ETL data pipeline using distributed streaming platform

SAZAL KANTI KUNDU, Brac University, Dhaka, Bangladesh

We are living in a world of data. Without data it is hard to do almost anything. Data is the central part of technologies like artificial intelligence, machine learning, data analytic, data engineering etc. So it becomes vital for the data to flow from one system to another system and in between get transformed as needed. Also, it needs to flow very fast so that the target system can make use of it in real time. Distributed streaming platform like kafka can sit in between and keep the data in motion reliably.

1 INTRODUCTION

ETL stands for extracting, transforming and loading. It is a process to get data from one or more sources to one or more target systems. The widespread use of internet, social media, IoT devices etc are producing a huge amount of data. Technology companies are taking data seriously. From machine learning to analytic engines consuming data to make sense out of it. So it is needed to move data in real time and process them as quickly as possible to make intelligent decision. To keep data in motion distributed streaming platforms like kafka need to be revisited. While data is in motion it also needed to keep them in the platform for a while for other processing instead of pulling data again from the sources. So the idea is streaming data and at the same time storing them as per need without using yet another data store. As kafka itself is distributed so fault tolerance, reliability, scalability is there by design.

2 RELATED WORKS

With the emergence of big data ETL process was revisited from time to time. X. Liu, C. Thomsen, and T. B. Pedersen adopted MapReduce paradigm and implemented a prototype called ELTMR which is a MapReduce version of the PygramETL prototype. Researchers also tried to implement ETL using Apache Hadoop. C. Thomsen, and T. B. Pedersen also proposed CloudETL framework which uses Hadoop to parallelize ETL process and Apache Hive to process the data.

3 METHODOLOY

Without using complex frameworks like MapReduce we can use Apache Kafka which is a distributed streaming platform. Extractor, Transformer and Loader components will be connected by Kafka. ETL components can be handled in a microservice based architecture. This will allow to scale each of the ETL components independently based on the load. As kafka will sit in between, all the ETL components will be decoupled. Producer will produce contents without any knowledge of consumer. Consumer might be in maintenance mood. Kafka will act as a buffer and store the data reliably for the consumers. As kafka is distributed in nature so it will have its own cluster and data will be distributed across the cluster. Consumers can be scaled independently as the load increases. Kafka will distribute the data among the consumer process as it scales up or down.

4 CONCLUSION

The ETL is the core component of the systems that rely on data heavily. All the data meant for analysis pass through this process. So it should be adapted so that it can handle the diverse data coming from heterogeneous sources very rapidly. In this paper we proposed a distributed approach for ETL process where our priority was data. Our approach will be able to store the data reliably and stream it in real time.