ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Optimized Balanced Scheduling Based Data Anonymization Using Two-Phase TDS Approach on Cloud

Vidyavati Bannatti
PG Scholar,
Computer Science and Engineering,
Bheema institute of Technology and Science.

K.Arjun
Assistant Professor,
Computer Science and Engineering,
Bheema institute of Technology and Science.

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.

1.1 Motivation:

It explains a highly scalable two-phase TDS approach for data anonymization based on Map Reduce on cloud. To make full use of the parallel capability of Map Reduce on cloud, specializations required in an anonymization process are split into two phases. In the first one, original data sets are partitioned into a group of smaller data sets, and these data sets are anonym zed in parallel, producing intermediate results. In the second one, the intermediate results are integrated into one, and further anonym zed to achieve consistent k-anonymous data sets. We leverage Map Reduce to accomplish the concrete computation in both phases. A group of Map Reduce jobs is deliberately designed and coordinated to perform specializations on data sets collaboratively. We evaluate our approach by conducting experiments on real-world data sets. Experimental results demonstrate that with our approach, the scalability and efficiency of TDS can be improved significantly over existing approaches.

1.2 Problem Definition:

We analyze the scalability problem of existing TDS approaches when handling large-scale data sets on cloud. The centralized TDS approaches in exploits the data structure TIPS to improve the scalability and efficiency by indexing anonymous data records and retaining statistical information in TIPS. The data structure speeds up the specialization process because indexing structure avoids frequently scanning entire data sets and storing statistical results circumvents recomputation overheads. On the other hand, the amount of metadata retained to maintain the statistical information and linkage information of record partitions is relatively large compared with data sets themselves, thereby consuming considerable memory. Moreover, the overheads incurred by maintaining the linkage structure and updating the statistic information will be huge when date sets become large. Hence, centralized approaches probably suffer from low efficiency and scalability when handling large-scale data sets. There is an assumption that all data processed should fit in memory for the centralized approaches . Unfortunately, this assumption often fails to hold in most data-intensive cloud applications nowadays.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

In cloud environments, computation is provisioned in the form of virtual machines (VMs). Usually, cloud compute services offer several flavors of VMs. As a result, the centralized approaches are difficult in handling large-scale data sets well on cloud using just one single VM even if the VM has the highest computation and storage capability. A distributed TDS approach is proposed to address the distributed anonymization problem which mainly concerns privacy protection against other parties, rather than scalability issues. Further, the approach only employs information gain, rather than its combination with privacy loss, as the search metric when determining the best specializations.

1.3.Context Diagram Of Project:

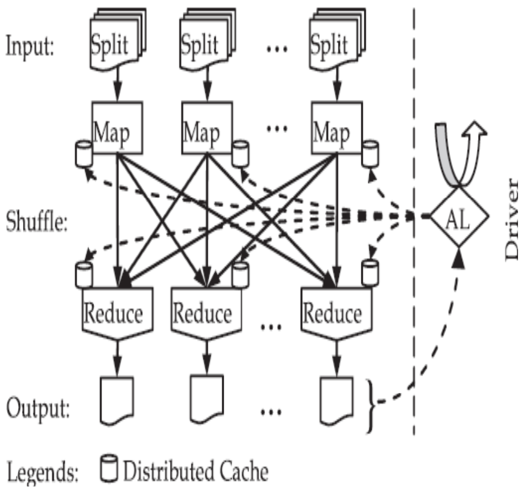


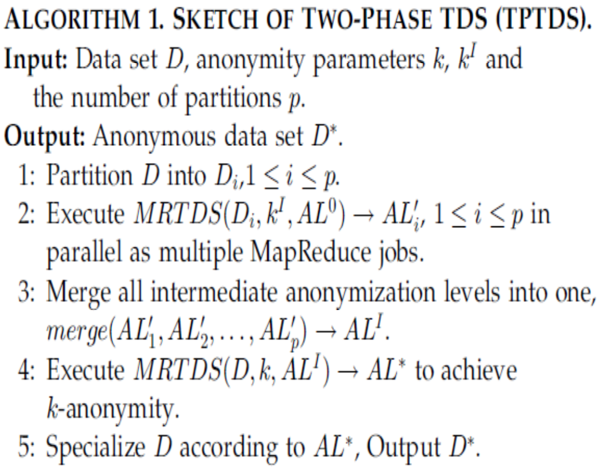
Fig .1: Context Diagram of Execution framework overview of MRTDS.

TPTDS approach to conduct the computation required in TDS in a highly scalable and efficient fashion. The two phases of our approach are based on the two levels of parallelization provisioned by Map Reduce on cloud. Basically, Map Reduce on cloud has two levels of parallelization, i.e., job level and task level. Job level parallelization means that multiple Map Reduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, Map Reduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, for example, Amazon Elastic Map Reduce service. Task level parallelization refers to that multiple Mapper/reducer tasks in a Map Reduce job are executed simultaneously over data splits.

To achieve High scalability, we parallelizing multiple jobs on data partitions in the first phase, but the resultant anonymization levels are not identical. To obtain finally consistent anonymous data sets, the second phase is necessary to integrate the intermediate results and further anonymize entire data sets. Details are formulated as follows. Then, we run a subroutine over each of the partitioned data sets in parallel to make full use of the job level parallelization of MapReduce. The subroutine is a MapReduce version of centralized TDS (MRTDS) which concretely conducts the computation required in TPTDS. MRTDS anonymizes data partitions to generate intermediate anonymization levels. An intermediate anonymization level means that further specialization can be performed without violating k-anonymity. MRTDS only leverages the task level Parallelization of MapReduce.

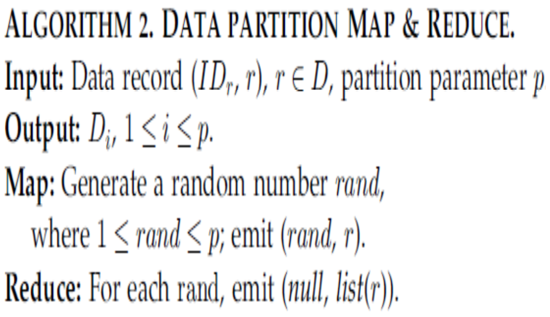
2.Algorithms And Flowcharts

Tptds Algorithm



```
ALGORITHM 1. SKETCH OF TWO-PHASE TDS (TPTDS).
Input: Data set $D$, anonymity parameters $k$, $k^I$ and
    the number of partitions $p$.
Output: Anonymous data set $D^*$.
  1: Partition $D$ into $D_i, 1 \le i \le p$.
  2: Execute $MRTDS(D_i, k^I, AL^0) \rightarrow AL'_i$, $1 \le i \le p$ in
     parallel as multiple MapReduce jobs.
  3: Merge all intermediate anonymization levels into one,
     $merge(AL'_1, AL'_2, \ldots, AL'_p) \rightarrow AL^I$.
  4: Execute $MRTDS(D, k, AL^I) \rightarrow AL^*$ to achieve
     $k$-anonymity.
  5: Specialize $D$ according to $AL^*$, Output $D^*$.
```

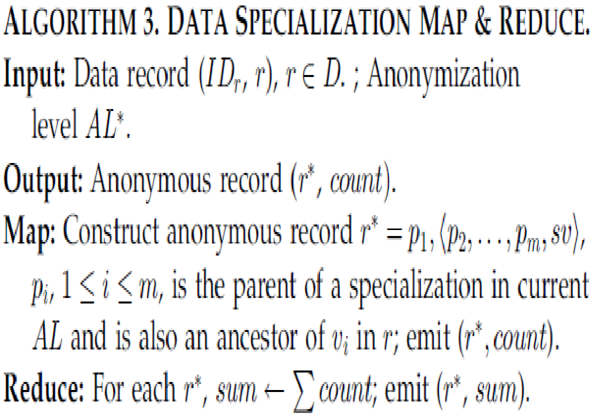
Data Partition Algorithm



```
ALGORITHM 2. DATA PARTITION MAP & REDUCE.
Input: Data record $(ID_r, r)$, $r \in D$, partition parameter $p$
Output: $D_i$, $1 \le i \le p$.
Map: Generate a random number $rand$,
    where $1 \le rand \le p$; emit $(rand, r)$.
Reduce: For each rand, emit $(null, list(r))$.
```

ISSN No: 2348-4845
International Journal & Magazine of Engineering,
Technology, Management and Research
A Peer Reviewed Open Access International Journal

Data Specialization Algorithm



```
ALGORITHM 3. DATA SPECIALIZATION MAP & REDUCE.
Input: Data record $(ID_r, r)$, $r \in D$. ; Anonymization
    level $AL^*$.
Output: Anonymous record $(r^*, count)$.
Map: Construct anonymous record $r^* = p_1, \langle p_2, \ldots, p_m, sv \rangle$,
    $p_i, 1 \leq i \leq m$, is the parent of a specialization in current
    $AL$ and is also an ancestor of $v_i$ in $r$; emit $(r^*, count)$.
Reduce: For each $r^*$, $sum \leftarrow \sum count$; emit $(r^*, sum)$.
```

3.Modules:

* Data Partition
* Anonymization
* Merging
* Specialization
* Obs

Modules Description:

Data Partition:

» In this module the data partition is performed on the cloud.
» Here we collect the large no of data sets.
» We are split the large into small data sets.
» Then we provides the random no for each data sets.

Anonymization:

» After geting the individual data sets we apply the anonymization.
» The anonymization means hide or remove the sensitive field in data sets.
» Then we get the intermediate result for the small data sets
» The intermediate results are used for the specialization process.
» All intermediate anonymization levels are merged into one in the second phase. The merging of anonymization levels is completed by merging cuts. To ensure that the merged intermediate anonymization level ALI never violates privacy requirements, the more general one is selected as the merged one

Merging:

» The intermediate result of the several small data sets are merged here.
» The MRTDS driver is used to organizes the small intermediate result
» For merging, the merged data sets are collected on cloud.
» The merging result is again applied in anonymization called specialization.

Specialization:

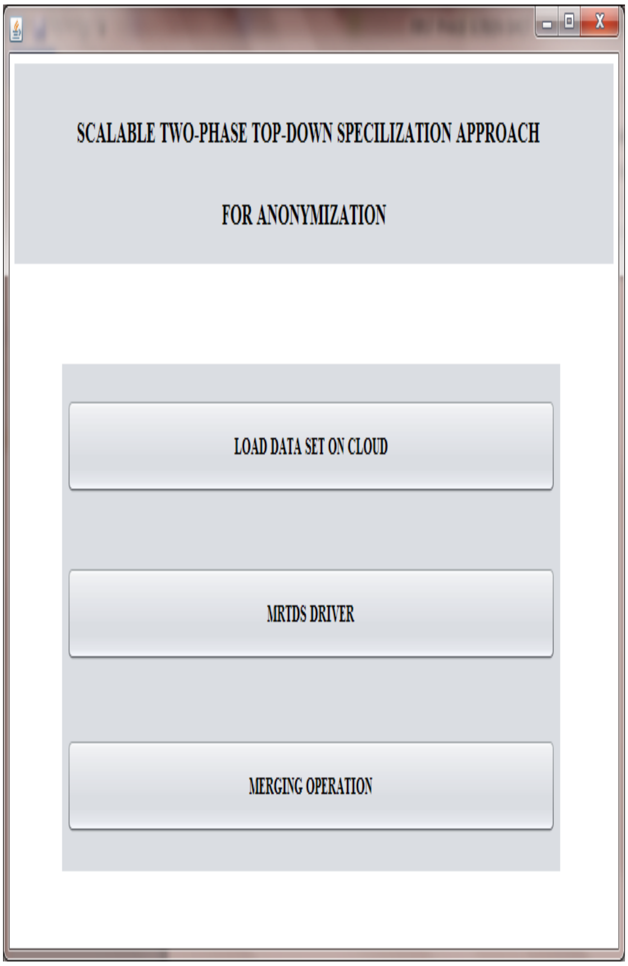
» After geting the intermediate result those results are merged into one.
» Then we again applies the anonymization on the merged data it called specialization.
» Here we are using the two kinds of jobs such as IGPL UPDATE AND IGPL INITIALIZATION.
» The jobs are organized by web using the driver.

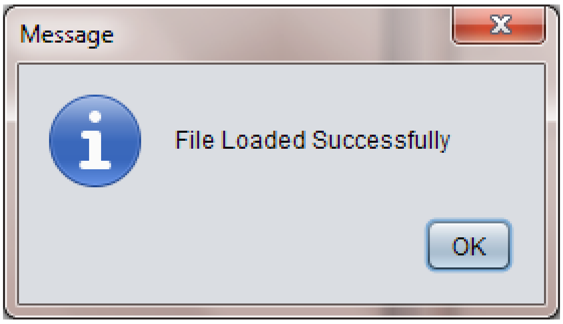
Obs:

» The OBS called optimized balancing scheduling.
» Here we focus on the two kinds of the scheduling called time and size.
» Here data sets are split in to the specified size and applied anonymization on specified time.
» The OBS approach is to provide the high ability on handles the large data sets.

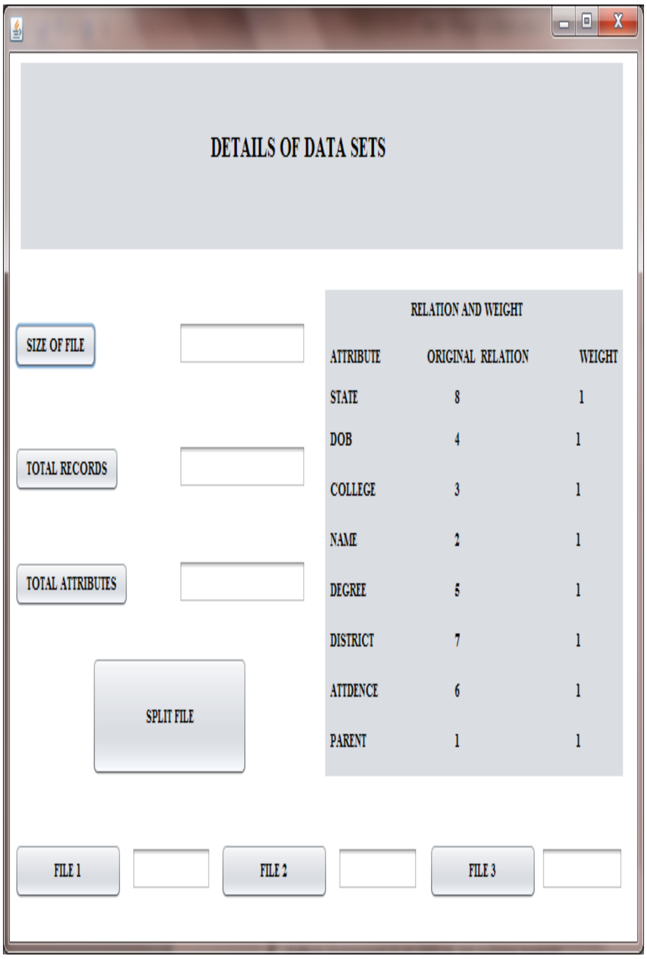
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

4.Simulation Results:


SCALABLE TWO-PHASE TOP-DOWN SPECILIZATION APPROACH FOR ANONYMIZATION

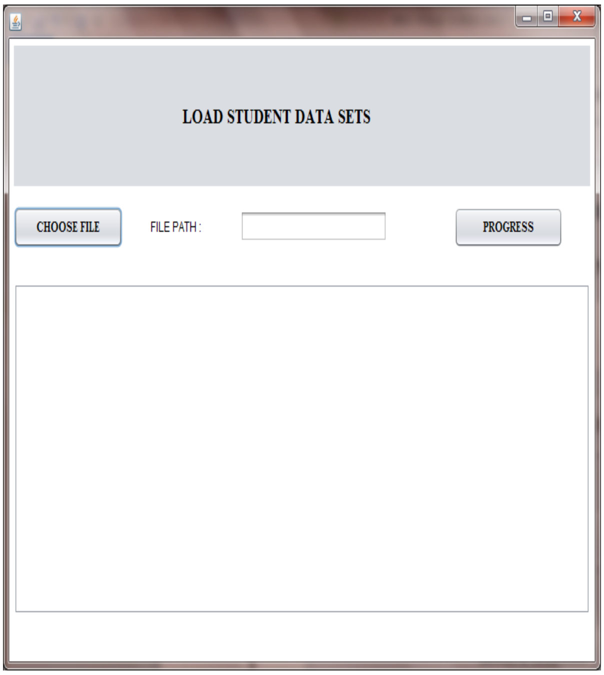


File Loaded Successfully

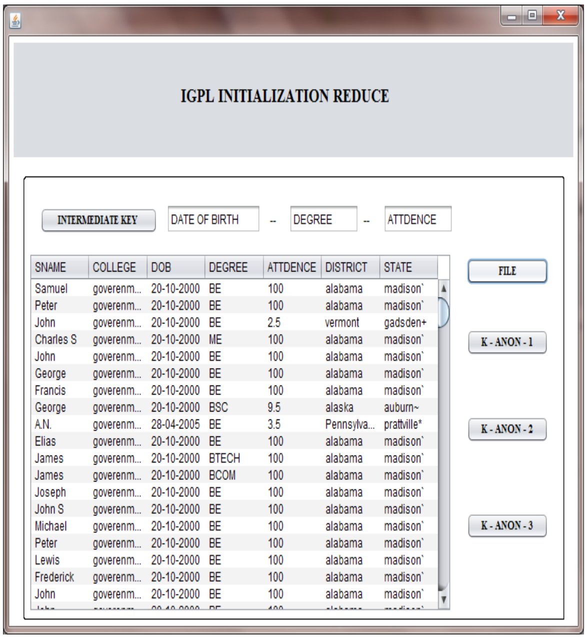

DETAILS OF DATA SETS

RELATION AND WEIGHT

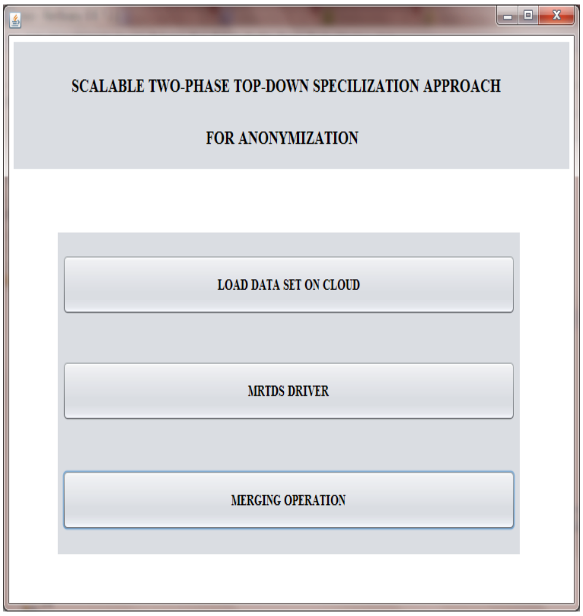
ATTRIBUTE	ORIGINAL RELATION	WEIGHT
STATE	8	1
DOB	4	1
COLLEGE	3	1
NAME	2	1
DEGREE	5	1
DISTRICT	7	1
ATTDENCE	6	1
PARENT	1	1


LOAD STUDENT DATA SETS

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal


IGPL INITIALIZATION REDUCE

SNAME	COLLEGE	DOB	DEGREE	ATTDENCE	DISTRICT	STATE
Samuel	goverenm...	20-10-2000	BE	100	alabama	madison`
Peter	goverenm...	20-10-2000	BE	100	alabama	madison`
John	goverenm...	20-10-2000	BE	2.5	vermont	gadsden+
Charles S	goverenm...	20-10-2000	ME	100	alabama	madison`
John	goverenm...	20-10-2000	BE	100	alabama	madison`
George	goverenm...	20-10-2000	BE	100	alabama	madison`
Francis	goverenm...	20-10-2000	BE	100	alabama	madison`
George	goverenm...	20-10-2000	BSC	9.5	alaska	auburn~
A.N.	goverenm...	28-04-2005	BE	3.5	Pennsylva...	prattville*
Elias	goverenm...	20-10-2000	BE	100	alabama	madison`
James	goverenm...	20-10-2000	BTECH	100	alabama	madison`
James	goverenm...	20-10-2000	BCOM	100	alabama	madison`
Joseph	goverenm...	20-10-2000	BE	100	alabama	madison`
John S	goverenm...	20-10-2000	BE	100	alabama	madison`
Michael	goverenm...	20-10-2000	BE	100	alabama	madison`
Peter	goverenm...	20-10-2000	BE	100	alabama	madison`
Lewis	goverenm...	20-10-2000	BE	100	alabama	madison`
Frederick	goverenm...	20-10-2000	BE	100	alabama	madison`
John	goverenm...	20-10-2000	BE	100	alabama	madison`


SCALABLE TWO-PHASE TOP-DOWN SPECILIZATION APPROACH

FOR ANONYMIZATION

CONCLUSION & FUTURE ENHANCEMENT:

Here, we have investigated the scalability problem of large-scale data anonymization by TDS and proposed a highly scalable two-phase TDS approach using Map Reduce on cloud. Data sets are partitioned and anonym zed in parallel in the first phase, producing intermediate results. Then, the intermediate results are merged and further anonym zed to produce consistent k-anonymous data sets in the second phase. We have creatively applied Map Reduce on cloud to data anonymization and deliberately designed a group of innovative Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way.

Experimental results on real-world data sets have demonstrated that with our approach, the scalability and efficiency of TDS are improved significantly over existing approaches. In cloud environment, the privacy preservation for data analysis, share and mining is a challenging research issue due to increasingly larger volumes of data sets, thereby requiring intensive investigation. We will investigate the adoption of our approach to the bottom-up generalization algorithms for data anonymization.

Future Enhancement:

Based on the contributions herein, we plan to further explore the next step on scalable privacy preservation aware analysis and scheduling on large-scale data sets. Optimized balanced scheduling strategies are expected to be developed towards overall scalable privacy preservation aware data set scheduling.

REFERENCES:

[1] S. Chaudhuri, “What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud,” Proc. 31st Symp. Principles of Database Systems (PODS ’12), pp. 1-4, 2012.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “A View of Cloud Computing,” Comm. ACM, vol. 53, no. 4, pp. 50-58, 2010.

[3] L. Wang, J. Zhan, W. Shi, and Y. Liang, “In Cloud, Can Scientific Communities Benefit from the Economies of Scale?,” IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 2, pp.296-303, Feb. 2012.

[4] H. Takabi, J.B.D. Joshi, and G. Ahn, “Security and Privacy Challenges in Cloud Computing Envirnments,” IEEE Security and Privacy, vol. 8, no. 6, pp. 24-31, Nov. 2010.

[5] D. Zissis and D. Lekkas, “Addressing Cloud Computing Security Issues,” Future Generation.

[6] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, “A Privacy Leakage Upper-Bound Constraint Based Approach for Cost- Effective Privacy Preserving of Intermediate Data Sets in Cloud,” IEEE Trans. Parallel and Distributed Systems, to be published, 2012.

[7] L. Hsiao-Ying and W.G. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, 2012.

[8] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM, pp. 829-837, 2011.

[9] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "Gupt: Privacy Preserving Data Analysis Made Easy," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '12), pp. 349-360, 2012.

[10] Microsoft HealthVault, http://www.microsoft.com/health/ww/ products/Pages/healthvault.aspx, 2013.