World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:10, No:03, 2016

RA-Apriori: An Efficient and Faster MapReduce-Based Algorithm for Frequent Itemset Mining on Apache Flink

Authors : Sanjay Rathee, Arti Kashyap

Abstract : Extraction of useful information from large datasets is one of the most important research problems. Association rule mining is one of the best methods for this purpose. Finding possible associations between items in large transaction based datasets (finding frequent patterns) is most important part of the association rule mining. There exist many algorithms to find frequent patterns but Apriori algorithm always remains a preferred choice due to its ease of implementation and natural tendency to be parallelized. Many single-machine based Apriori variants exist but massive amount of data available these days is above capacity of a single machine. Therefore, to meet the demands of this ever-growing huge data, there is a need of multiple machines based Apriori algorithm. For these types of distributed applications, MapReduce is a popular fault-tolerant framework. Hadoop is one of the best open-source software frameworks with MapReduce approach for distributed storage and distributed processing of huge datasets using clusters built from commodity hardware. However, heavy disk I/O operation at each iteration of a highly iterative algorithm like Apriori makes Hadoop inefficient. A number of MapReduce-based platforms are being developed for parallel computing in recent years. Among them, two platforms, namely, Spark and Flink have attracted a lot of attention because of their inbuilt support to distributed computations. Earlier we proposed a reduced- Apriori algorithm on Spark platform which outperforms parallel Apriori, one because of use of Spark and secondly because of the improvement we proposed in standard Apriori. Therefore, this work is a natural sequel of our work and targets on implementing, testing and benchmarking Apriori and Reduced-Apriori and our new algorithm ReducedAll-Apriori on Apache Flink and compares it with Spark implementation. Flink, a streaming dataflow engine, overcomes disk I/O bottlenecks in MapReduce, providing an ideal platform for distributed Apriori. Flink's pipelining based structure allows starting a next iteration as soon as partial results of earlier iteration are available. Therefore, there is no need to wait for all reducers result to start a next iteration. We conduct in-depth experiments to gain insight into the effectiveness, efficiency and scalability of the Apriori and RA-Apriori algorithm on Flink.

Keywords : apriori, apache flink, Mapreduce, spark, Hadoop, R-Apriori, frequent itemset mining
Conference Title : ICPP 2016 : International Conference on Parallel Processing
Conference Location : Paris, France
Conference Dates : March 14-15, 2016