ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 6, July 2017

A Novel Apriori Improved Algorithm for MapReduce Architecture

*Chandana G, *Pavithra V, *Meghana M , *Nitin Kumar, #Vandana B

*UG Students, Dept. of CSE, RRCE, Bengaluru, India

#Assistant Professor, Dept. of CSE, RRCE, Bengaluru, India

ABSTRACT: This Under the environment of big data, efficiency is low and there are many candidates when the traditional serial Apriori algorithm in dealing with massive data. This paper proposes a parallel better algorithm based on MapReduce distributed architecture. Based on the basic Apriori algorithm on MapReduce, this paper makes a reconstruction of the original transaction database, and implements parallel in data set fragmentation. The algorithm optimizes the transaction database; candidate item sets counting and pruning strategy. The experimental results show that the improved algorithm proposed in this paper can reduce the candidate items and improve the efficiency.

KEYWORDS: Big data, Apriori, MapReduce.

I. INTRODUCTION

With the rapid development of computer technology and the further application of the Internet, the data show explosive growth. Hundreds to thousands of PB level data make electricity, finance and scientific computing areas distress, because the traditional computing technology and systems have been unable to cope and meet the requirements of the calculation. Big data technology brings a new way for the discovery of the potential value of massive data. The mining of association rules can make people find a lot of potential and valuable information from the huge and complex data in large data environment. Association rule mining has become a very important research direction in data mining [1]. A challenging topic is to study the association rule algorithm in the big data technology. The classical algorithm of association rule mining algorithm includes Apriori algorithm and FP-Growth algorithm. Traditional Apriori algorithm multiply scanned repeatedly the original database and had low efficiency when dealing with massive data. Communication is large and node failure is easy to occur of the Apriori parallel algorithm based on MPI programming model which may use expensive computer. In the big data environment, these shortcomings of the apriori algorithm are particularly prominent. Although many scholars have carried out a variety of parallelization in the traditional apriori algorithm, but still it need to repeatedly scan the original database, and the efficiency is not high.

II. SYSTEM ARCHITECTURE

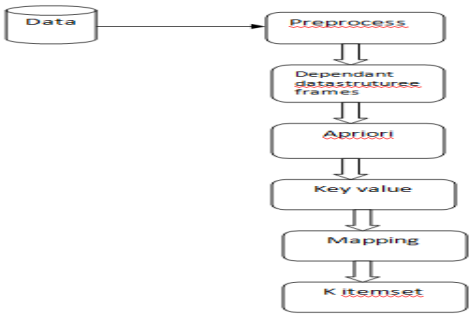

Fig. 1: System Architecture

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 6, July 2017

Data: means that for the process of this algorithm first they required data to it. So we first gather the data from the user that data are the inputs to this flow.

Preprocess: Preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Mapping: data mapping is the process of creating data element mappings between two distinct data models. Data mapping issued as a first step for a wide variety of data integration tasks..

Apriori Algorithm: it is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending then to larger item sets as long as those item set appear sufficiently often in the database.

III. CONCLUSION

Aiming at shortcomings of Apriori algorithm whose I/O load, multi candidate at the large-scale data processing, this paper analyse the advantages and disadvantages of various solutions, and propose an Apriori parallel algorithm based on MapReduce.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Concepts andtechniques of Data mining, Beijing, 2007, pp. 45-46.
[2] R Agrawal, R Srikant,"Fast algorithms for mining association rules," Proceedings of the 1994 International Conference on Very Large Databases, Santiago, Chile, 1994, pp.487-499.
[3] Chen Jing, Zhang Yan, "Visualization of Apriori-Partition algorithm based on association rules," Microcomputer Information, 25, pp.190-191,Mar, 2009.
[4] J. S. Park, M. S. Chen, and P.S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules," Proceedings ACM SIGMOD International Conference on Management of Data, SanJose, 1995, pp.175-186.
[5] Lv Wanqi, Zhong Cheng, and Tang Yinhu, "Parallel Mining under Hadoop distributed architecture of large data sets," Computer Technology and Development, 24, pp.22-26, Jan, 2014.
[6] Yu Qian, WEI Cheng, and Wang Kai, "MapReduce resource scheduling algorithm based on machine learning," Application Research of Computers, 33, pp.111-114, Jan, 2016.
[7] Li Aiguo, She Xiangyang, Data Mining Theory, Algorithms and Applications, Xi'an, 2012, pp.30-31.
[8] Xie Pengjun, "Parallel Research of Frequent Item Sets Mining Based on MapReduce," Nanjing University,Nanjing, 2012.