MRMondrian: Scalable Multidimensional Anonymisation for Big Data Privacy Preservation

Abstract—Scalable data processing platforms built on cloud computing becomes increasingly attractive as infrastructure for supporting big data applications. But privacy concerns are one of the major obstacles to making use of public cloud platforms. Multidimensional anonymisation, a global-recoding generalisation scheme for privacy-preserving data publishing, has been a recent focus due to its capability of balancing data obfuscation and usability. Existing multidimensional anonymisation methods suffer from scalability problems when handling big data due to the impractical serial I/O cost. Given the recursive feature of multidimensional anonymisation, parallelisation is an ideal solution to scalability issues. However, it is still a challenge to use existing distributed and parallel paradigms directly for recursive computation. In this paper, we propose a scalable approach for big data multidimensional anonymisation based on MapReduce, a state-of-the-art data processing paradigm. Our basic idea is to partition a data set recursively into smaller partitions using MapReduce until all partitions can fit in the memory of a computing node. A tree indexing structure is proposed to achieve recursive computation. Moreover, we show the applicability of our approach to differential privacy. Experimental results on real-life data demonstrate that our approach can significantly improve the scalability of multidimensional anonymisation over existing methods.

CONCLUSIONS

In this paper, we have proposed a highly scalable and efficient approach named MRMondrian for multidimensional anonymisation over big data based on the MapReduce paradigm. Following the basic idea of dividing data sets into small data partitions to make them fit into the main memory of a single node, we have proposed to conduct iterative data partitioning in a parallel manner. Concrete data
partitioning is performed when all data partitions can fit into main memory. Then, each data partition is further divided recursively by the traditional serial Mondrian method on a single node. To support iterative data partitioning, a tree structure named PID-tree has been proposed to index data partitions for searching partition IDs. Coefficient of variation has been leveraged to select splitting attributes for a categorical attribute, while a scalable method based on the idea of the median of medians and the histogram technique has been proposed to find the median for a numerical attribute. We have also extended our approach to differential privacy.

SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 512 Mb

SOFTWARE REQUIREMENTS:

- Operating system : Windows 7/UBUNTU.
- Coding Language : Java 1.7, Hadoop 0.8.1
- IDE : Eclipse
- Database : MYSQL

REFERENCES
