International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

THE EVOLUTION OF KAFKA STREAMS - FROM A MESSAGING SYSTEM TO REAL-TIME DATA STREAMING PIPELINES

Hareesh Kumar Rapolu
hareeshkumar.rapolu@gmail.com

Abstract

The research paper has properly highlighted the way in which Kafka Streams has transformed itself from being just a messaging system to a fully functional data-processing platform. The different applications of Kafka Streams have been thoroughly assessed within this research paper. Furthermore, the study has also provided insights into the different advantages and disadvantages of using Kafka Streams. Finally, a number of recommendations have been highlighted in the research paper that can help resolve the drawbacks of the platform. Thus, Kafka streams have proved to be effective in delivering outstanding results for processing real-time data streaming. This has ultimately served with valuable outputs in generating input and output data which has been stored in Kafka platforms.

Keywords- Kafka Streams, real-time, stream-processing applications, pipelines, client library

I. INTRODUCTION

Kafka Streams are used by both individuals and organisations in order to process large volumes of data in real-time. It is a powerful library that is used for crafting stream-processing applications with the help of Apache Kafka. This particular research paper will critically evaluate the way in which Kafka Streams gradually evolved from a simple messaging system to a platform that helps to process real-time data streaming pipelines. It will further analyse the different kinds of applications of Kafka Streams in the modern scenario. The study will also outline a few advantages and disadvantages of Kafka Streams. In the final portion, a few recommendations will be provided that can help mitigate the drawbacks of the Kafka Streams client library.

II. EVOLUTION OF KAFKA STREAMS

Initially, Kafka Streams served as a messaging system which used the Apache Kafka platform. Messages could be reliably delivered between different applications. Therefore, this is how

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

producers were able to publish different kinds of data on various topics. The consumers were able to subscribe to these different topics according to their preferences and get relevant messages on them. However, with the passage of time, Kafka Streams transformed into one of the most powerful tools that is effective for building real-time data streaming pipelines[1]. It provides high throughput, efficient and distributed storage of data. Additionally, there is a stream processing layer that enables the users to process data in real time when it passes through different Kafka topics. Therefore, it is evident that Kafka Streams is not just about sending a particular message and receiving another. Over time, it has rapidly evolved and transformed itself into a proper platform that provides stream processing services in real-time[2]. Therefore, it can now be used for sophisticated processing of large streams of data. Most importantly, real-time analytics can be gained, which is very important in the modern world. The gradual evolution of the messaging system into a full-fledged platform with a range of services is mainly due to the changing demands of the customers. The dynamic nature of Kafka Streams can solve problems related to data and process large volumes of it in a jiffy[3]. If S is the input stream and certain mapping functions f are applied to it, then S′ will be the transformed output stream. Additionally, Kafka Streams can effectively summarise or find the aggregate A(k) of certain values with the help of a function f.

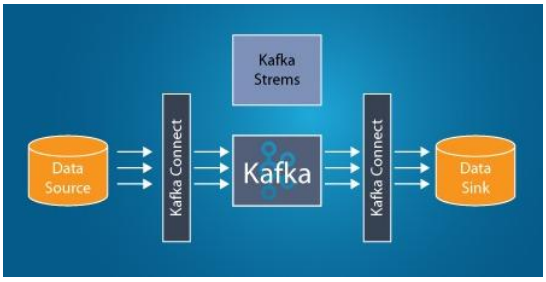

Figure 1: Design of Kafka Streams

III. APPLICATIONS OF KAFKA STREAMS

Kafka Streams play an important role in properly monitoring the live data streams in real time. Since it has the capability of generating accurate insights on such data, the users are able to monitor their different activities like website traffic, system performances, or fluctuating stock

market prices. It helps them to make accurate decisions by leveraging the authentic insights generated by the Kafka Streams.In the modern world, the online world is largely plagued by hackers who can steal personal information. In this context, Kafka Streams can really help to provide protection from these fraudulent activities. Since Kafka Streams has an advanced algorithm, it helps users to identify unconventional patterns in the system[4]. Therefore, the users are able to take appropriate actions on time. This helps to minimise potential data and financial losses due to unethical online activities.Observing a particular business, Kafka Streams is extremely crucial since it is able to analyse the behaviours of the customers. Therefore, businesses can obtain personalised product recommendations that can considerably improve their potential.

IV.ADVANTAGES AND DISADVANTAGES OF KAFKA STREAMS

The first advantage is that the design of Kafka enables it to handle both batch and real-time data in a flawless manner. This efficiency is the main reason why it is widely used by individuals and businesses alike. The second advantage is that at times, a lot of data is transmitted every day in a very fast-paced environment. Therefore, Kafka Streams' ability to process large volumes of data in real-time helps the users to optimise their operations[5].Security is identified as the third advantage. There are a number of security features that are integrated into Kafka Streams like authorisation, encryption and authentication.The first disadvantage with respect to Kafka Streams is that the users can find it difficult to learn all the operations of Kafka Streams due to its complexity. A good amount of time is to be invested by the user to master the platform.The second disadvantage is inefficient failure recovery. There is no checkpoint mechanism in Kafka Streams which is detrimental in the case of a total system failure. In those instances, the users can find it difficult to retrieve the lost data.

V. RECOMMENDATIONS

The first recommendation is to provide systematic learning. Here, the user needs to understand the basic concepts within Kafka in a sequential manner. It can help to process and retain the acquired knowledge for a longer amount of time[6]. Different online resources can be utilised that provide real-life practical applications of the Kafka Streams platform.Using topic replication is observed as the second recommendation. This method can be instrumental in retrieving lost

data if a server fails suddenly. In this technique, each topic is divided into a few partitions. These partitions are then replicated and stored on different servers. Therefore, if one server fails, the data can be recovered from another one.

Figure 2: Ways to mitigate the drawbacks of Kafka Streams

VI. CONCLUSION

From the discussion, it can be concluded that Kafka Streams is one of the widely used platforms in different business sectors. It utilises Apache Kafka in order to efficiently process data in real-time and build stream-processing applications[7]. There is a high-level domain-specific language which is beneficial for transforming and evaluating long streams of data.

A. Abbreviations and acronyms

KStream - Kafka Stream
KTable - Kafka Table
DSL - Domain-specific language

B. Units

- KStream - Stream of records
- KTable - Changelog stream of the latest state

C. Equations

Stream processing - $S'=f(S)$
Aggregate of values $A(k)=f(v1,v2,\ldots,vn)$

REFERENCES

1. B. Leang, S. Ean, G.-A. Ryu, and K.-H. You, "Improvement of Kafka Streaming Using Partition and Multi-Threading in Big Data Environment," *Sensors*, vol. 19, no. 1, Jan. 2019, doi: https://doi.org/10.3390/s19010134
2. B. V. S. Srikanth and V. K. Reddy, "Efficiency of Stream Processing Engines for Processing BIGDATA Streams," *Indian Journal of Science and Technology*, vol. 9, no. 14, Apr. 2016, doi: https://doi.org/10.17485/ijst/2016/v9i14/84797
3. H. Isah, T. Abughofa, S. Mahfuz, D. Ajerla, F. Zulkernine, and S. Khan, "A Survey of Distributed Data Stream Processing Frameworks," *Ieee.org*, vol. 7, Oct. 2019, doi: https://doi.org/10.1109/ACCESS.2019.2946884. Available: https://ieeexplore.ieee.org/iel7/6287639/8600701/08864052.pdf
4. K. M. M. Thein, "Apache Kafka: Next Generation Distributed Messaging System KHIN ME ME THEIN," *International Journal of Scientific Engineering and Technology Research*, vol. 3, no. 47, Dec. 2014, Available: https://ijsetr.com/uploads/ 436215IJSETR3636-621.pdf
5. M. H. Javed, X. Lu, and D. K. (DK) Panda, "Characterization of Big Data Stream Processing Pipeline," *Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, Dec. 2017, doi: https://doi.org/10.1145/3148055.3148068
6. M. Laska, S. Herle, R. Klamma, and J. Blankenbach, "A Scalable Architecture for Real-Time Stream Processing of Spatiotemporal IoT Stream Data—Performance Analysis on the Example of Map Matching," *ISPRS International Journal of Geo-Information*, vol. 7, no. 238, Jun. 2018, doi: https://doi.org/10.3390/ijgi7070238
7. O. Soumaya, T. Mohamed Amine, A. Soufiane, D. Abderrahmane, and A. Mohamed, "Real-time Data Stream Processing - Challenges and Perspectives," *International Journal of Computer Science Issues*, vol. 14, no. 5, pp. 6–12, Sep. 2017, doi: https://doi.org/10.20943/01201705.612