SAURIDA: CLOUD COMPUTING BASED
Data Mining System in Telecommunication Industry

Qing Ke, Bin Wu
School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China

Yuxiao Dong, Lei Qin
School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China

Keywords: Cloud Computing, Data Mining, Data Flow, Telecommunication.

Abstract: Telecommunication data mining has been often used as a background application to motivate many technical problems in data mining research. However, traditional mining algorithms face new challenges which are tremendous amount of data and high time and space complexity of algorithms. Recently, Map-Reduce parallel computing model has been emerging. In this paper, we combine data mining with Map-Reduce based cloud computing to meet the challenges and showcase our applied system named Saurida. As a full functionality system, we provide data flow oriented preprocessing utilities which achieve almost linear speedup and extensively support for user defined functions, and we also provide many data mining algorithms. More importantly, we elaborate several application scenarios as real-word requirements of telecom industry by employing a large volume of data obtained from telecom operator. And we validate our system has a good scalability, effectiveness and efficiency.

1 INTRODUCTION

Telecommunication data analysis has stimulated great interests in recent years. Typical application scenarios are customer churn prediction and customers' relationship management.

However, these analysis methods face new challenges. Firstly, the telecom industry generates and stores a tremendous amount of data. Secondly, many data mining algorithms have high time and space complexity.

Traditional business solutions of data mining are commercial database or data warehouse systems or commercial data mining tools. However, these systems or tools are low scalability and high cost. In research areas, Wang et.al (Wang, 2009) developed a working data mining system on real mobile communication data, but the system mainly focused on algorithms in research such as sequential patterns mining and community detection.

Recently, the Map-Reduce (Dean, 2004) computational model and its open-source implementation, Hadoop, are widely applied both in research and industry areas. The model mainly focuses on share-nothing parallelism, and its storage system focuses on scalability. These advantages are very suitable for telecom data mining.

In this paper, we combine data mining with Map-Reduce based cloud computing to meet the challenges and introduce our applied system, *Saurida*. The system is built on distributed cluster infrastructure as hardware and Hadoop distributed computing platform as fundamental software. As a full functionality system, we provide data preprocessing utilities, data mining algorithms. More importantly, we elaborate several application scenarios as real-word requirements of telecom industry by employing a large volume of data obtained from telecom operator, we validate our system from the view of scalability, effectiveness and efficiency. In summary, Saurida takes the following challenges as its destination as well as the contributions to this work:

- Data flow oriented and almost linear speedup of preprocessing.
- Extensive support for user defined functions.
- Nearly linear speedup of data mining algorithm.

Ke Q., Wu B., Dong Y. and Qin L..
SAURIDA: CLOUD COMPUTING BASED - Data Mining System in Telecommunication Industry.
DOI: 10.5220/0003387905160519
In *Proceedings of the 1st International Conference on Cloud Computing and Services Science* (CLOSER-2011), pages 516-519
ISBN: 978-989-8425-52-2

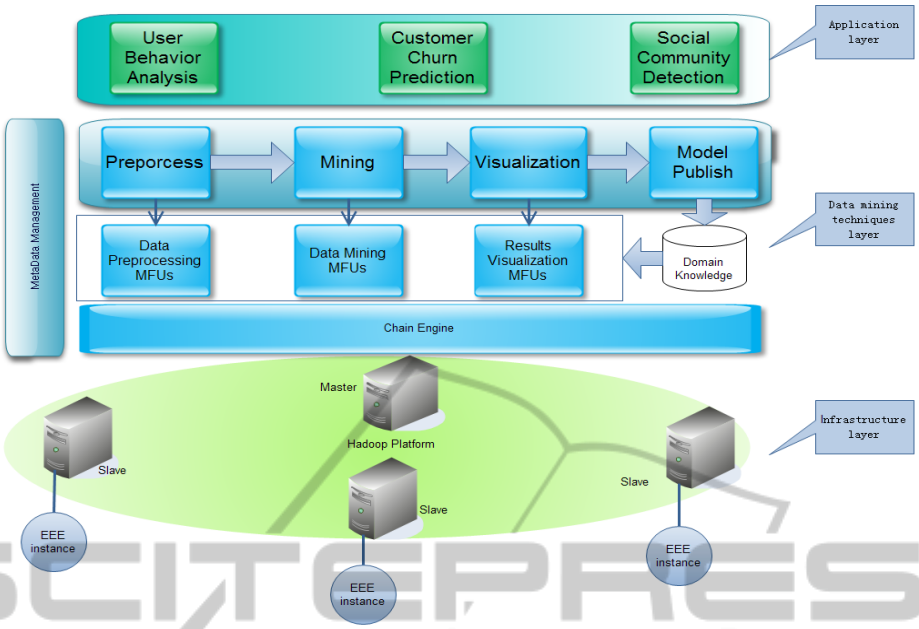

Figure 1: Saurida architecture.

- Real-word application in telecom industry.

The rest of this paper is structured as follows. Section 2 describes system architecture. Section 3 discusses implementation issues. In Section 4, we present several applications. Finally, we draw the conclusion and discuss future work.

2 SYSTEM ARCHITECTURE

The architecture of Saurida is depicted in Figure 1. As can be seen, the system consists of three layers. The functions and features of each layer are described as follows:

- Application layer implements business applications of telecom industry, such as user behavior analysis, customer churn prediction and social community detection.
- Data mining techniques layer completes functions mainly including preprocessing, data mining and results visualization, and a very important component named Chain Engine which is responsible for chaining the preprocessing utilities together and submitting to Hadoop, we will discuss it in detail in Section 3.2.

  Infrastructure layer consists of Hadoop platform where every slave node runs an Expression Evaluation Engine (EEE) instance which is the essential component to implement custom processing.

3 IMPLEMENTATION

We set up a cluster environment, composed of one master node and 21 computing nodes (Intel Xeon 2.50GHz×4, 8GB RAM, 250GByte×4 SATA II disk, Linux RH4 OS). The cluster is interconnected through 1000Mbps Switch. And deployed Hadoop platform version is 0.20.0.

3.1 Preprocessing Utilities

Our system provides many preprocessing utilities which are mainly categorized into operations on record and operations on attribute. All of them are implemented through running Map-Reduce jobs.

We complete a performance benchmark which processes a terabyte data by running some typical utilities on 128 nodes. Figure 2 depicts the running time. We can see that all of them complete duration 1,500 seconds except the Merge because it actually process total 2 terabyte data and transferring intermediate data from mapper to reducer also consumes some time. Figure 3 shows derived scalability on 32, 64 and 128 nodes. The experiment results indicate that the parallel data preprocessing has excellent scalability.

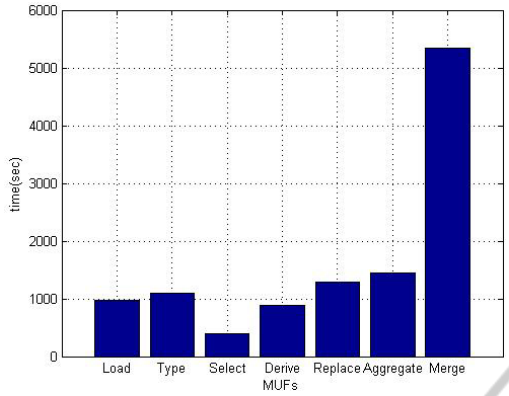

Figure 2: Benchmark results of preprocessing MFUs.

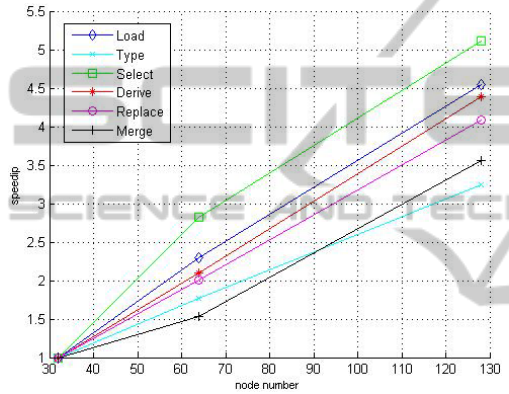

Figure 3: Scalability of preprocessing Utilities.

3.2 Chain Engine

We use native APIs provided by Hadoop, *ChainMapper* and *ChainReducer* to develop chain engine which is working as following steps:

1. Chaining all the user selected utilities together;
2. Changing the logical data flow into Map-Reduce jobs;
3. Submitting the jobs to Hadoop;

The immediate benefit of chained pattern is a dramatic reduction in disk I/O because the output of the first utility becomes the input of the second one, and so on until the last one, all the intermediate results do not need to flush to disk. So the execution time of the data flow dramatic reduced.

3.3 EEE

To accommodate specialized data processing tasks, Saurida has extensive support for User Defined Functions (UDFs). Essentially all aspects of preprocessing utilities in Saurida including Select, Derive, Replace and so on can be customized through the use of UDFs.

When the Map-Reduce job is executing, all the UDFs are put into EEE instance. Every DataNode of the Hadoop cluster is running an instance, we can achieve and the engine will output the result of each UDF. EEE uses traditional Reverse Polish Notation algorithm to evaluate the expression.

4 APPLICATION SCENARIOS

4.1 Ad-hoc Query

In this Section, we describe a sample ad-hoc data analysis tasks. The SQL is:

```
SQL: Select ID, fee_A, fee_B,
       case when fee_A>100 then 1
            When fee_A>200 then 2
            else 0 as fee_A_interval
  from fee_info where fee_A>50;
```

Figure 4 is shown with the number of nodes increasing from 6, 9 to 17, the chain which process 12GB data has excellent scalability, that is to say, the speedup ratio increases nearly linearly with the number of nodes.

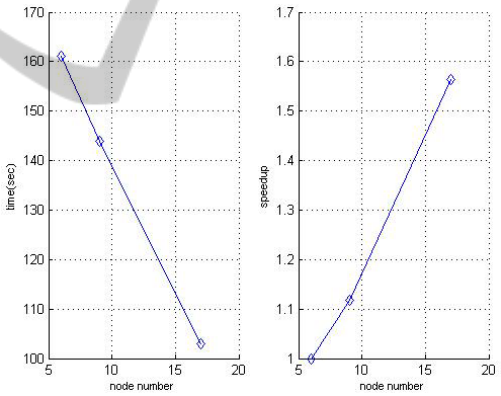

Figure 4: Running time and scalability of a sample ad-hoc query.

4.2 PCA

PCA (Wold, 1987) transforms a number of possibly correlated variables into a number of uncorrelated variables called principal components, commonly by an orthogonal transformation based on variance. PCA is very useful both in research and industry area. In many telecom data mining applications, the training data set may be as many as hundreds of feature items. However, many features are correlated, and these relevant features can be removed.

We run the parallel implementation of PCA on a real-world data set and on different number of nodes

to test performance and scalability. The data is 12 GB and contains 49 fields. We set the number of principal component to 10. The results are shown in Figure 5. And we can see that the PCA algorithm achieves good scalability.

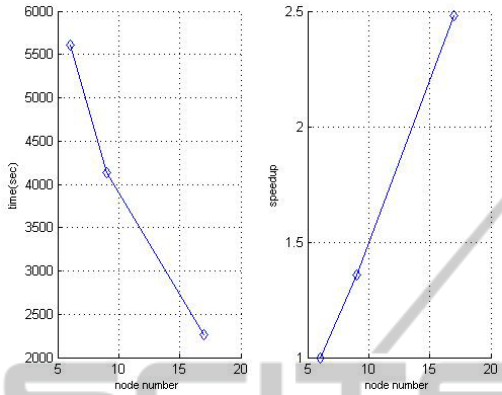

Figure 5: Running time and scalability of PCA.

4.3 NN

We run the parallel implementation of Feed-forward back-propagation neutral network (BP) (Williams, 1986) on a real-world data set and on different number of nodes to test performance and scalability. The data set is 14 GB, containing total 67 fields, we choose 55 fields of them as training attributes and the classification attribute contains 2 values. Figure 6 shows the results of running time on 6, 9 and 17 nodes and corresponding derived scalability. And the parallel NN algorithm also achieves nearly linear speedup.

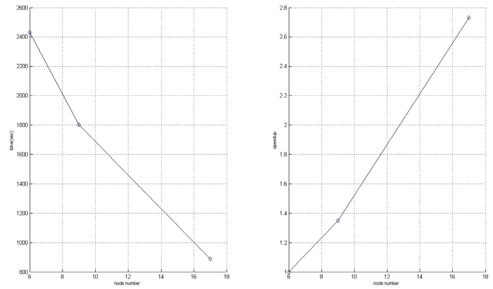

Figure 6: Running time and scalability of NN.

5 CONCLUSIONS AND FUTURE WORK

Motivated by recently increasing request for the capability of large scale data computing in telecommunications industry, in this paper, we introduce our system, *Saurida*, and demonstrate the system has advantages that open-source or commercial data mining tools do not have. These advantages include ability to process terabytes scale of data, high performance, linear speedup, cost-effective and custom processing. From the industrial view, we describe several application scenarios over large scale data as real-word requirements.

Nonetheless, *Saurida* is an experimental framework at present. Further development and improvement is needed at aspects such as functionality, performance and reliability to meet telecom industry requirements. Essentially, we hope our system can serve as a practical data mining system in telecom industry.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation of China (Grant No.60905025, 90924029, 61074128).

REFERENCES

Wold, S., Esbensen, K., Geladi, P., 1987. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 2, pp. 37-52.

T. Wang, B. Yang, J. Gao, D. Yang, S. Tang, H. Wu, K. Liu, and J. Pei, 2009. MobileMiner: A Real World Case Study of Data Mining in Mobile Communication. In *SIGMOD'09, Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data.*

Dean, J., Ghemawat, S., 2004. MapReduce: Simplified data processing on large clusters. In *OSDI '04, Sixth Symposium on Operating System Design and Implementation*.

Williams R. J., Rumelhart D. E., Hinton G. E., 1986. Learning representation by back-propagating errors. *Nature*, vol. 323, pp. 533–536.