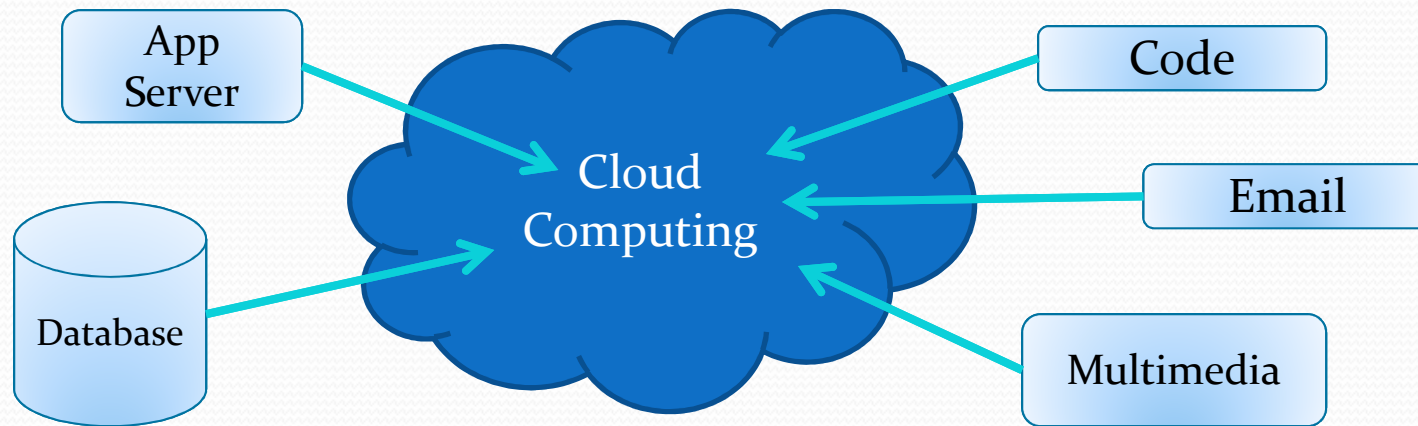


# Risk Aware Data Processing on the Cloud

Murat Kantarcioglu,

Joint work with (Sharad Mehrotra (UCI), Bhavani  
Thuraisingham, Kerim Oktay (UCI), Vaibhav Khadilkar)

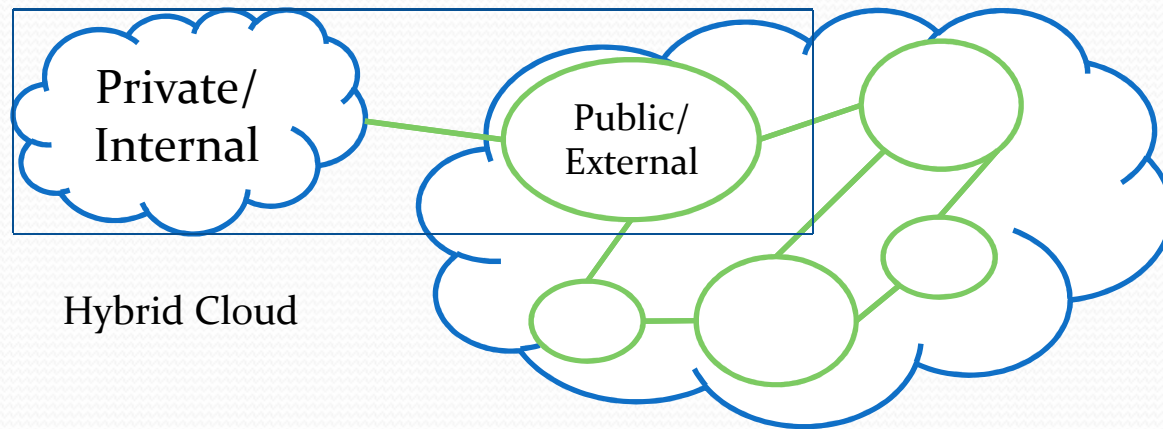
# Cloud Computing



- Like Software as a service and DAS model offers many advantages
  - Better availability
  - Reduced Costs
  - Unlimited scalability and elasticity

# Hybrid Cloud

- Integrates local infrastructure with public cloud resources



- **Extra Advantages**
  - The flexibility of shifting workload to *public cloud* when the *private cloud* is overwhelmed (Cloud Bursting)
  - Utilizing in-house resources along with public resources
- **Cons**
  - Sensitive data exposure
  - Public Cloud Resource Allocation Cost (both storage and computing)




THE ECS ARCHIVE: Storage, Security, Mobility and more...

## 2013: Year of the hybrid cloud

Hybrid clouds, cloud brokers, big data and software-defined networking (SDN) predicted to be the major trends in cloud computing in 2013.

By *Christine Burns*, Network World

December 03, 2012 12:08 AM ET

 3 Comments  Print

      Like 90   + Briefcase  More

Network World -

The time for dabbling in cloud computing is over, say industry analysts. 2013 is the year that companies need to implement a hybrid cloud strategy that puts select workloads in the public cloud and keeps others in-house.

"Next year has to be the year that enterprises get serious about having real cloud operations as part and parcel of their IT operations," says John Treadway, vice president at Cloud Technology Partners, a consultancy.

[10 cloud predictions for 2013](#)

[Careers in the cloud](#)

Treadway says that in the last year, he and his colleagues have worked with many large



COLLAGE ILLUSTRATION: STEPHEN SAUER

# Data & Computation Partitioning Challenge

Q1: SELECT name, ssn from Student

Q2: SELECT dept, count(\*) FROM Student  
GROUP\_BY dept

Sensitive

Student

s_id	name	ssn	dept
1	James	1234	CS
2	Charlie	4321	EE
3	John	5645	CS
4	Matt	8743	ECON

How to split computation?

How to partition the table ?

- Q1 contains sensitive information
- Q2 execution is more expensive

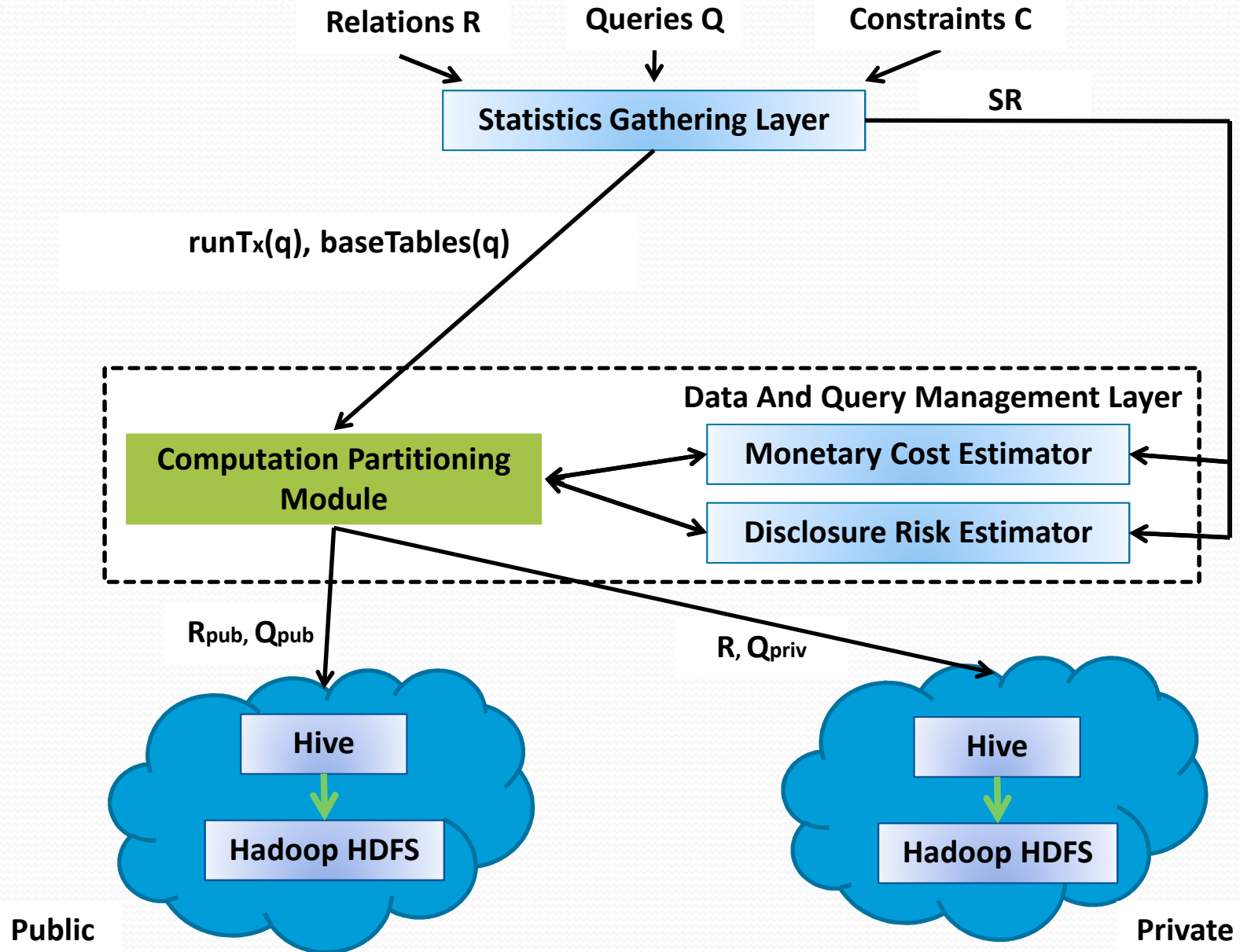
## Constraints



# Design Spectrum

- Data Model
  - **Relational**, Semi-structured, Key-Value Stores, Text
- Sensitivity Model
  - **Attribute Level**, Privacy Associations, View-Based
- Partitioning Models
  - **Workload Partitioning**, Intra-query Parallelism, Dynamic Workload
- Minimization Priority
  - **Running Time**, Sensitive Data Disclosure, Monetary Cost

# Detailed Hybrid Cloud Architecture



# Computation Partitioning Problem (CPP)

- Find a **subset of given query workload**,  $Q_{pub} \subseteq Q$  and **subset of the given dataset**  $R_{pub} \subseteq R$  where

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

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

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

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

- MC, DC are user defined constraints**



# Metrics in CPP

- Query Execution Time (**runT<sub>x</sub>(q)**)

$$\text{runT}_x(q) = \frac{\sum_{\substack{\forall \text{ operator} \\ \rho \in q}} \text{inpSize}(\rho) + \text{outSize}(\rho)}{w_x}$$

- Monetary Costs

- **stor(R<sub>pub</sub>)** : Storage monetary cost of the public cloud partition
- **proc(q)** : Processing monetary cost of a public side query q

- Sensitive Data Disclosure Risk (**sens(R<sub>pub</sub>)**)

- Estimated number of sensitive cells within R<sub>pub</sub>

# 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 and Statistic Collection
  - 100GB TPC-H Data
- Query Workload
  - 40 queries containing modified versions of Q1, Q3, Q6, Q11



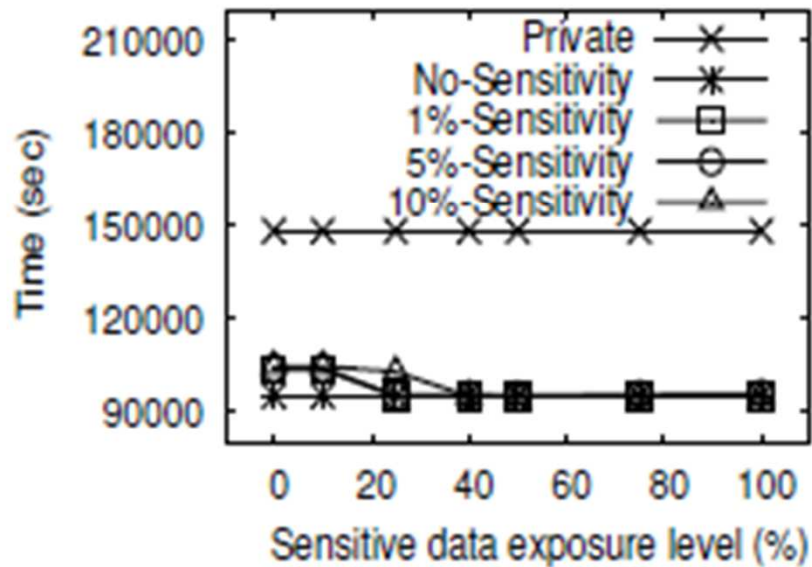
# Experimental Setting

- Estimation of Weight ( $w_x$ )
  - Running all 22 TPC-H queries for a 300GB dataset
  - $w_{\text{pub}} \approx 40\text{MB/sec}$  ,  $w_{\text{priv}} \approx 8\text{MB/sec}$
- 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

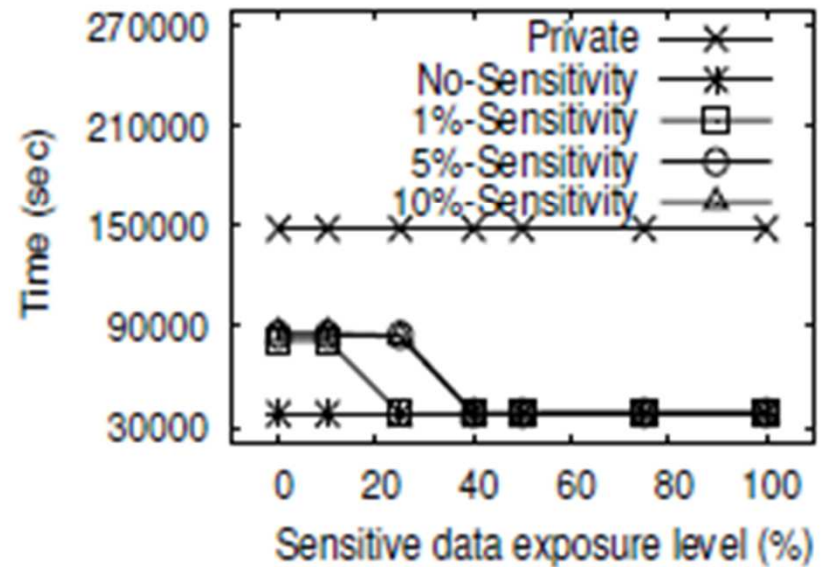


# Experimental Results

Resource Allocation Cost (25%)



Resource Allocation Cost (50%)



# Conclusions

- Hybrid clouds offer interesting security and load balancing alternatives
- We focused on inter-query distribution based approach
- Public clouds could be leveraged in a secure manner efficiently.

# Our Other Work in Cloud Security

- Develop efficient access control for map-reduce type systems (e.g., Hadoop)
- Cloud Auditing
- Encryption support for key-value stores (e.g., encrypted data support for HBASE)