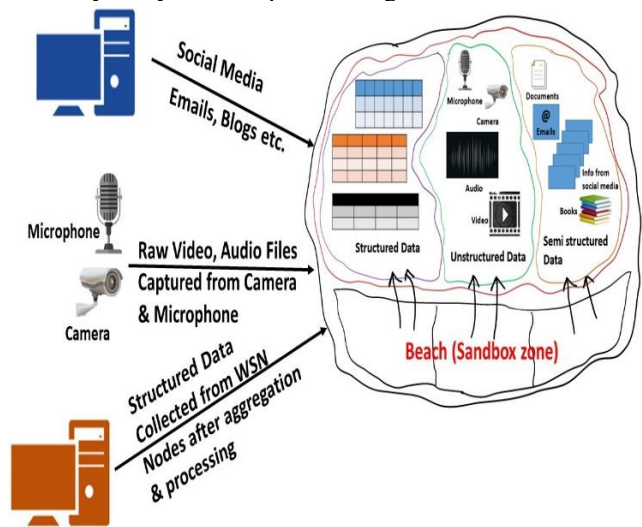


Fig. 6. WSN Data Lake

The four zones of WSN data lake are raw data zone which holds the data when it first enters the lake. Here the data is in its raw form. No transformation, error correction etc. is performed on data. In this zone, the major objective is to import data at big volumes at high speed to get data in to lake as quickly as possible. The task of retiring the data or error correction is left for later stages. Second zone is refined data zone here the data is stored after it is refined, made error free but is still in its original form. Not all data from raw data zone may be cleansed and stored in refined data zone. Only the data that may be required by the analytics for particular application may be cleansed and moved to refined data zone. The final zone is curated data zone here the cleared, error free data is stored in groups and structured form as per the requirements of high value analytics. The data zones are linked together by the beach (sandbox area). The beach zone provides easy access to data in all zones. The data from three zones can be loaded for short term or experimental work for performing analytics. The data lake isolates beach zone from the data pipeline to zones, thereby making it available for experimentation without interfering the data organization or primary analytical work being formed three data lake zones.

3.1 Amazon AWS data lake

AWS provides a platform for building data lake, which is secure, scalable and comprehensive. It supports to build data lake in cloud with facility to analyse all data coming from WSN nodes with a variety of analytical approaches. Figure-7 shows the simplified WSN Data Lake implemented with AWS lake formation.

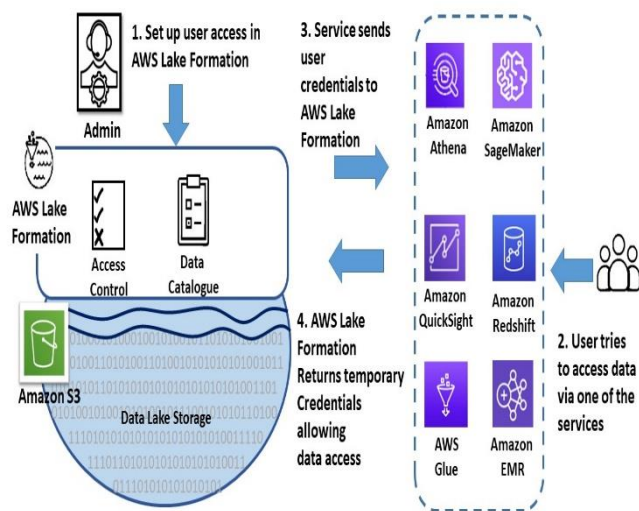


Fig. 7. Simplified WSN Data Lake implemented with AWS lake formation

3.2 Data Lake Storage Model

Figure 8 gives data lake storage model. It shows the data generation from WSN nodes which is passed to Amazon AWS Data Lake and used for analysis purpose by analyst using dashboards, reports at various levels of granularity and complexity. The data thus obtained is analyzed and the parameters like data aggregation, data analysis accessibility, error free report generation, system overhead and processing and administrative overhead were analyzed. The comparison of these features were done with data warehouse technique.

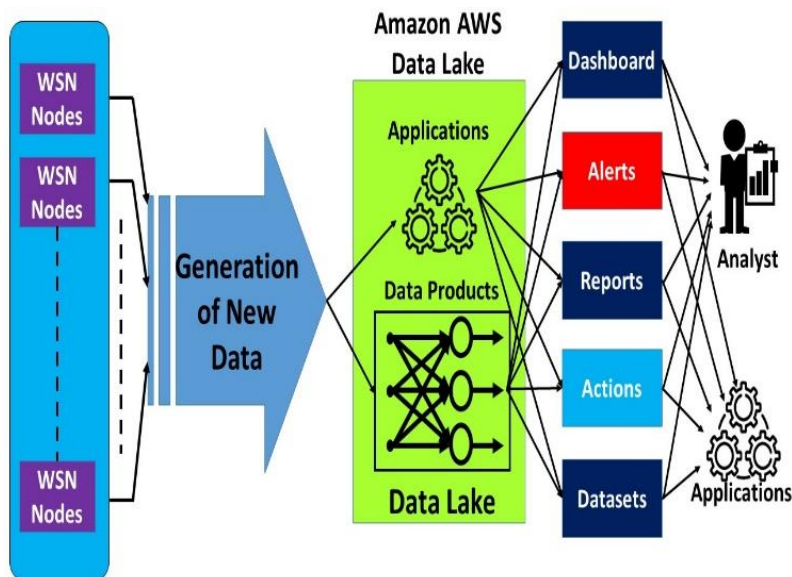


Fig. 8. The data lake storage model

4 Results & Discussion

The figure 9 gives a simplified AWS data lake architecture for WSN node. The data is acquired from various nodes in raw format then it is shifted to loading zone, to exploration zone, to refined zone in our case referred as raw data zone, processed data zone and curated data zone respectively. Finally the data is available for analytics purpose through Amazon SageMaker, Amazon Prediction, Amazon QuickSight or visualizations. The advantages of Amazon AWS data lake are scalability, the storing capacity can be scaled up instantly as per requirement, accessibility easy access to data, security against failures, errors and easy integration with third party APIs.

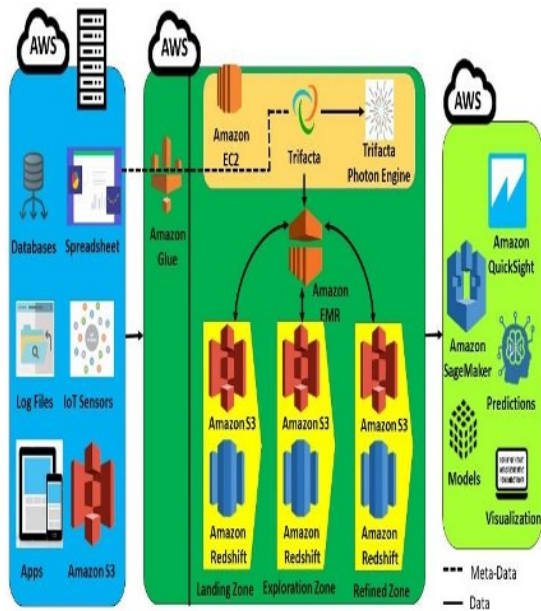


Fig. 9. Architecture of an AWS based WSN Data Lake

The figure 9 gives a simplified AWS data lake architecture for WSN node. The data is acquired from various nodes in raw format then it is shifted to loading zone, to exploration zone, to refined zone in our case referred as raw data zone, processed data zone and curated data zone respectively. Finally the data is available for analytics purpose through Amazon SageMaker, Amazon Prediction, Amazon QuickSight or visualizations. The advantages of Amazon AWS data lake are scalability, the storing capacity can be scaled up instantly as per requirement, accessibility easy access to data, security against failures, errors and easy integration with third party APIs.

The proposed design for WSN applications employs Amazon AWS for storing large amount of data generated at the nodes. The designed data lake serves as substitute to data warehouse. Earlier for this type of applications (like surveillance, monitoring etc.) data has to be stored in original version for future use. Use of Amazon AWS system solves a lot of problem posed by data warehouse:

1. Earlier during data aggregation a lot of useful details were curtailed resulting in limited types of analysis that can be performed. Figure 10 predicts proposed system gives 65% better results than data warehouse.

- 2. Many a times some analysis requires to be done at coarse granularity, which requires data to be stored in original version. Proposed design gives 85% better results as shown in figure 11.
- 3. The system is capable of generating error free reports, creating queries at different levels of complexity and granularity of data. The figure 12 shows proposed system is less error prone.
- 4. The system overhead was reduced for managing data at various levels of granularity and complexity of data. 68% lesser data overhead shown in figure13.

Processing and administrative overhead was also reduced for varied levels of complexity and granularity of data. Figure 14 shows 87% less processing and administrative overhead.

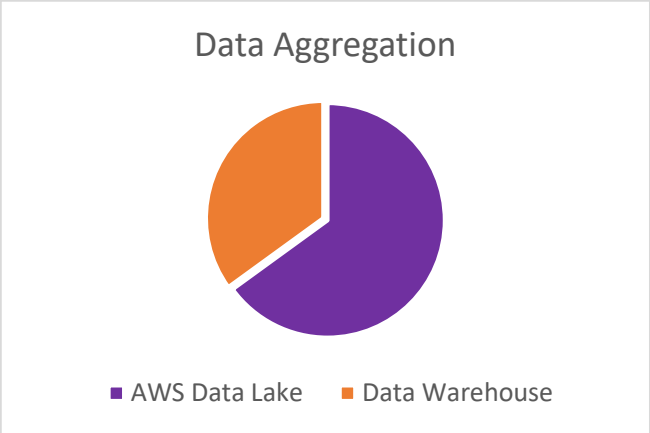


Fig. 10. Data aggregation in AWS data lake and data warehouse

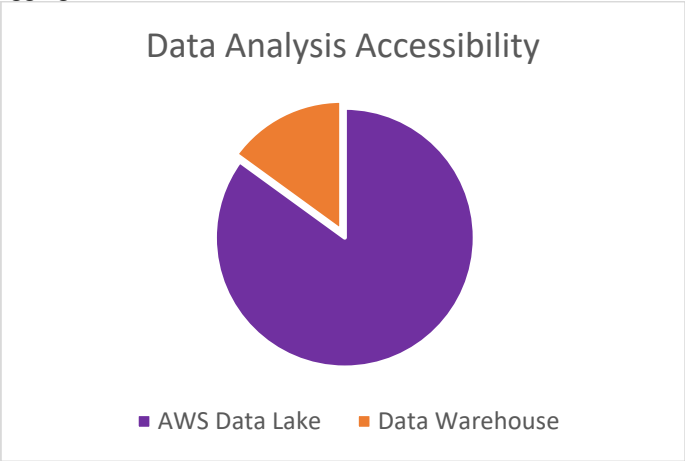


Fig. 11. Data analysis accessibility in AWS data lake and data warehouse

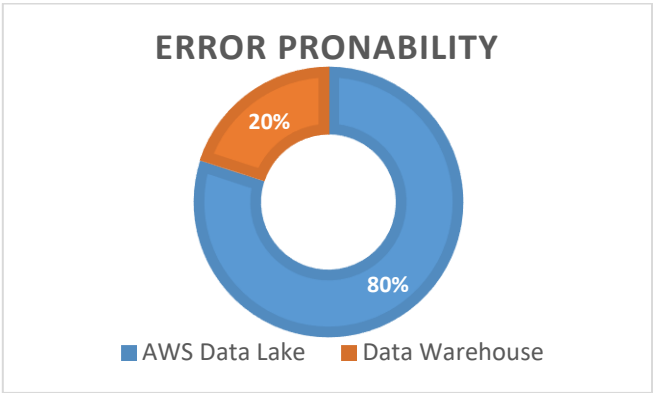


Fig. 12. Error pornability in AWS data lake and data warehouse

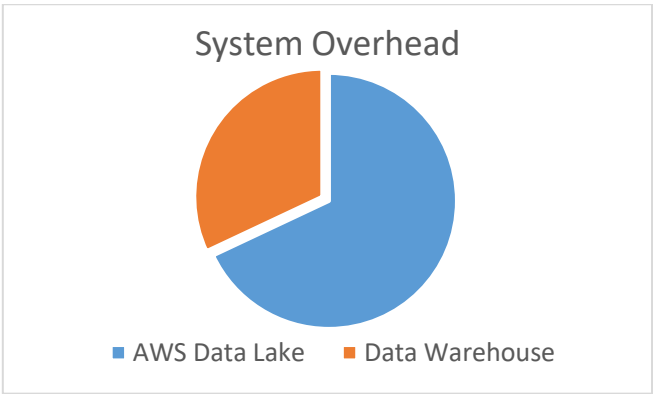


Fig. 13. System overhead in AWS data lake and data warehouse

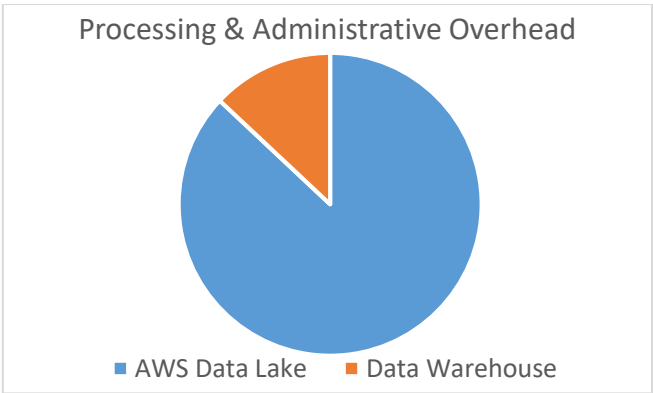


Fig. 14. Processing and administrative overhead in AWS data lake and data warehouse

5 Conclusion and Future work

In early days for typical applications which required data storage data warehousing was used. However, as the requirement for analytics is growing and certain applications need that even the tiniest bit of data should be available for analytics purpose, data warehouse becomes out dated and of no use, there the data lake comes to rescue. The data acquired from the WSN

nodes can be easily stored in the three zones of data lakes. The proposed data lake framework is implemented on Amazon AWS lake formation services. However the same data lake can be implemented using Microsoft Azure also. With the increase in the data volume speed and variety, the Microsoft Azure platform can also a low cost effective alternative to Amazon web based cloud services. The implementation of the WSN data lake house architecture can be done on Azure by utilizing the capabilities of Apache Spark and Delta lake thereby creating and implementing the required WSN Data lake solution on Data bricks, Synapse Analytics, and Snowflake. The data lake implementation using Microsoft Azure which employs Apache Spark and Delta lake are widely used for implementing now big data ELTs i.e. extraction loading and transformation.

References

1. E. Zagan and M. Danubianu, *Cloud DATA LAKE: The new trend of data storage*, 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1-4, doi: 10.1109/HORA52670.2021.9461293. (2021)
2. Abdallah Saleh Ali Shatat, Md. Mobin Akhtar, Abu Sarwar Zamani, Sara Dilshad & Faizan Samdani, *Big Data Driven Map Reduce Framework for Automated Flood Disaster Detection Based on Heuristic-Based Ensemble Learning*, Cybernetics and Systems, DOI: 10.1080/01969722.2022.2145664. (2022)
3. R. Chaudhry and N. Kumar, *A Multi-Objective Meta-Heuristic Solution for Green Computing in Software-Defined Wireless Sensor Networks*, in IEEE Transactions on Green Communications and Networking, vol. **6**, no. 2, pp. 1231-1241, doi: 10.1109/TGCN.2021.3122078. (2022)
4. A. M. Olawoyin, C. K. Leung and A. Cuzzocrea, "Open Data Lake to Support Machine Learning on Arctic Big Data", 2021 IEEE International Conference on Big Data (Big Data), pp. 5215-5224, doi: 10.1109/BigData52589.2021.9671453. (2021)
5. S. Bagwari, A. Gehlot, R. Singh, N. Priyadarshi and B. Khan, *Low-Cost Sensor-Based and LoRaWAN Opportunities for Landslide Monitoring Systems on IoT Platform: A Review*, in IEEE Access, vol. **10**, pp. 7107-7127, doi: 10.1109/ACCESS.2021.3137841. (2022)
6. H. Fang, *Managing data lakes in big data era: What's a data lake and why has it become popular in data management ecosystem*, 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 820-824, doi: 10.1109/CYBER.2015.7288049. (2015)
7. I. Ali, I. Ahmedy, A. Gani, M. U. Munir and M. H. Anisi, *Data Collection in Studies on Internet of Things (IoT), Wireless Sensor Networks (WSNs), and Sensor Cloud (SC): Similarities and Differences*, in IEEE Access, vol. **10**, pp. 33909-33931, doi: 10.1109/ACCESS.2022.3161929. (2022)
8. S. Ramchand and T. Mahmood, *Big data architectures for data lakes: A systematic literature review*, 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1141-1146, doi: 10.1109/COMPSAC54236.2022.00179. (2022)
9. Fei Wang, Hongxia Wang & Omid Ranjbar Dehghan, *Machine Learning Techniques and Big Data Analysis for Internet of Things Applications: A Review Study*, Cybernetics and Systems, DOI: 10.1080/01969722.2022.2103231. (2022)
10. S. A. Yadav and T. Poongodi, *A Review of ML Based Fault Detection Algorithms in WSNs*, 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), pp. 615-618, doi: 10.1109/ICIEM51511.2021.9445384. (2021)

11. R. Wrembel, *Still Open Problems in Data Warehouse and Data Lake Research: extended abstract*, 2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS), pp. 01-03, doi: 10.1109/SNAMS53716.2021.9732098. (2021)
12. N. Tripathi and K. K. Sharma, *Analytical Review of Different Routing Methodologies and Clustering Techniques in WSN/WCSN leading to Cooperative Communication*, 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1-6, doi: 10.1109/INOCON50539.2020.9298208. (2020)
13. C. Liu, *Research on the Method of Constructing Distributed Data Lake Driven by Virtualization Model*, 2020 International Conference on Computers, Information Processing and Advanced Education (CIPAE), pp. 223-226, doi: 10.1109/CIPAE51077.2020.00064. (2020)
14. Anne Laurent; Dominique Laurent; Cédrine Madera, *Introduction to Data Lakes: Definitions and Discussions*, in *Data Lakes* , Wiley, pp.1-20, doi: 10.1002/9781119720430.ch1. (2020)
15. Housseem Chihoub; Cédrine Madera; Christoph Quix; Rihan Hai, *Architecture of Data Lakes*, in *Data Lakes* , Wiley, pp.21-39, doi: 10.1002/9781119720430.ch2. (2020)
16. O. R. Ahutu and H. El-Ocla, *MAC Routing Protocol for Improving Efficiency in IEEE 802.15.4 Wireless Sensor Networks*, 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 0026-0030, doi: 10.1109/IEMCON51383.2020.9284899. (2020)
17. V. Sadhu, X. Zhao and D. Pompili, *Energy-Efficient Analog Sensing for Large-Scale and High-Density Persistent Wireless Monitoring*, in *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6778-6786, doi: 10.1109/IIOT.2020.2984484. (2020)
18. M. F. Khan, R. K. Dwivedi and R. Kumar, *Energy Efficient Data Transmission in Sensor Cloud : A Review*, 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 308-313, doi: 10.1109/ICOEI.2019.8862759. (2019)
19. R. K. Dwivedi, S. Singh and R. Kumar, *Integration of Wireless Sensor Networks with Cloud: A Review*, 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 114-119, doi: 10.1109/CONFLUENCE.2019.8776968. (2019)
20. D. Tracey and C. Sreenan, *How to see through the Fog? Using Peer to Peer (P2P) for the Internet of Things*, 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 47-52, doi: 10.1109/WF-IoT.2019.8767275. (2019)
21. D. Tracey and C. Sreenan, *Using a DHT in a Peer to Peer Architecture for the Internet of Things*, 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 560-565, doi: 10.1109/WF-IoT.2019.8767261. (2019)
22. G. M. Dias, C. B. Margi, F. C. P. de Oliveira and B. Bellalta, *Cloud-Empowered, Self-Managing Wireless Sensor Networks: Interconnecting Management Operations at the Application Layer*, in *IEEE Consumer Electronics Magazine*, vol. 8, no. 1, pp. 55-60, doi: 10.1109/MCE.2018.2868110. (2019)