Big Data Streaming using Kafka

Rehana Naguru

Northwest Missouri State University, Maryville MO 64468, USA
S545514@nwmissouri.edu, rehananaguru@gmail.com

Abstract. Big Data, itself refers to enormous amount of data. Due to increase of technology a lot of data is being generated. The need to generate the techniques to operate this voluminous data is increased. In this modern age, Apache Kafka is the approved architecture acclimated to process the data stream. Kafka provides the similar API to the message system.

Keywords: kafka · big data · apache · data · stream.

1 Introduction

In recent time, Technology is improving very fast. Digital products have become a part of everyone life, which in-turn generates the continuous data, which is called Data Stream. To process this large amount of data, standard methods are inefficient. Security issues will arise. Many researchers are aimed to overcome the problems with Big data.

Kafka is the better solution to handle data. Kafka, is first implemented by LinkedIn team. It was used to gather all the data operated by the users. It is the end to end message delivery system.

2 Review of the existing problem

In the early days, all the data in the companies are recorded in hard copies. This is ineffective as it is consuming huge space and time. It also causes security issues and more money is wasted to maintain. We invented the systems to operate with huge data. From past two decades the data utilization is touching the roofs. A prediction of 175 Zettabyte of data will be used in 2025. To overcome this issue, Big data is helpful to store vast amount of data. This solved many issues.

In this fast moving era, technology is growing at a rapid pace. There is a constant update in Big Data. This concept is applied in many industries and are becoming smart factories.

3 Requirement

In the cloud computing,the generalized process is collecting the data and sending to the System which is Centralized for storing and to perform computations. As

there is a rapid increase in data complexity, data storage and processing of data is challenging. The main requirement is to collect, process and perform aggregation on a local device to avoid the huge traffic of storing back to the cloud.

4 How Kafka Works

Apache Kafka is the open-source software. It is preowned to process the streams of data. For each data record generated, Key, value and Timestamp is present. All these records are grouped to the Topics, which consists of the partitions of data. These are all read by different consumers.

Fig. 2 Working of Kafka.

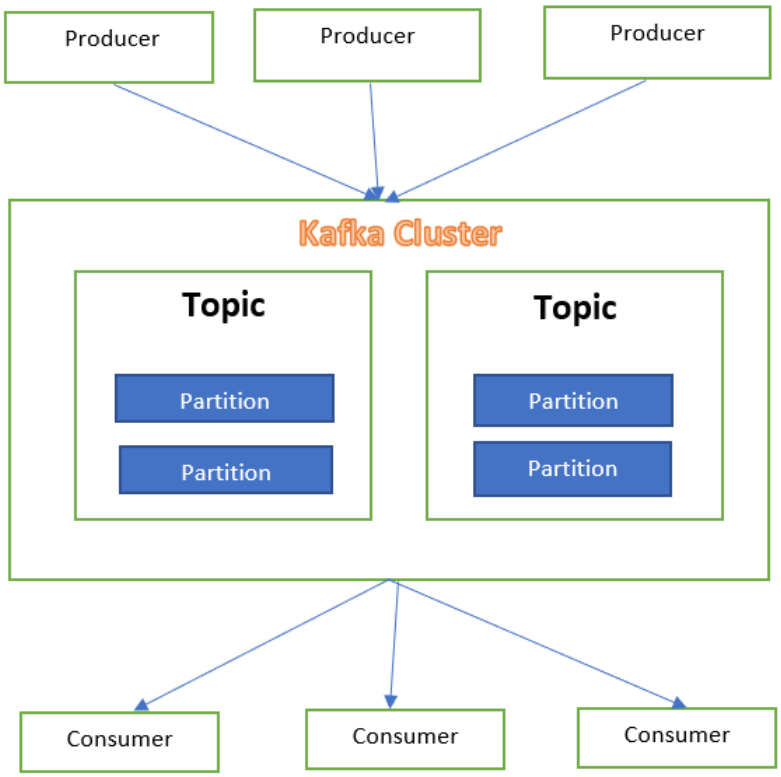

Fig. 1. Working of Kafka

5 System Architecture

The entire system is the integration of Hadoop ecosystem with Kafka and is classified into three major parts: Data Collection Layer, Streaming and Processing Layer, and Data Storage Layer, which is represented in figure 1.

Fig. 2 System Architecture.

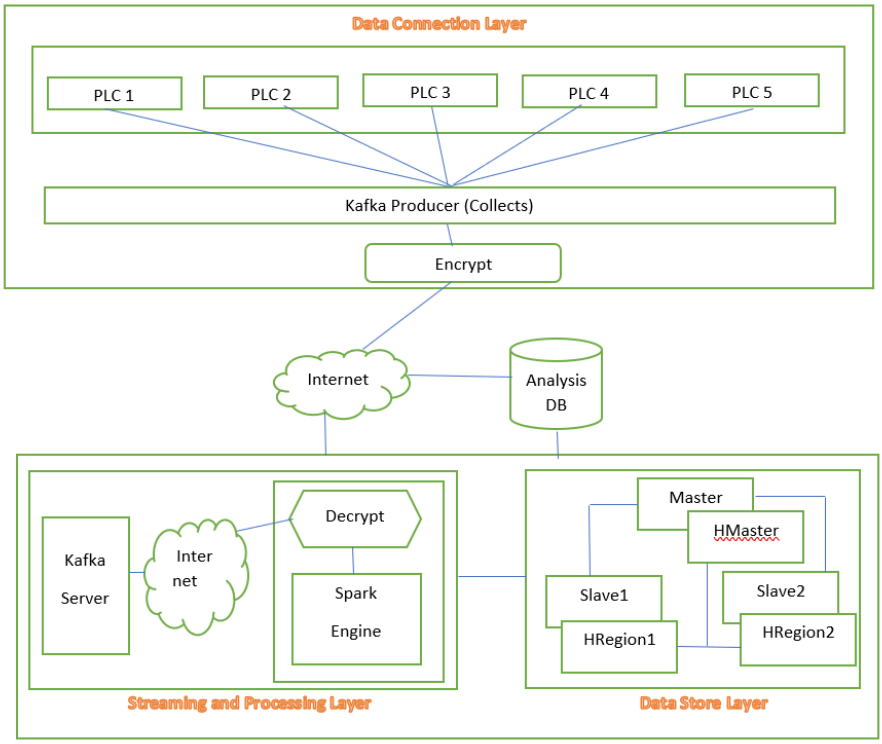

Fig. 2. System Architecture

Primarily, the Data Collection Layer, collects the data from the PLCs(Programming Logic Controllers) and it transfers the data to the Kafka Producer. To reduce message size and and to encrypt the message, a public key is included in this Kafka Producer. In the Kafka server the message is stored temporarily before sending to the other locations. During this a notification is send to kafka consumer, that it received a message. Later, a private key is comprised in kafka consumer for decryption. It extricate the message and perform real time data processing through Apache Spark Cluster.Further, data is stored in the HDFS and HBase of Data Store Layer. It contains two nodes and one master. All the data collected through PLCs, contains status of machine such as Product infor-

mation, stop, wait and run. On the whole, to authorize the Big Data, this three layers have to function in conjunction.

6 How Kafka is different?

Kafka is most reliable. During the failure, it naturally balances the consumers. It offers the elevated performance. Even if handling the huge terabytes of data,it always gives the consistency in execution. Kafka is vastly scalable and also extremely durable.

7 Use Cases

- Apache Kafka can be used for the operational monitoring data. It requires collecting the statistics from the scattered applications to generate the centralized feeds of the operational data.

- Kafka operates the data in the processing pipelines which consists of several stages. For this original data is consumed from the Kafka topics and later collected and transformed to the new topics for later operations.

- Apache Kafka can also be used over the organization for gathering the logs from various services and it will make them accessible in a common format to the different consumers.

8 Kafka Advantages

- Kafka is having many advantages. The main advantage is the Fault-tolerance, which is an intrinsic competency in the Kafka. This is impenetrable to the failure of machine in the cluster.

- Kafka is Distributed. With Kafka we can manage many messages at a time. With high concurrency Kafka is able to read and write the messages.

- Inside of a Kafka Cluster, manipulation of the message is highly translucent and are consistent. It is scalable.

- Kafka is consumer friendly. By using this we can incorporate diversification of consumers. On the report of consumer Kafka can work differently. In variation of languages, Kafka can work well.

- Data pipelines can be operated by Kafka. From the real- time applications, we can also operate concurrent messages.

- Kafka is able to operate the messages which are reduced latency with the reach of milliseconds.

- Difference of use cases are normally vital for the Data Lake, which can be managed by Kafka. It can also perform the tasks of the conventional ETL.

- Messages are always safe with the Kafka. No way of losing messages. Elevated velocity and volume of data can be easily operated by Kafka. This will support the production of many messages for each second.

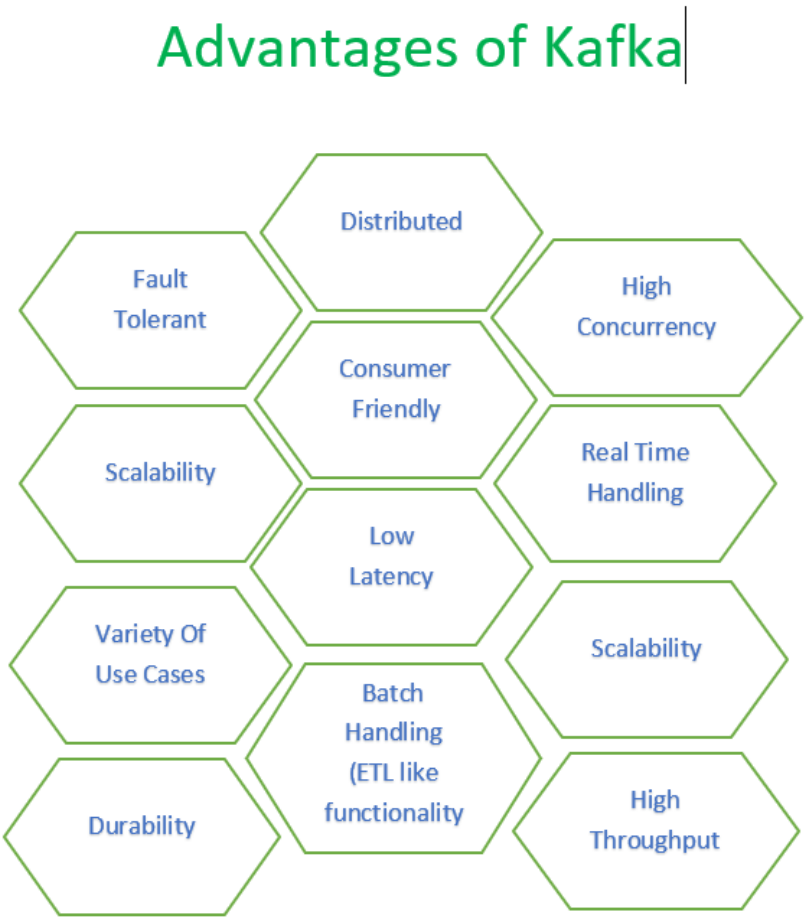

Fig. 3. Advantages of Kafka

9 Kafka Disadvantages

Even having many advantages, Kafka have some limitations. If the message required any improvement, then the performance of the Kafka will diminish remarkably. As it uses the proficiency of the system, Kafka will perform skillfully if the data is not changed.

If the aggregate of queues increases, then the Kafka will perform a bit relaxed. As Kafka works with different languages, they require API's , which are prolonged by discrete individuals. So the problem arises with slowdown of speed.

For reinforcing eventually, Kafka is not supportive due to the need of tools for monitoring and management. Some times it also reduces the performance.

10 Conclusion

This paper mainly discusses about the existing issues with the Big Data, resolving of these issues using the Kafka. Kafka is the better solution used to handle the huge volumes of data. This is the end to end message delivery system. Apache Kafka is the open-source software. It is predefined to process the streams of data. It also discussed about the streaming of data through kafka and its use cases. It clearly explained about the System Architecture of Kafka. Advantages and disadvantages are clearly explained in this paper.

11 Acknowledgements

I would like to thank Dr. Ajay Bandi, Sowmya Yalamarthi and Varshith Bairy for helping me in this Research Paper.

References

1. A containerized big data streaming architecture for edge cloud computing on clustered single-board devices.
2. Aparna, K., Sudeep, T., Sudhanshu, T., Neeraj, K.: Verification and validation techniques for streaming big data analytics in internet of things environment. IET Networks **8**(3), 155 – 163
3. Bandi, A., Hurtado, J.A.: Big data streaming architecture for edge computing using kafka and rockset. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). pp. 323–329 (April 2021). https://doi.org/10.1109/ICCMC51019.2021.9418466
4. Boyanov, L.: System components for data extraction and processing from internet of things. Knowledge: International Journal **47**(3), 463 – 467
5. D., A., R.I., M.: Edge computing based surveillance framework for real time activity recognition. ICT Express **7**(2), 182 – 186
6. Erum, M., Tayyaba, A.: Challenges and solutions for processing real-time big data stream: A systematic literature review. IEEE Access **8**, 119123 – 119143

7. Hiraman, B.R., Viresh M., C., Abhijeet C., K.: A study of apache kafka in big data stream processing. In: 2018 International Conference on Information , Communication, Engineering and Technology (ICICET). pp. 1–3 (Aug 2018). https://doi.org/10.1109/ICICET.2018.8533771
8. Leang, B., Ean, S., Ryu, G.A., Yoo, K.H.: Improvement of kafka streaming using partition and multi-threading in big data environment. Sensors (14248220) **19**(1), 134
9. Shafqat, S., Kishwer, S., Rasool, R.U., Qadir, J., Amjad, T., Ahmad, H.F.: Big data analytics enhanced healthcare systems: a review. Journal of Supercomputing **76**(3), 1754 – 1799
10. Zhu, R., Liu, L., Song, H., Ma, M.: Multi-access edge computing enabled internet of things: advances and novel applications. Neural Computing Applications **32**(19), 15313 – 15316