SEDIC: PRIVACY-AWARE DATA INTENSIVE COMPUTING ON HYBRID CLOUDS

Kehuan Zhang, Xiaoyong Zhou, Yangyi Chen and XiaoFengWang
*Yaoping Ruan **

*Indiana University, * IBM T.J.Watson Research Center*

CCS 2011

Presented By Dong Yuan & Zhihui Deng

(2015210938 2015210926)

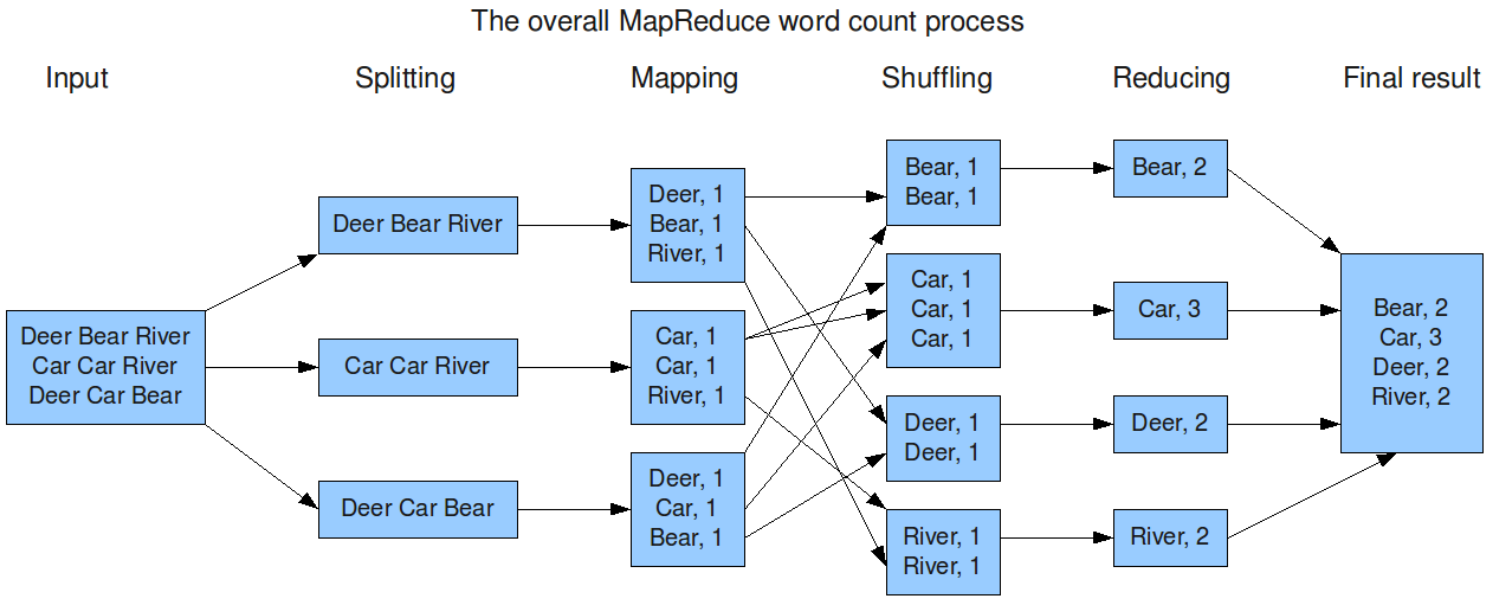
BACKGROUND

- Rapid growth of computing tasks [Map-Reduce Task]
- Popularity of the public cloud
- The processing of the sensitive data

- Demanding for Secure hybrid-cloud computing

MAP-REDUCE

- A simple mapreduce example:



OBJECTIVES

- High privacy assurance
- Moving workload to the public cloud when possible
- Scalability
- Limited inter-cloud data transfer
- Ease to use

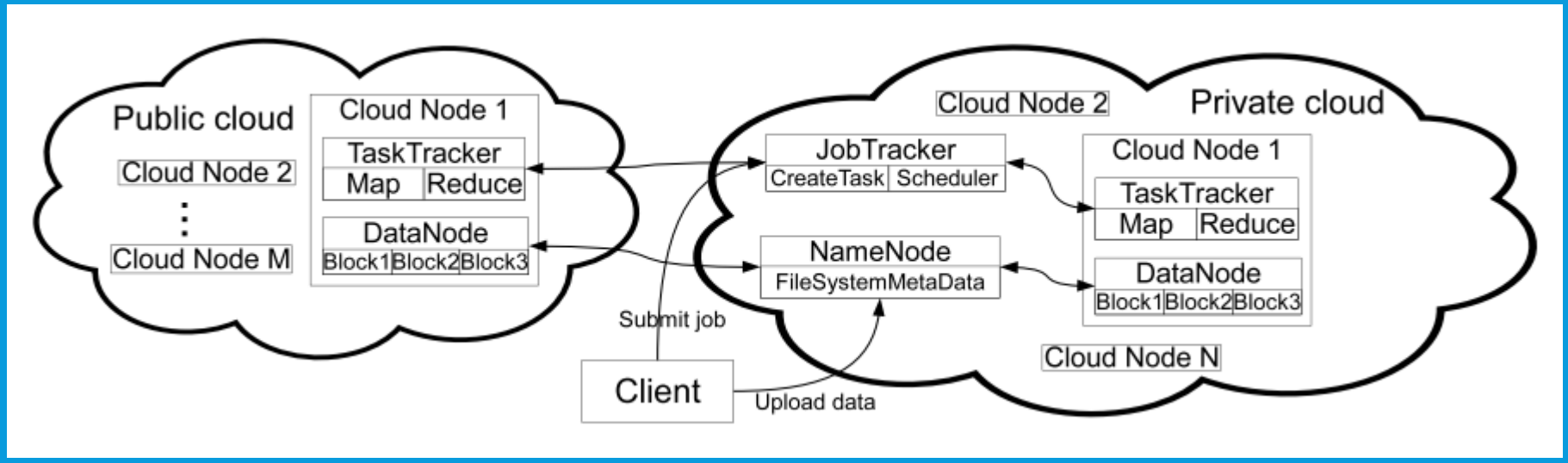
- Adversary Model:
  - An adversary who intends to acquire sensitive user information and has a full control of the public cloud

OVERVIEW

- Users:
  - Label sensitive data, which can be done through a data-tagging tool
  - Submit to Sedic labeled data and a MapReduce job.
- Sedic:
  - Analyze and transform the reduction structure of the job
  - Partition and replicate the data according to security labels
  - Create and schedule mappers across the public/private clouds
  - Combine the results on the public cloud and complete the reduction on the private cloud

THE EXECUTION FRAMEWORK

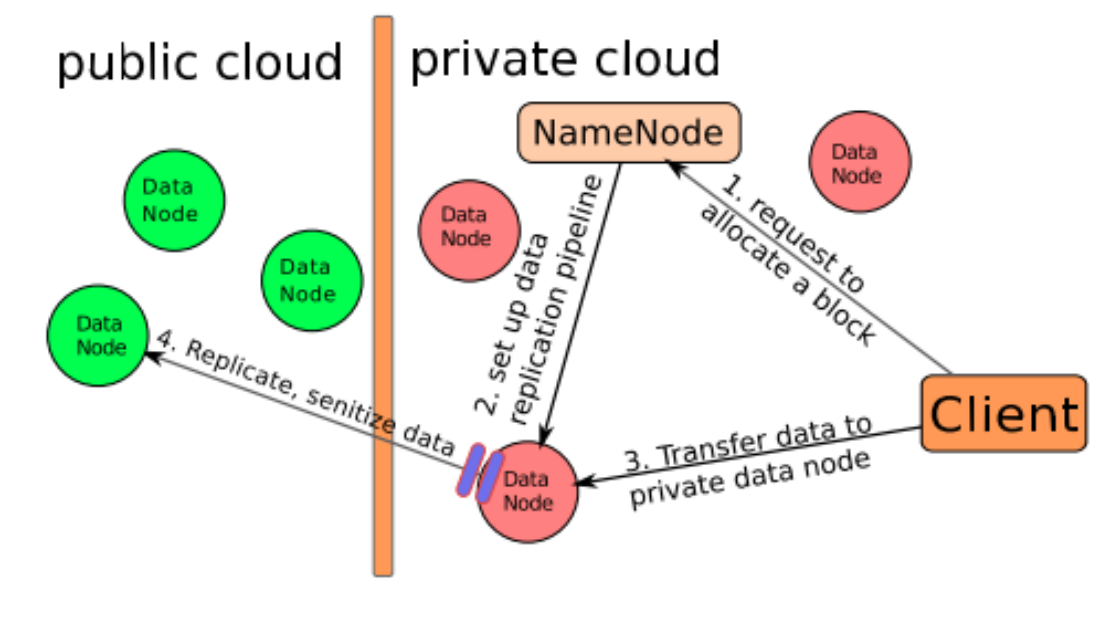
- Data Labeling and Replication
- Map Task Management
- Reduction Planning



DATA LABELING AND REPLICATION

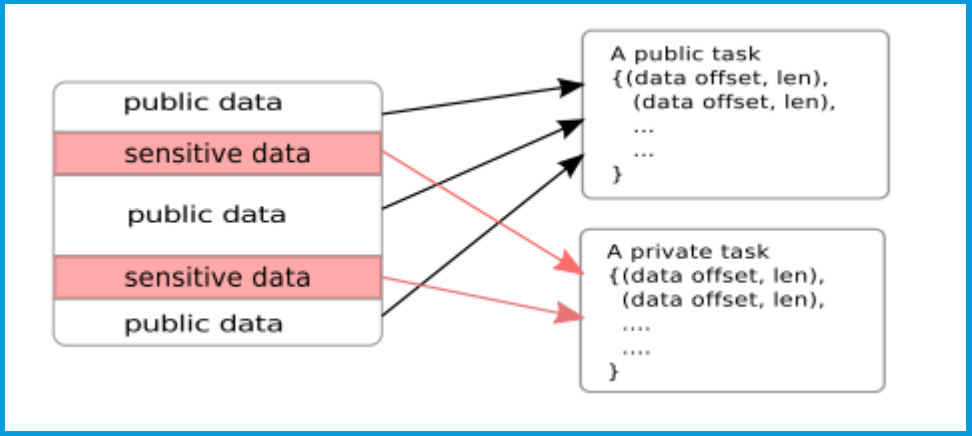
- Sensitive data labeling
  - A tool is provided
  - Label info: (<filename, offset, length>)
- Data uploading
  - Data block with sensitive stored in private node
  - Normal data block stored in public node
  - Sensitive label stored in private name node
- Data Replication
  - Replicate inside of the public cloud or private cloud

DATA LABELING AND REPLICATION



MAP TASK MANAGEMENT

- Task creation and submission
  - Create different for private and public block
  - For interleaved segaments, use one map task to handle all public data of the block and the other to process the sensitive one.
- Task scheduling
  - a sensitive task is always scheduled to a private datanode
- Task execution



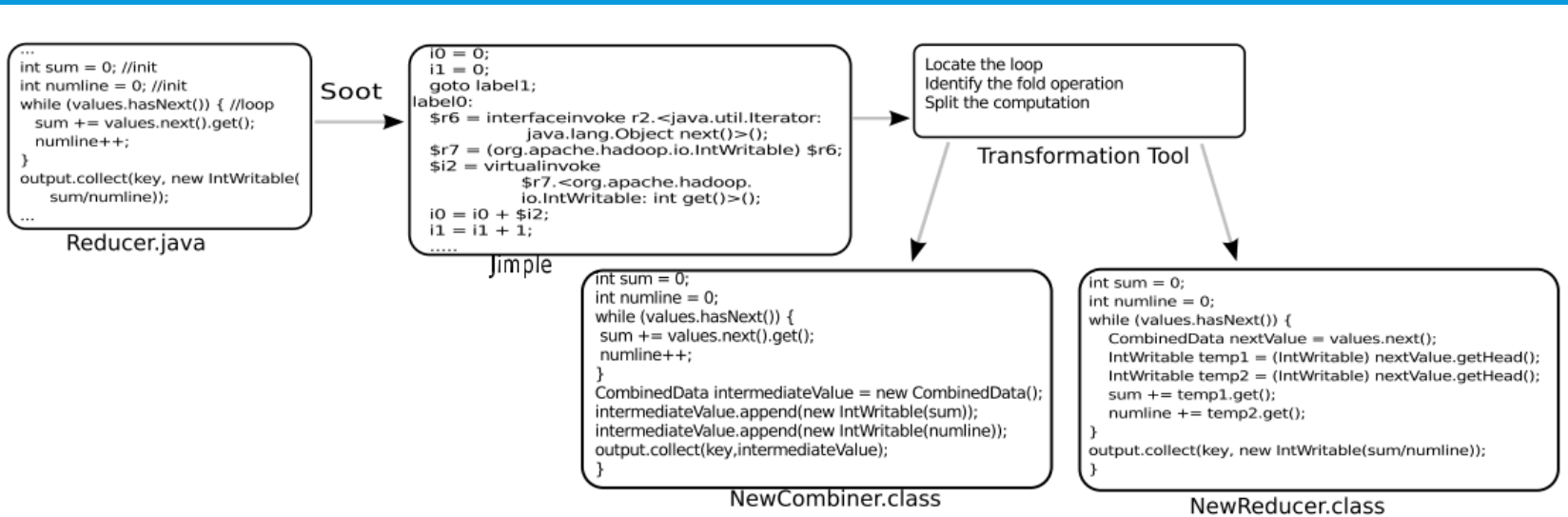
REDUCTION PLANNING

- All the task should be reduced in private node
- Sedic carefully plans the scheduling of map tasks to ensure that the total amount of the map output to be generated by the public cloud does not exceed an upper limit set by the user according

- A little bit stupid???
  - Go on…

AUTOMATIC REDUCER ANALYSIS AND CODE TRANSFORMATION

- Obersevation:
  - f([a1, a2, a3, a4, a5, a6]) = f([f([a1, a2]), f([a3, a4]), f([a5, a6])])
  - f([a1, a2, a3, a4, a5, a6]) = f([f([f([a1, a2]), f([a3, a4])]), f([a5, a6])])
- We can reduce 'a little' in public cloud



EVALUATION

- Experimental Setting
  - Build on FutureGrid:
    - 3 public nodes
    - 3 private nodes

Table 2: Descriptions of Hadoop Jobs

Job	Data set	Descriptions
Port Scan Detection	IDS data set	Find the TCP ports connected by each host
Traffic Statistics	IDS data set	Count the total amount of the traffic generated by each host (for detecting denial of service attacks)
Email Word Count	Spam data set	Count the total number of words in the spam dataset (for calculating Bayes probability)
Spam Keyword Count	Spam date set	Count the occurrences of each keyword on a given spam keyword list file(for calculating Bayes probability)
Grep	Twitter	Search for word patterns according to predefined regular expressions within the dataset, e.g., brand names and comment words such as awesome, wonderful, worst etc.

Table 3: Descriptions of Datasets

Name	Sensitive Data	Public Data	Size of Sensitive Data	Size of Public Data	Percentage of Sensitive Content
IDS data set	The tcpdump files for inside	The tcpdump files for outside	17GB	15GB	54%
Spam data set	The Enron Email Dataset	SPAM archive download from http://untroubled.org/spam/	1.3GB	0.8GB	62%
Twitter data set	Tweets from randomly chosen users who are assumed to prefer to protect their tweets	Tweets from other users	123MB	491MB	20%

RESULTS

- Baseline: run the whole job in private cloud

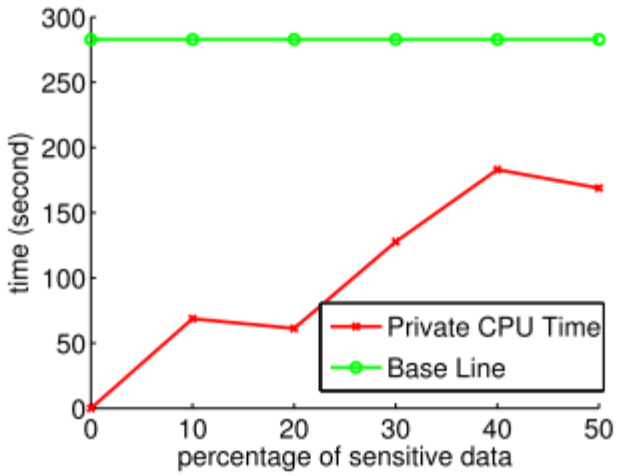


Figure 6: Performance vs. Sensitive Data Ratio

COMMENTS

- The reduce strategy is still rough
  - Can it integrate some confusion teches to reduce the comm. cost
- It's hard to support the iteratively map&reduce???
- It does not consider the communication latency to computation latency results.

Thank you!