Secure Data Processing in the Cloud by Managing Risks

Sharad Mehrotra
Dept. of Computer Science
University of California, Irvine

&

Anoop Singhal
Computer Security Division
National Institute of Standards and Technology
Gaithersburg MD

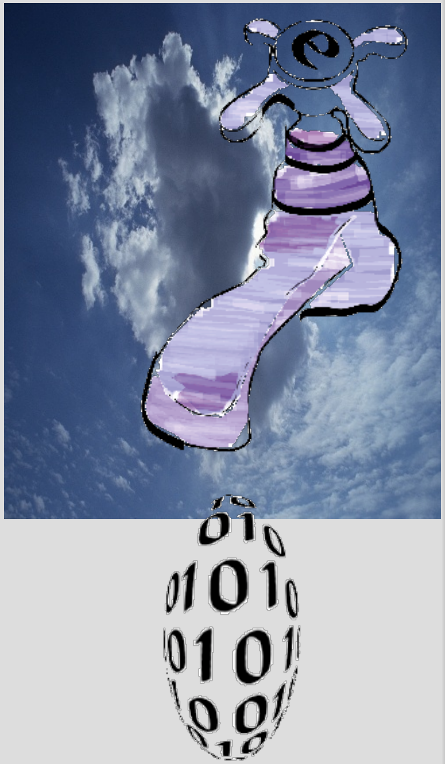
Outline

- Risk-based Approach to mgmt in cloud [Sharad]
  - Motivation

- Two main challenges:
  - Modeling Risks [Anoop]
    - State-of-the-art In enterprise networks
    - Thoughts on generalizing to cloud data
  - Given risks, data and workload partitioning problem [Sharad]
    - Some initial results

Cloud Computing

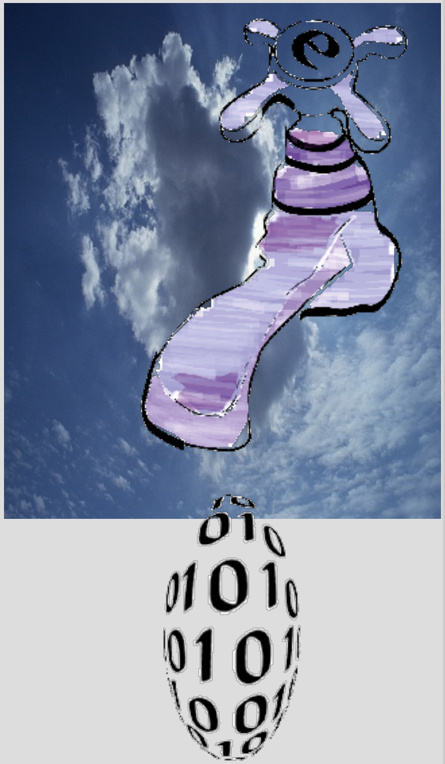
- **X as a service**, **where X is:**
  - Infrastructure, platforms, Software,
  - Storage, Application, test environments…
- **Characteristics:**
  - **Elastic** -- Use as much as your needs
  - Pay for only what you use
  - Don’t worry about failure
  - No system management headaches
    - E.g., loss of data due to failures
  - Hopefully cheaper due to economy of scale
    - Better control over IT investment



Utility model

Cloud Computing

- X as a service, where X is:
  - Infrastructure, platforms, Software,
  - Storage, Application, test environments…
- Characteristics:
  - Elastic -- Use as much as **your** needs
  - Pay for **only** what you use
  - Don’t **worry** about failure
  - No system management headaches
    - E.g., **loss of** data due to failures
  - Hopefully cheaper due to economy of scale
    - Better **control** over IT investment



Utility model

Loss of control

- Loss of Control: ***Inability to restrict (and monitor) other entities from accessing ones data*.**

- **Factors leading to loss of control**
  - Data resides in **shared systems** administration of which is not in owners control.
  - **Unknown applications and processes** share resources with your apps and data.
  - Data owners have **no control over CSP's internal** data security **personnel**, **policies or their enforcement**.
    - Insider attacks
    - Data mining attacks leading to information leakage

Implications of Loss of Control

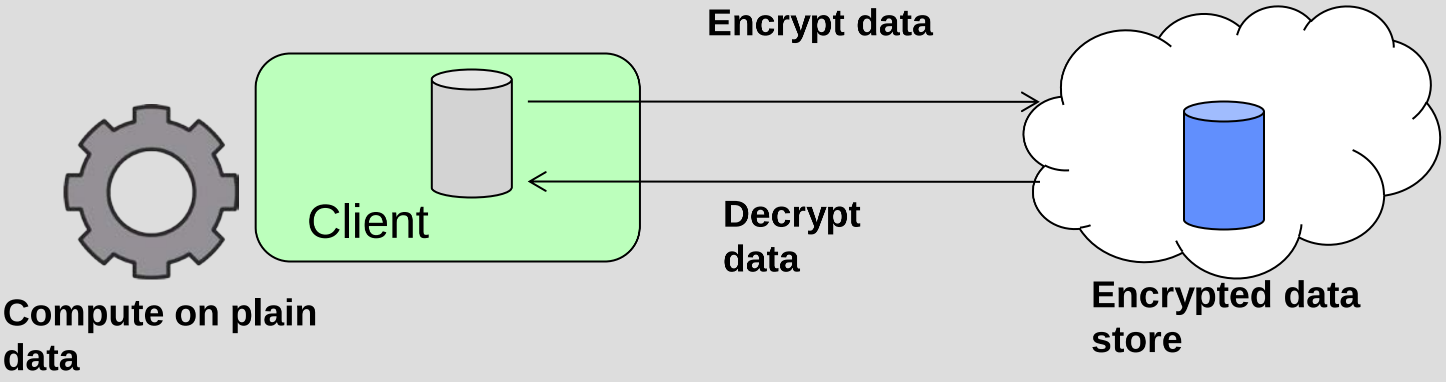
End Users

- Integrity
  - Will the CSP serve my data correctly?
  - Can my data get corrupted?
- Availability
  - Will I have access to my data and services at all times?
- Security
  - Will the CSP implement its own security policies appropriately?
- Privacy & confidentiality
  - Will sensitive data remain confidential?
  - Will my data be vulnerable to misuse? By other tenants? By the service provider?

What is the solution?

Encrypt sensitive data before uploading to the cloud

2 models of querying/Computing on encrypted data



Encrypt data

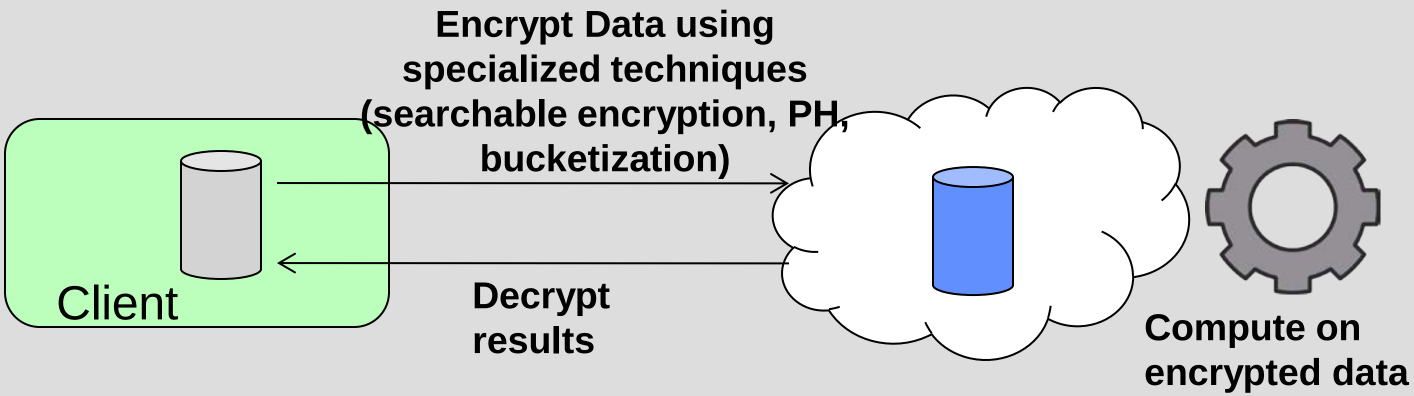
Client

Decrypt data

Compute on plain data

Encrypted data store

Most work done at the client; limited utility of cloud



Encrypt Data using specialized techniques (searchable encryption, PH, bucketization)

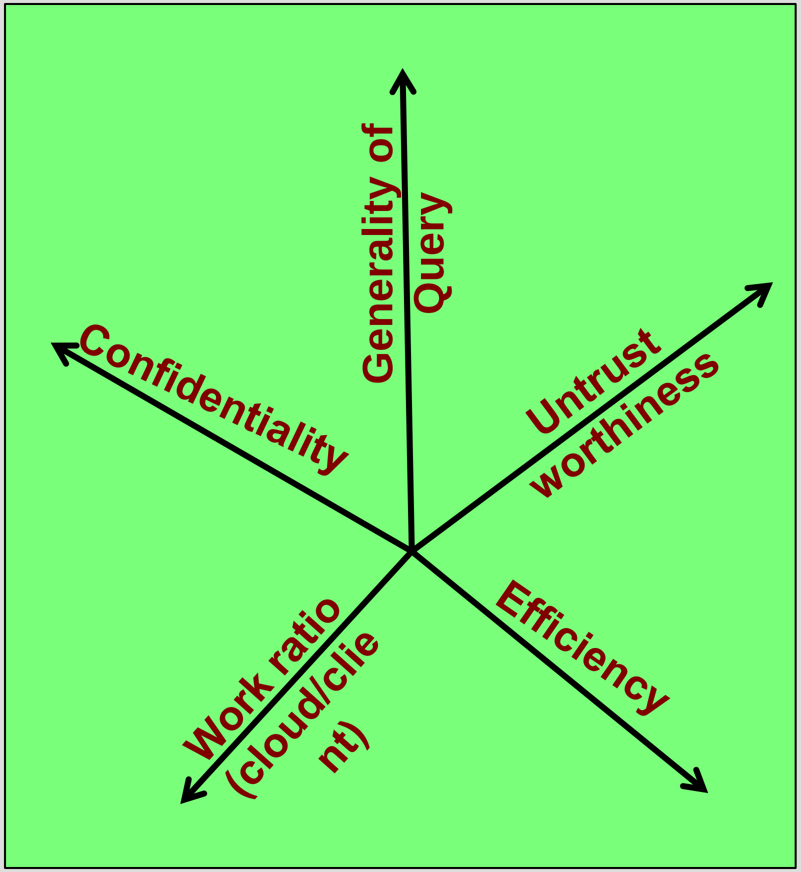
Client

Decrypt results

Compute on encrypted data

Can utilize techniques for computing on encrypted data (15 years worth of work)

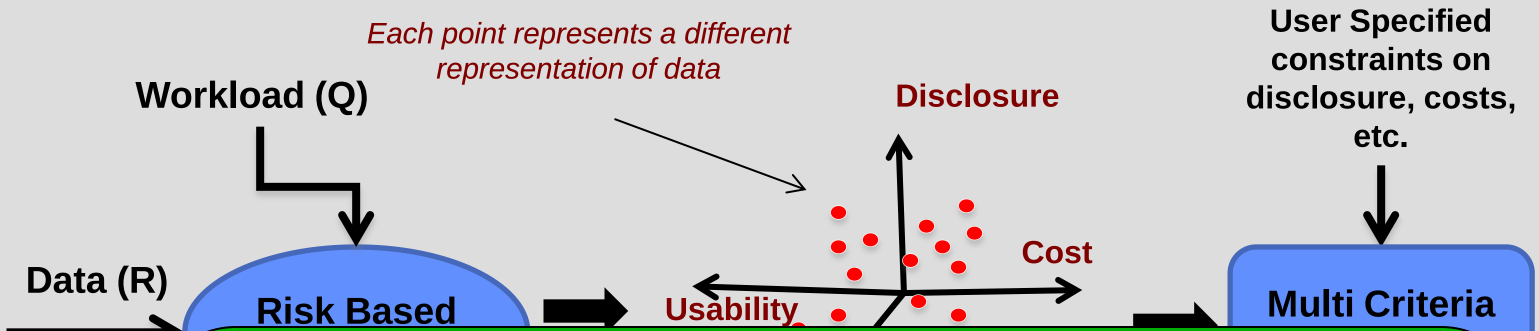
Search over Encrypted Data



- **Existing solutions**
  - can be characterized along multiple dimensions.
  - Represent points in the spectrum of possibilities
  - Explore different tradeoffs.
- **Example**:
  - Cloud as storage → poor work ratio
  - Homomorphic encryption → too inefficient to be practical
- **Mix-n-Match**
  - Many existing methods can be "mixed-n-matched" to provide practical solutions for specific problems

Computing on encrypted data remains an active research area!

Risk Based Data Processing in Clouds (Radicle Project)


Each point represents a different representation of data
Workload (Q)
Data (R)
Risk Based
Disclosure
Cost
Usability
User Specified constraints on disclosure, costs, etc.
Multi Criteria

Radicle exploits the hypothesis that 100% security is neither required nor achievable. Users may be willing to tolerate risks for improved performance, reduced costs, etc.

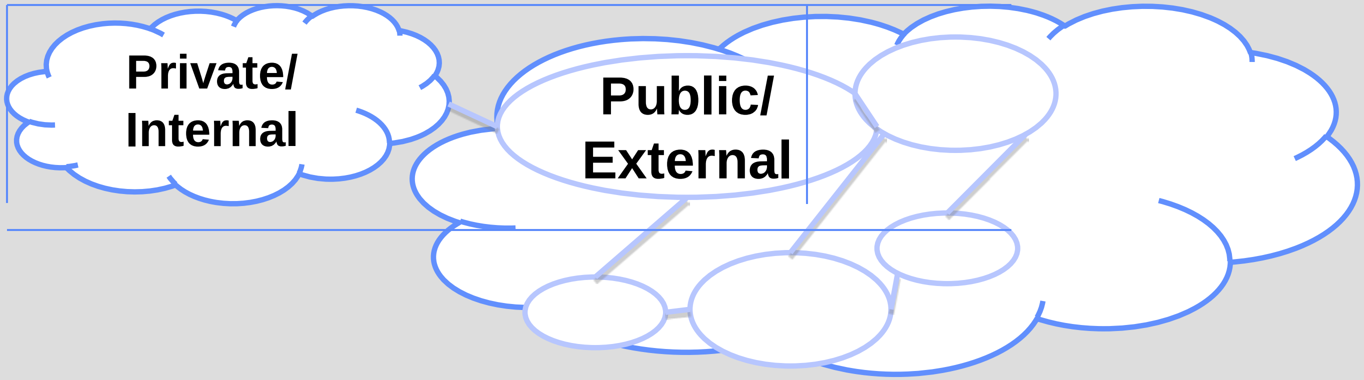
ions
) and
Plan

- **Supp**
  - Stro … orphic enc … tation.
- **Model exposure-risks of representation**
  - # sensitive data items exposed on public cloud, The representation of data on cloud-side, Duration of exposure, The trustworthiness of service-provider, ..
- **Partition computation and data between server and client**
  - such that owner can strike a desired balance between exposure risk, performance, usability and monetary costs incurred.

Design Spectrum

- **Input:**
  - **Data Model -** How is data represented?
    - Relational, Semi-structured, Key-Value Stores, Text…
  - **Workload Model -** What type of workload is given?
    - (Dynamic or Batch) SQL or HIVE Queries, MapReduce Jobs…
  - **Sensitivity Model -** How is sensitivity specified?
    - Attribute Level, Privacy Associations, View-Based…
- **Metrics:**
  - **Risk Model-** How is disclosure measured?
    - Number of exposed sensitive cells, Inference Exposure...
  - **Resource Allocation costs -** How is cost measured?
    - Based on elastic pricing model of public cloud providers
  - **Performance**
  - **…**
- **Solutions Space:**
  - **Data Representation Model -** How is data on public cloud partitioned and represented?
  - **Workload Partitioning Model -** How should workload be partitioned?
    - Inter-query Partitioning, Intra-query Partitioning…

Hybrid Clouds



- Hyb
  - Ir
  - E es
  - P he
    pl

Two Main Challenges:

1. *Modeling Risks*
2. *Data & Workload Partitioning*

- **Examples…**
  - http://www-01.ibm.com/software/tivoli/products/hybrid-cloud/
  - http://www.emc.com/campaign/global/hybridcloud/index.htm

Outline

- Risk-based Approach to data management in cloud [Sharad]
  - Motivation – focusing on why
- Modeling Risks [Anoop]
  - State-of-the-art In enterprise networks
  - Thoughts on generalizing to cloud data
- Given risks, example data and workload partitioning problem [Sharad]
  - Some initial results

Security Risk Modeling for Cloud Computing

Anoop Singhal

Computer Security Division
National Institute of Standards and Technology
Gaithersburg MD

Email: psinghal@nist.gov

Enterprise Systems Security Management

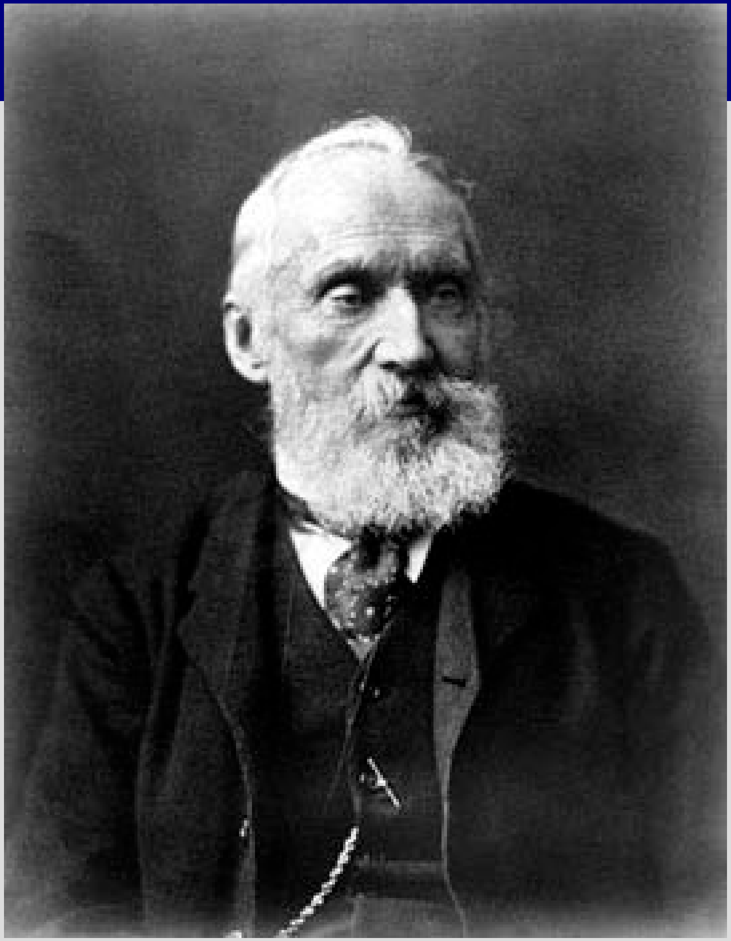
- Network Systems are getting large and complex
- Vulnerabilities in software are constantly discovered
- System Security Management is a challenging task
- Even a small system can have numerous attack paths

Enterprise System Security Management

- Currently, security management is more of an art and not a science
- System administrators operate by instinct and learned experience
- There is no objective way of measuring the security risk in a networked system
- “If I change this network configuration setting will my network become more or less secure?”

Challenges in Modeling Security Risk

- Typical issues addressed in the literature
  - How can a database server be secured from intruders?
  - How do I stop an ongoing intrusion?
- Better questions to ask:
  - How secure is the database server in a given network configuration?
  - How much security does a new configuration provide?
  - How can I plan on security investments so it provides a certain amount of security?
- For this we need a model for security risk



If you cannot measure (or model) it, you cannot improve it.

---Lord Kelvin

Challenges in Security Risk Metrics

- Metric for individual vulnerability exists
  - Impact, exploitability, temporal, environmental, etc.
  - E.g., the Common Vulnerability Scoring System (CVSS) v2 released on June 20, 2007[1]
- However, how to compose individual measures for the overall security of a network?
  - Our work focuses on this issue

[1] Common Vulnerability Scoring System (CVSS-SIG) v2, http://www.first.org/cvss/

Challenges in Security Risk Metrics

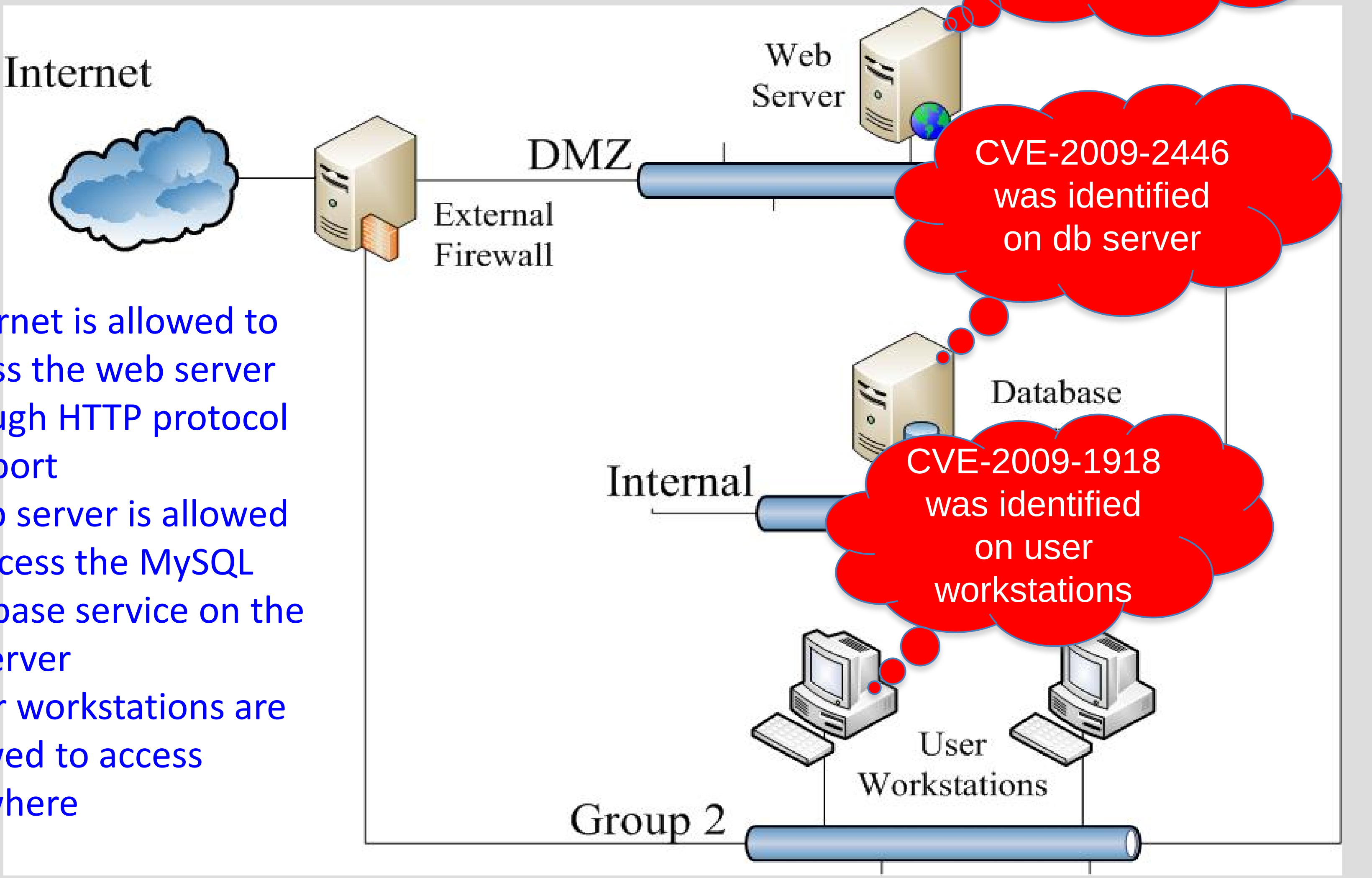
- Counting the number of vulnerabilities is not enough
  - Vulnerabilities have different importance
  - The scoring of a vulnerability is a challenge
    - Context of the Application
    - Configuration of the Application
- How to *compose* vulnerabilities for the overall security of a network system

What is an Attack Graph

- A model for
  - How an attacker can *combine* vulnerabilities to stage an attack such as a data breach
  - *Dependencies* among vulnerabilities

Example

- Internet is allowed to access the web server through HTTP protocol and port
- Web server is allowed to access the MySQL database service on the db server
- User workstations are allowed to access anywhere

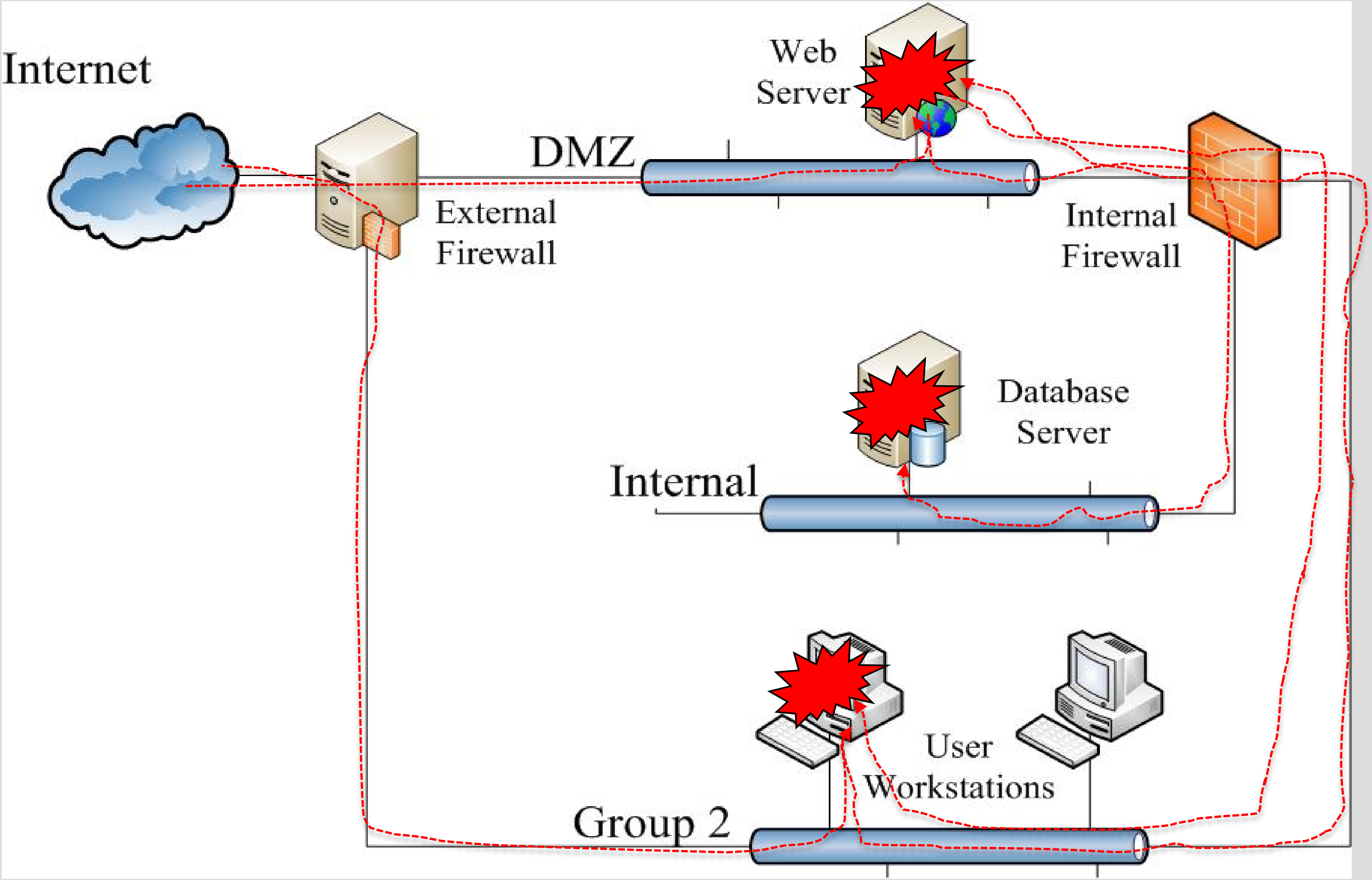


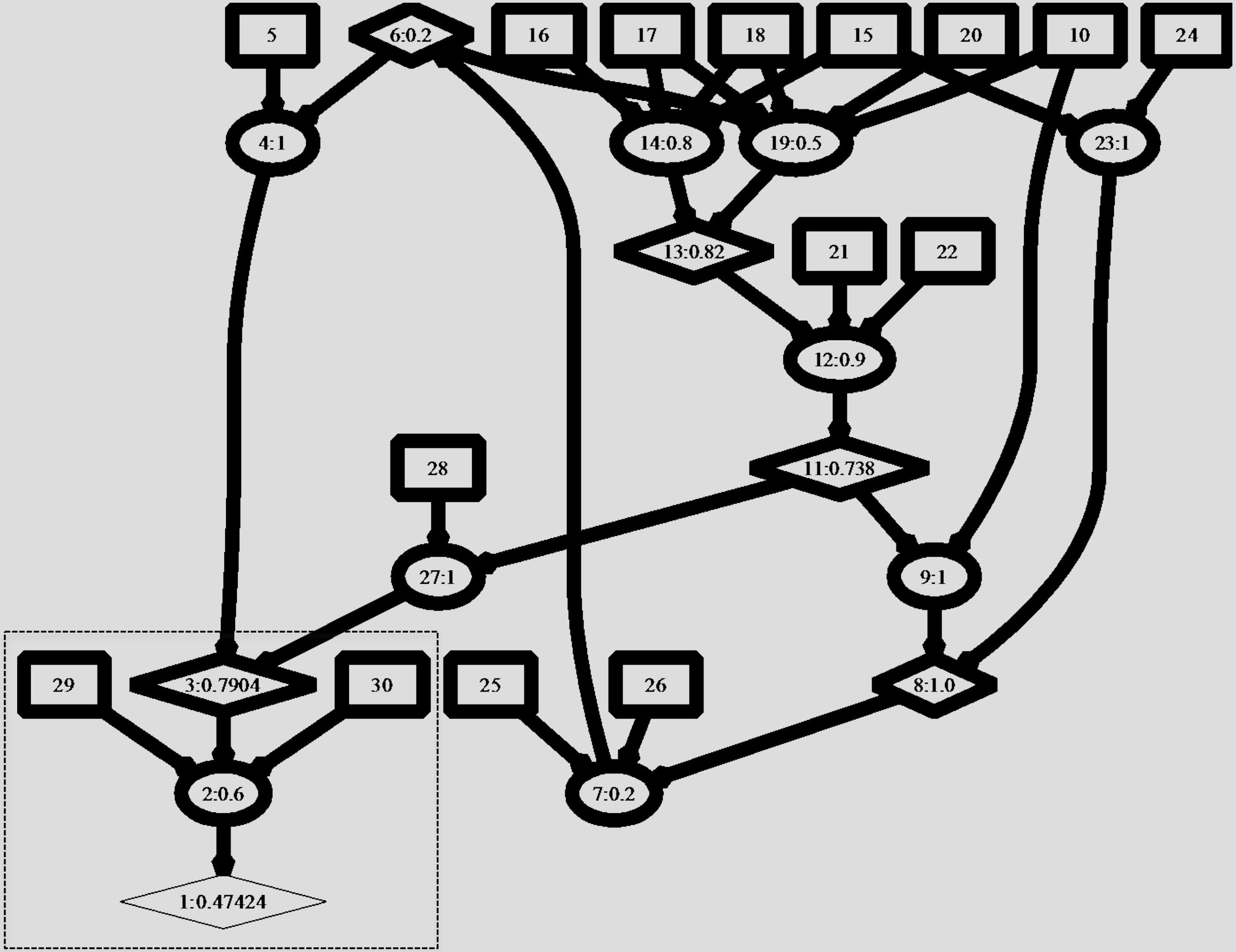
CVE-2006-3747 was identified on web server

CVE-2009-2446 was identified on db server

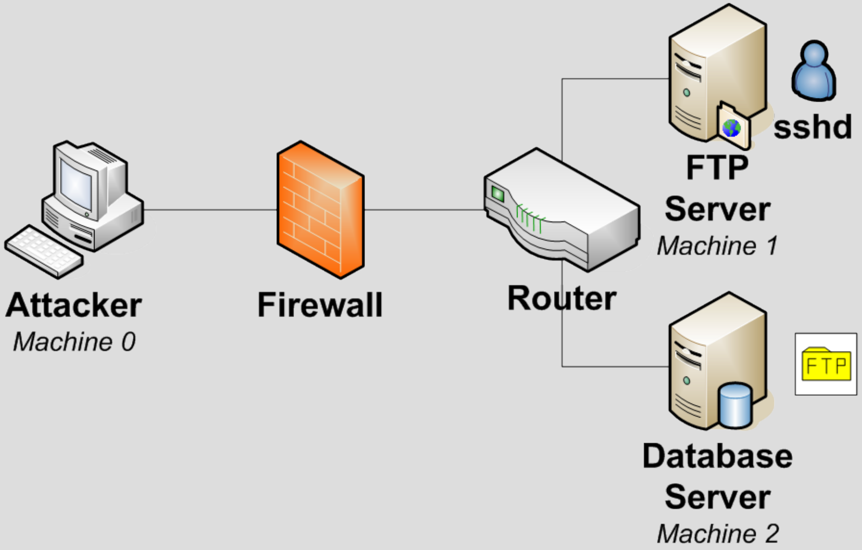
CVE-2009-1918 was identified on user workstations

Possible attack paths




5
6:0.2
16
17
18
15
20
10
24
4:1
14:0.8
19:0.5
23:1
13:0.82
21
22
12:0.9
28
11:0.738
27:1
9:1
29
3:0.7904
30
25
26
8:1.0
2:0.6
7:0.2
1:0.47424

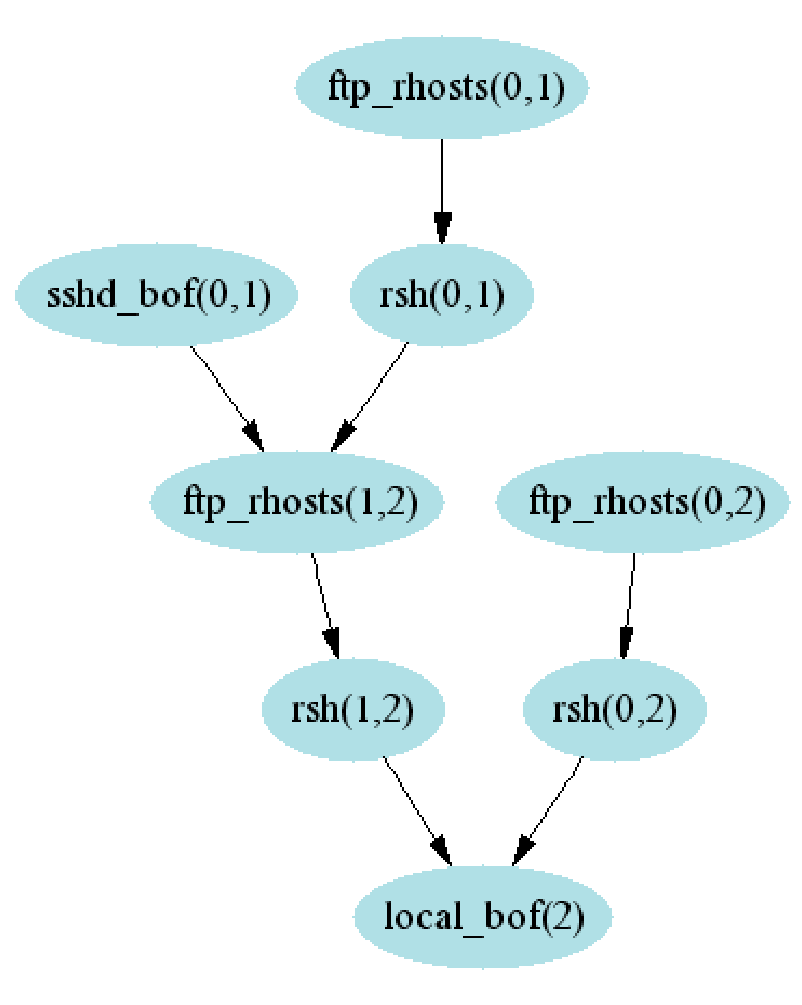
Attack Graph (Another Example)



Different Paths for the Attack

- *sshd_bof(0,1)* → *ftp_rhosts(1,2)* → *rsh(1,2)* → *local_bof(2)*
- *ftp_rhosts(0,1)* → *rsh(0,1)* → *ftp_rhosts(1,2)* → *rsh(1,2)* → *local_bof(2)*
- *ftp_rhosts(0,2)* → *rsh(0,2)* → *local_bof(2)*

Attack Graph from machine 0 to DB Server



Summary on Risk Modeling

- Based on attack graphs, we have proposed a model for security risk analysis of information systems
- The metric meets intuitive requirements
- We plan to extend this model for hybrid cloud environment

Outline

- Risk-based Approach to data management in cloud [Sharad]
  - Motivation – focusing on why
- Modeling Risks [Anoop]
  - State-of-the-art In enterprise networks
  - Thoughts on generalizing to cloud data
- **Given risks, example data and workload partitioning problem [Sharad]**
  - Some initial results [IEEE Cloud, 2012-a, 2012-b]

Data & Computation Partitioning Problem

- **Given**

Relational Data

Set of constraints and desired goals on sensitivity, performance, costs, etc.

Student

Sensitive

s_id	name	Course	dept
1	James	123	CS
2	Charlie	123	EE
3	John	987	CS
4	Matt	245	ECON

HIVE/SQL Queries

Q1: SELECT name, Course from Student where dept = CS

Q2: SELECT dept, count(*) FROM Student GROUP_BY dept HAVING dept != CS

Q3: SELECT * FROM Student WHERE course != 987

How to partition the table ?

How to represent data on the public machines?

How to split computation?

Q1 has the most sensitive exposure

Q2 execution is the most expensive

Computation Partitioning Problem (CPP)

- Find a **subset of given query workload**, $Q_{pub} \subseteq Q$ and **subset of the** such that

$$ORunT(Q',Q'') = \max\begin{cases}\sum_{q \in Q''} freq(q)\ x\ runT_{pub}(q) \\ \sum_{q \in Q'-Q''} freq(q)\ x\ runT_{priv}(q)\end{cases}$$

minimize $ORunT(Q,Q_{pub})$

subject to (1) $store(R_{pub}) + \sum_{q \in Q_{pub}} freq(q)\ x\ proc(q) \leq MC$

(2) $sens(R_{pub}) \leq DC$

(3) $\forall q \in Q_{pub}\ baseTables(q) \subseteq R_{pub}$

The estimated # of sensitive cells exposed

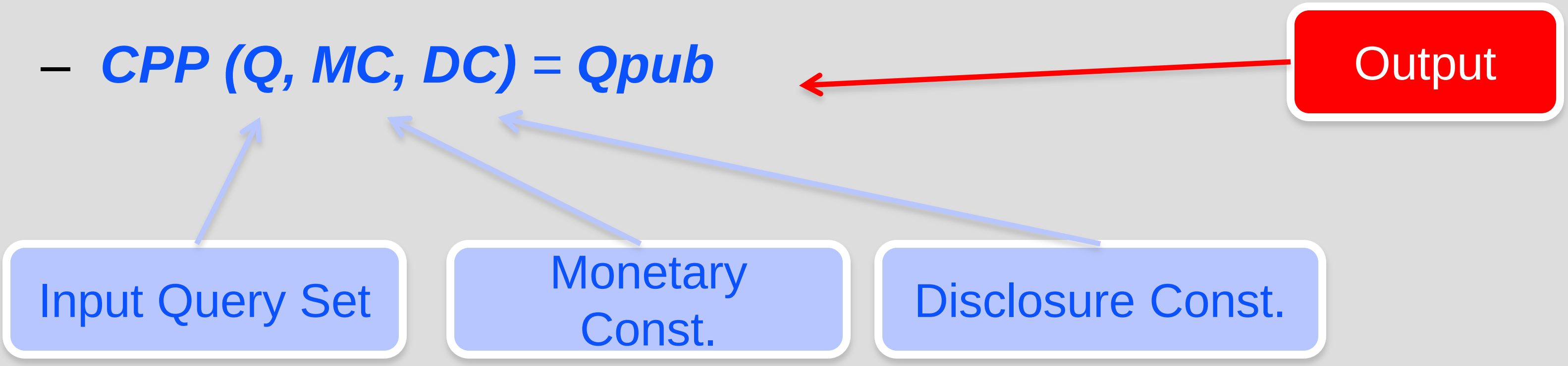
The estimated minimum set of data items necessary to answer query q Є Q

- MC , DC **are user defined constraints**

Solution to CPP

- **CPP can be simplified to only finding *Q_{pub}***

- **Dynamic Programming Approach**
  - *CPP (Q, MC, DC) = Qpub*



Output

Input Query Set

Monetary Const.

Disclosure Const.

Experimental Setting

- **Experimental Setting**
  - Private Cloud: 14 Nodes, located at UTD, Pentium IV, 4GB Ram, 290-320GB disk space
  - Public Cloud: 38 Nodes, located at UCI, AMD Dual Core, 8GB Ram, 631GB disk space
  - Hadoop 0.20.2 and Hive 0.7.1

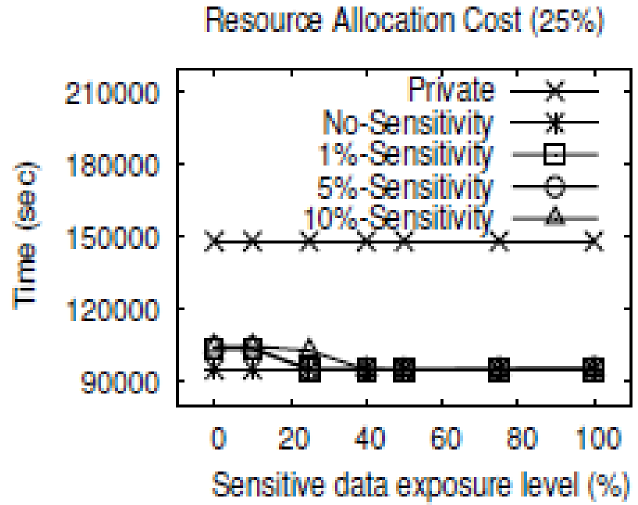
- **Dataset**
  - 100**GB TPC-H Data**

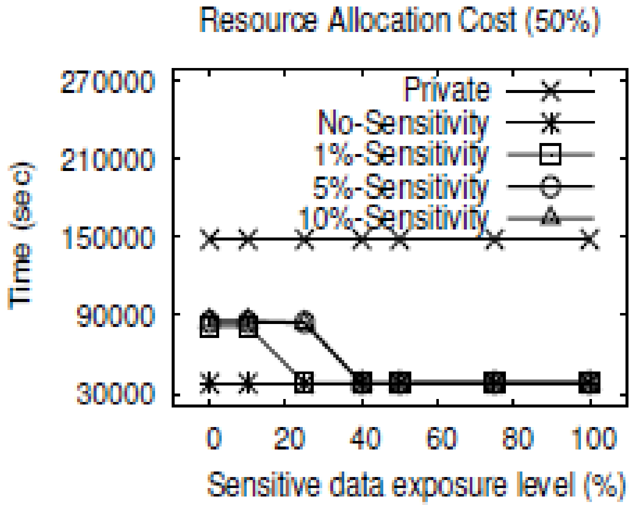
- **Query Workload**
  - 40 queries containing modified versions of Q1, Q3, Q6, Q11 of TPC-H Queries

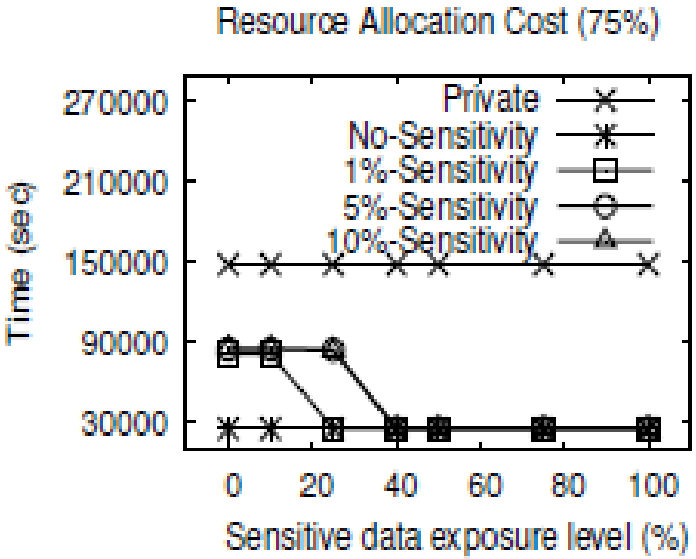
Experimental Setting

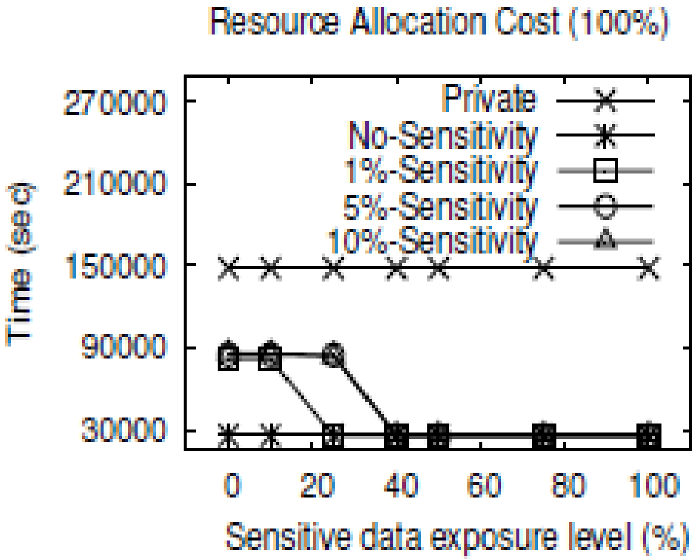
- Estimation of Weight (w_x)
  - Running all 22 TPC-H queries for a 300GB dataset
  - w_{pub} ≈ 40MB/sec , w_{priv} ≈ 8MB/sec
- Resource Allocation Cost
  - Amazon S3 Pricing for storage and communication
    - Storage = $0.140/GB + PUT, Communication= $0.120/GB + GET
    - PUT=$0.01/1000 request, GET=$0.01/10000 request
  - Amazon EC2 and EMR Pricing for processing
    - $0.085 + $0.015 = $0.1/hour
- Sensitivity
  - Customer : *c_name, c_phone, c_address attributes*
  - Lineitem: All attributes in %1-5-10 of tuples

Experimental Results









Summary

- Challenge in adopting cloud-based solutions → loss of control over data
- Leads to privacy & security concerns
- Owners need tools that empower them to manage their sensitive information in the cloud
  - Cryptography offers only limited solutions. It is part of, but not the whole solution.
- risk-minimization based approach offers an attractive possibility. Empowers users to control
  - how data is represented in cloud
  - When to release more and when to scale back
  - Supports mechanism to strike the required balance between utility and data loss (exposure) risk.

Radicle Publications

- Building Disclosure Risk Aware Query Optimizers for Relational Databases, Mustafa Canim, Murat Kantarcioglu, Bijit Hore, Sharad Mehrotra, VLDB 2010.
- Secure Multidimensional Range Queries over Outsourced Data, Bijit Hore, Mustafa Canim, Murat Kantarcioglu, Sharad Mehrotra, VLDBJ 2012.
- CloudProtect: Managing Data Privacy in Cloud Applications, Mamadou Diallo, Bijit Hore, Ee-Chien Chang, Sharad Mehrotra, Nalini Venkatasubramanian, IEEE CLOUD 2012.
- Risk-aware Workload Distribution in Hybrid Clouds, Kerim Oktay, Vaibhav Khadilkar, Bijit Hore, Murat Kantarcioglu, Sharad Mehrotra, Bhavani Thuraisingham, IEEE CLOUD 2012.
- Indexing Encrypted Documents for Supporting Efficient Keyword Search. Bijit Hore, Ee-Chien Chang, Mamadou Diallo, Sharad Mehrotra, SDM 2012.
- Secure Quasi-Realtime Collaborative Editing over Low-Cost Storage Services. Chunwang Zhang, Junjie Jin, Ee-Chien Chang, Sharad Mehrotra, SDM 2012.
- CloudProtect: A Middleware for Managing Privacy in Cloud Applications, Mamadou Diallo (Masters Thesis) UCI 2012.
- Hibrider: A Framework for Partitioning Workloads over Hybrid Cloud, Vaibhav Khadilkar,Kerim Oktay, Murat Kantarcioglu, Sharad Mehrotra, Bhavani Thuraisingham, TR '12
- Secure Data Processing in Hybrid Clouds, Vaibhav Khadilkar,Kerim Oktay, Murat Kantarcioglu, Sharad Mehrotra, IEEE Data Engineering Bulletin, Dec. 201**2.**