A MapReduce-based Inter-Organizational and Distributed Process Log Clustering Framework

Thanh-Hai Nguyen[†]
nguyenthanhhai@kgu.ac.kr
Data and Process Engineering Research Lab.
Department of Computer Science, Graduate School of
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Kyoung-Sook Kim[*]
khmjmc@kgu.ac.kr
Contents Convergence Software Research Institute
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Sang-Eun Ahn[‡]
sangeun1804@kgu.ac.kr
Contents Convergence Software Research Institute
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Kwanghoon Pio Kim[§]
kwang@kgu.ac.kr
Data and Process Engineering Research Lab.
Division of AI Computer Science and Engineering,
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

1 INTRODUCTION

The main research challenge of this paper is to develop a process log clustering technique based upon the MapReduce [4] platform, which will be eventually applied to performing the process mining system that was developed for a process model of structured information control nets from a large scale and inter-organizational process enactment event log dataset. We also assume that the typical process mining algorithm to be embedding the process log clustering framework proposed in this paper is the ρ-Algorithm [3] and its implemented system.

2 FRAMEWORK

We implemented the MapReduce-based inter organizational process BIG-log clustering algorithm and verify the functional correctness of the implemented framework. In terms of implementation considerations, we assume that the input and output files are supported by a block-based distributed file system. In other words, in the block-based distributed file system, each input/output file is managed in fixed-size block units (default 64MB), and each block basically has two copies for recovery in case of failure. Also, in carrying out an experimental verification, we use a process execution event log dataset of the Customer Summary process model, which is formatted in the IEEE-XES standardized schema [2] [2] and available at the 4TU.ResearchData [1] website.

To verify the MapReduce-based inter-organizational process BIG-log clustering system, we used three horizontally fragmented datasets from the original Customer Summary process dataset. A MapReduce job basically consists of a map-phase that executes the method Map( ) function and a reduce-phase that executes the method Reduce( ) function. The Combine( ) function is intended to reduce the I/O cost required to transfer data from the mapper to the reducer by performing the aggregate operation in advance in the map-step. Basically, the reducers receive intermediate results recorded on the local disk of each mapper through HTTPS when all mappers are terminated. At the same time, each reducer receives all lists of pattern-key values assigned to it to process from all mappers. This process is called shuffling because it is similar to the act of shuffling cards in a playing card, and all values for each pattern-key collected in this way are merged into a list. After that, the method Reduce( ) function is applied to each pattern-key value and the result is recorded in the distributed file system.

As a consequence, we successfully implemented the inter organizational process BIG-log clustering algorithm operating upon the temporal work-case driven MapReduce platform, and plugged this implemented algorithm in the process mining system as a preprocessing function. By using the process mining system, we carried out an verificational experiment with fragmenting the original Customer Summary process log dataset into three groups of horizontally distributed temporal work-case traces.

3 CONCLUSION

In this paper, we proposed a MapReduce-based inter-organizational process log clustering framework, which is one of the preprocessing techniques to be ultimately used in process mining technology to discover and rediscover process models from the distributed and fragmented execution event logs of large-scale inter-organizational process models. In particular, the proposed clustering framework satisfies the 5V-property of bigdata, which is specifically defined as process BIG-log in this paper, and is implemented by developing the MapReduce-based inter-organizational process log clustering algorithm as an efficient pre-processing method based on fragmented temporal work-cases.

REFERENCES

[1] Centre for Research Data 4TU. 2012, 2013, 2014, 2015, 2016, 2017, 2018. BPI Challenges. BPM.
[2] IEEE. 2016. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. IEEE 1849-2016.
[3] Kyoung-Sook Kim, Dinh-Lam Pham, and Kwanghoon Pio Kim. 2021. ρ-Algorithm: A SICN-Oriented Process Mining Framework. *IEEE Access* 9 (2021), 139852–139875. https://doi.org/10.1109/ACCESS.2021.3119011
[4] K.-H. Lee, W. J. Park, K. S. Cho, and W. Ryu. 2013. The MapReduce framework for Large-scale Data Analysis: Overview and Research Trends. *Electronics and telecommunications trends* 28, 6 (2013), 156–166. https://doi.org/10.22648/ETRI.2013.J.280616