

Market Basket Analysis Algorithm with no-SQL DB HBase and Hadoop

Jongwook Woo, Siddharth Basopia, Yuhang Xu
Computer Information Systems Department
California State University, Los Angeles, CA, USA
{jwoo5, sbasopi, yxu10}@calstatela.edu

Seon Ho Kim
Integrated Media Systems Center
University of Southern California, Los Angeles, CA, USA
seonkim@usc.edu

no-SQL DB has received highlights to maintain huge volumes of data based on the concept of Google's BigTable. HBase is one of no-SQL DBs, which run on Apache Hadoop Distributed File Systems (HDFS) and to handle big data. The paper presents a HBase schema to process transaction data for Market Basket Analysis algorithm. The algorithm runs on Hadoop Map/Reduce by reading data from both HBase and HDFS. Then, it sorts and converts the transaction data to data set with (key, value) pair, and stores the data to the HBase or HDFS. It is executed on Amazon EC2 platform instantiated by Whirr. The experimental results show that the code with Map/Reduce increases the performance as adding more nodes but at a certain point, there is a bottle-neck that does not allow the performance gain. Besides, executing the algorithm with data in HBase is slower than in HDFS.

Keywords: HBase, no-SQL, Map/Reduce, Market Basket Analysis, Data Mining, Hadoop, Whirr, Cloud Computing

1. INTRODUCTION

Data gets bigger and reaches peta- or tera-bytes as the web has grown in the world. Before Internet and Web did not exist, we did not have enough data to analyze people, society, and science etc with the limited volumes of data. However, since Internet and web came out, it has been more difficult to store and analyze data because of its huge volumes tera- or peta-bytes of data. *Google* faced to the issue as collecting big data and the existing file systems and Relational Database Management Systems (RDBMS) were not sufficient to store and handle the data efficiently. Besides, the legacy computing power and platforms were not useful for the big data. Thus, Google implemented Google File Systems (*GFS*), BigTable, and Map/Reduce parallel computing platform, which *Apache Hadoop* and *HBase* projects are motivated from.

Hadoop is the parallel programming platform built on Hadoop Distributed File Systems (*HDFS*) for Map/Reduce computation that processes data as (key, value) pairs. *HBase* runs on *HDFS* with *Hadoop Map/Reduce* to store and process big data. *HBase* and *Hadoop* have been adopted dramatically for the enterprise computing because business world always has the big data such as log files for web transactions, which is not easy to store and compute. Especially, when *Amazon AWS* supports *Hadoop* instances, it becomes much easy for people to run

- [3] “Building a business on an open source distributed computing”, Bradford Stephens , O'Reilly Open Source Convention (OSCON) 2009, July 20-24, 2009, San Jose, CA
- [4] “MapReduce Debates and Schema-Free”, Woohyun Kim, Coord, March 3 2010
- [5] “Data-Intensive Text Processing with MapReduce”, Jimmy Lin and Chris Dyer, Tutorial at the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), June 2010, Los Angeles, California
- [6] “ Introduction to Cloud Computing”, Jongwook Woo, the 10th KOCSEA 2009 Symposium, UNLV, Dec 18-19, 2009
- [7] “The Technical Demand of Cloud Computing”, Jongwook Woo, Korean Technical Report of KISTI (Korea Institute of Science and Technical Information), Feb 2011
- [8] “Market Basket Analysis Example in Hadoop”, Jongwook Woo, <http://dal-cloudcomputing.blogspot.com/2011/03/market-basket-analysis-example-in.html>, March 2011
- [9] “SQL MapReduce framework ”, Aster Data, <http://www.asterdata.com/product/advanced-analytics.php>
- [10] Apache HBase, “<http://hbase.apache.org/>”
- [11] “Data-Intensive Text Processing with MapReduce”, Jimmy Lin and Chris Dyer, Morgan & Claypool Publishers, 2010.
- [12] Apache Whirr, “<http://incubator.apache.org/whirr/>”
- [13] “How to set up Hadoop and HBase together with Whirr on Amazon EC2”, Jongwook Woo, <http://dal-cloudcomputing.blogspot.com/2011/06/how-to-set-up-hadoop-and-hbase-together.html>, June 2011
- [14] “Whirr Quick Start Guide”, <https://cwiki.apache.org/confluence/display/WHIRR/Quick+Start+Guide>, Apache Whirr Wiki, June 2011
- [15] “MapReduce: Simplified Data Processing on Large Clusters”, Jeffrey Dean and Sanjay Ghemawa, Google Labs, pp. 137–150, OSDI 2004
- [16] GNU Coord, <http://www.coordguru.com/>
- [17] “Integrated Information Systems Architecture in e-Business”, Jongwook Woo, Dong-Yon Kim, Wonhong Cho, MinSeok Jang, The 2007 international Conference on e-Learning, e-Business, Enterprise Information Systems, e-Government, and Outsourcing, Las Vegas (June 26-29, 2007)
- [18] “[Hbase - non SQL Database, Performances Evaluation](#)”, Dorin Carstoiu, Elena Lepadatu, Mihai Gaspar, International Journal of Advancements in Computing Technology, Volume 2, Number 5, December 2010
- [19] “Distributed Storage of Large Scale Multidimensional Electroencephalogram Data using Hadoop and HBase”, Haimonti Dutta, Alex Kamil, Manoj Pooleery, Simha Sethumadhavan and John Demme, [Book Chapter in Grid and Cloud Database Management](#), Editors Sandro Fiore and Giovanni Aloisio, Springer, 2011