A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud

ABSTRACT:

A large number of cloud services require users to share private data like electronic health records for data analysis or mining, bringing privacy concerns. Anonymizing data sets via generalization to satisfy certain privacy requirements such as k-anonymity is a widely used category of privacy preserving techniques. At present, the scale of data in many cloud applications increases tremendously in accordance with the Big Data trend, thereby making it a challenge for commonly used software tools to capture, manage, and process such large-scale data within a tolerable elapsed time. As a result, it is a challenge for existing anonymization approaches to achieve privacy preservation on privacy-sensitive large-scale data sets due to their insufficiency of scalability. In this paper, we propose a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the MapReduce framework on cloud. In both phases of our approach, we deliberately design a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way. Experimental evaluation results demonstrate that with our approach, the scalability and efficiency of TDS can be significantly improved over existing approaches.
EXISTING SYSTEM:

A widely adopted parallel data processing framework, to address the scalability problem of the top-down specialization (TDS) approach for large-scale data anonymization. The TDS approach, offering a good tradeoff between data utility and data consistency, is widely applied for data anonymization. Most TDS algorithms are centralized, resulting in their inadequacy in handling largescale data sets. Although some distributed algorithms have been proposed, they mainly focus on secure anonymization of data sets from multiple parties, rather than the scalability aspect.

PROBLEM DEFINITION:

- The MapReduce computation paradigm still a challenge to design proper MapReduce jobs for TDS.

PROPOSED SYSTEM:

- In this paper, we propose a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the MapReduce framework on cloud.
- In both phases of our approach, we deliberately design a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way.
ADVANTAGES OF PROPOSED SYSTEM:

- Accomplish the specializations in a highly scalable fashion.
- Gain high scalability.
- Significantly improve the scalability and efficiency of TDS for data anonymization over existing approaches.

SYSTEM ARCHITECTURE:
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 XP/7.
- Coding Language : JAVA/J2EE
- IDE : Netbeans 7.4
- Database : MYSQL