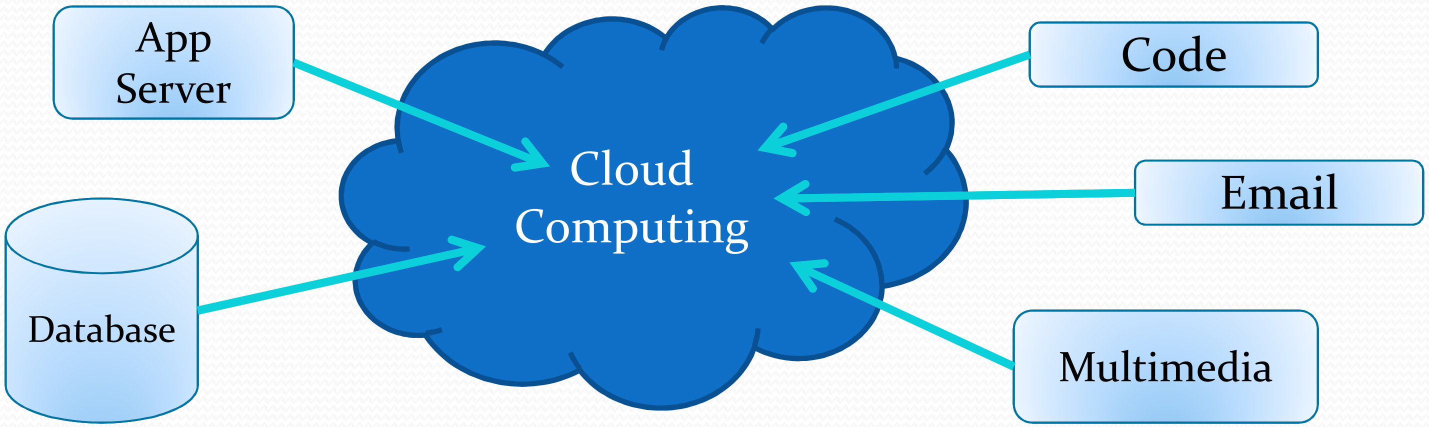
Risk Aware Data Processing over Hybrid Cloud

Murat Kantarcioglu,

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

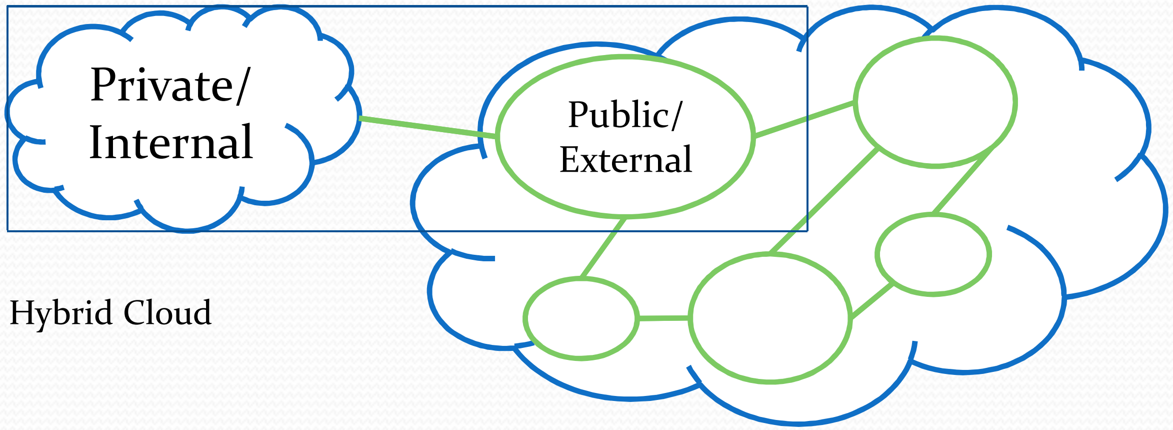
Cloud Computing



- 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



Hybrid Cloud

- **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
  - Provide better performance compared to pure crypto. solutions
- **Cons**
  - Increased risk ?
  - 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 +1 Like 90 + Briefcase More

Network World -

COLLAGE ILLUSTRATION: STEPHEN SAUER

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

Data & Computation Partitioning Challenge

Q1: SELECT name, ssn from Student

Q2: SELECT dept, count(*) FROM Student GROUP_BY dept

How to split computation?

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 partition the table ?

Constraints

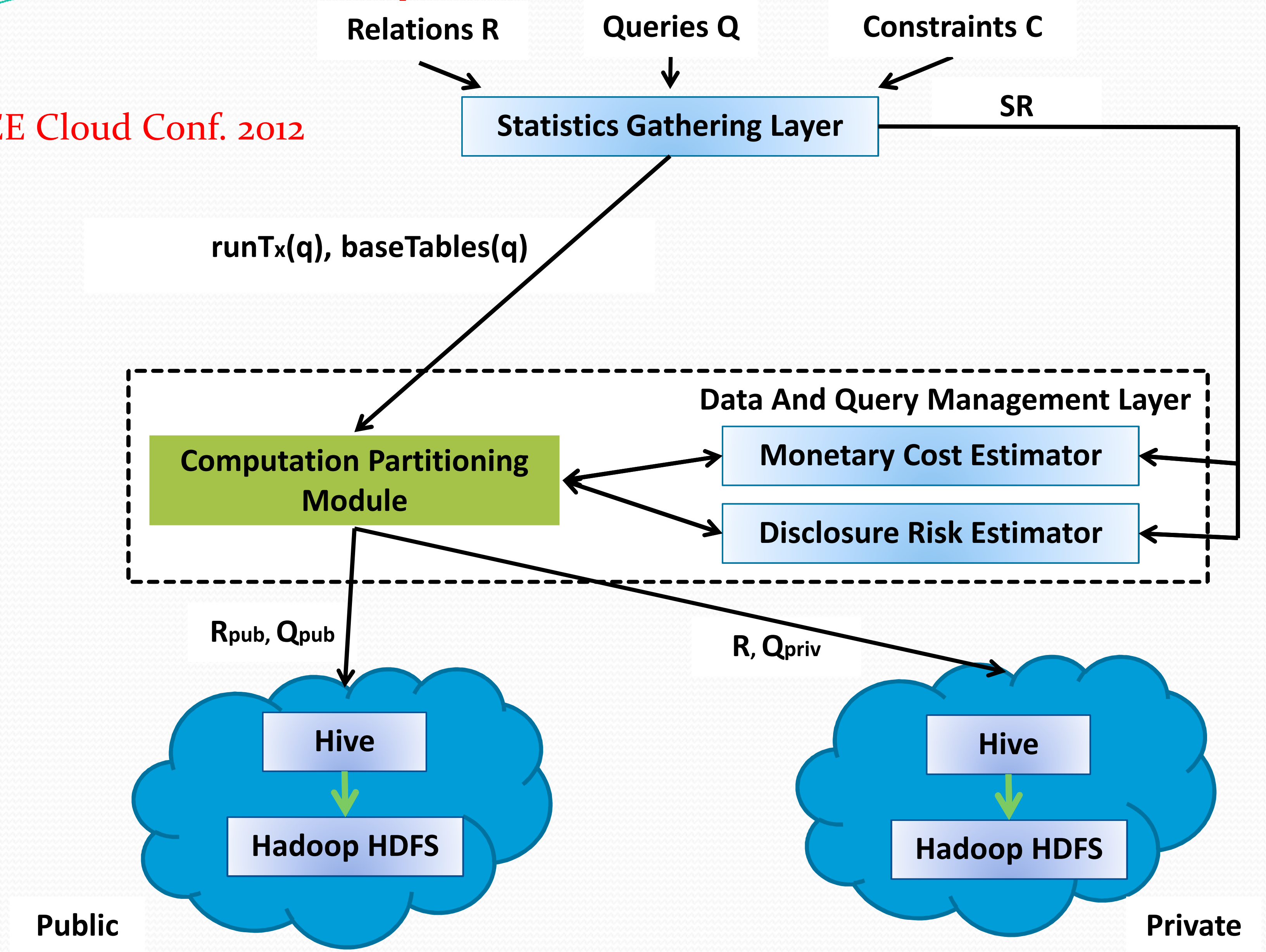
- **Q1 contains sensitive information**
- **Q2 execution is more expensive**

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

IEEE Cloud Conf. 2012



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

$$\begin{array}{ll} \textbf{minimize} & ORunT\,(Q, Q_{pub}) \\ \textbf{subject to} & (1)\ store(R_{pub}) + \sum_{q \in Q_{pub}} freq(q)\ x\ proc(q) \leq MC \\ & (2)\ sens(R_{pub}) \leq DC \\ & (3)\ \forall q \in Q_{pub}\ \ baseTables\,(q) \subseteq R_{pub} \end{array}$$

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

Metrics in CPP

- **Query Execution Time ($runT_x(q)$)**

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

- **Monetary Costs**
  - ***$stor(R_{pub})$:*** Storage monetary cost of the public cloud partition
  - ***$proc(q)$:*** Processing monetary cost of a public side query q
- **Sensitive Data Disclosure Risk ($sens(R_{pub})$)**
  - Estimated number of sensitive cells within R_{pub}

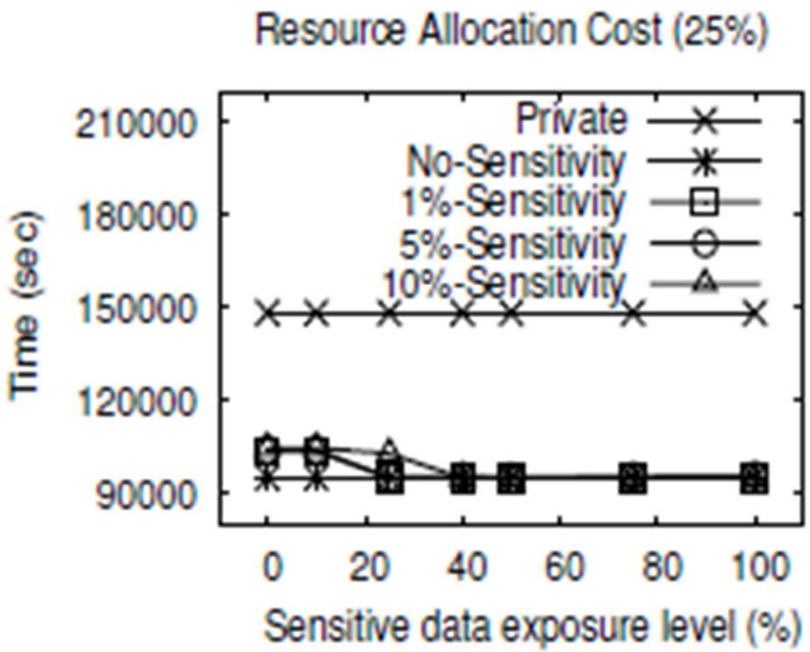
Experimental Setting

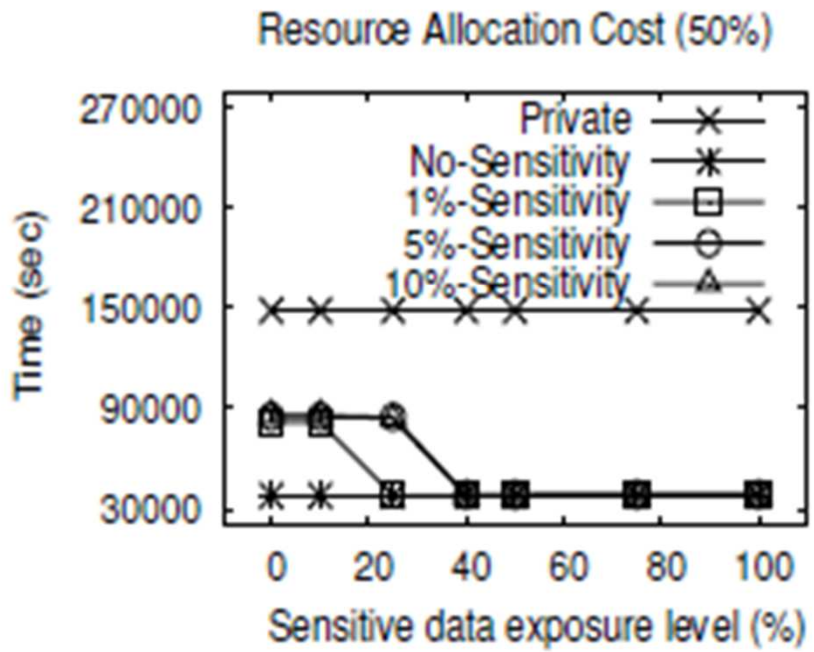
- Experimental Setting
  - Private Cloud: 14 Nodes, located at UTD, Pentium IV, 4GB Ram, 290-320GB disk space
  - Public Cloud: 38 Nodes, located at UCI, AMD Dual Core, 8GB Ram, 631GB disk space
  - Hadoop 0.20.2 and Hive 0.7.1
- Dataset and Statistic Collection
  - 100GB TPC-H Data
- Query Workload
  - 40 queries containing modified versions of Q1, Q3, Q6, Q11

Experimental Setting

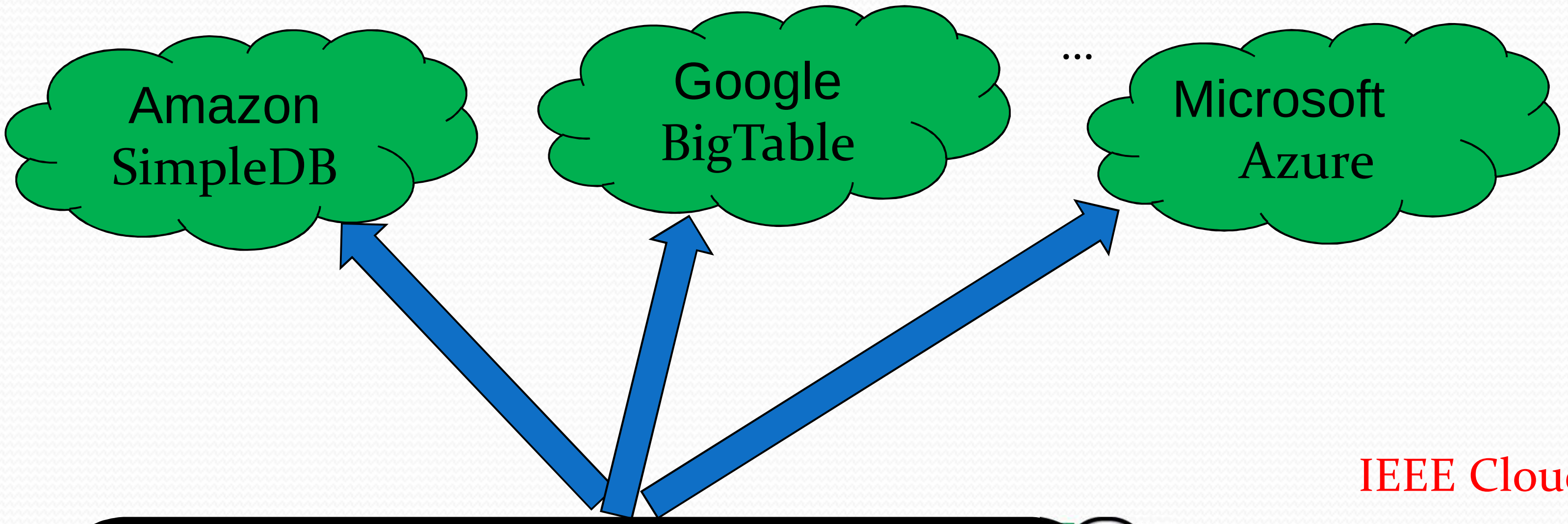
- Estimation of Weight (w_x)
  - Running all 22 TPC-H queries for a 300GB dataset
  - $w_{pub} \approx$ 40MB/sec , $w_{priv} \approx$ 8MB/sec
- Resource Allocation Cost
  - Amazon S3 Pricing for storage and communication
    - Storage = $0.140/GB + PUT, Communication= $0.120/GB + GET
    - PUT=$0.01/1000 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





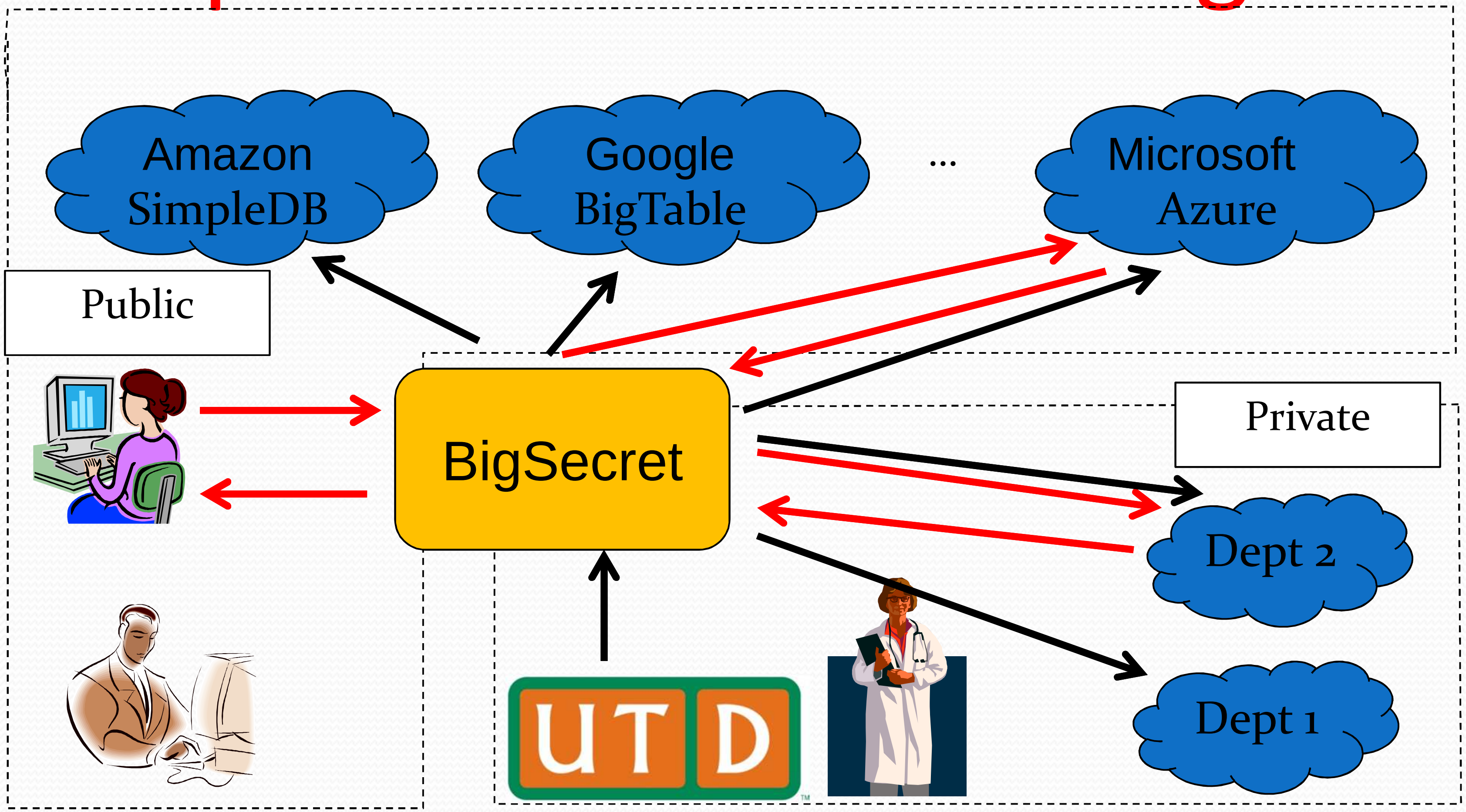
Application to Key-Value Stores



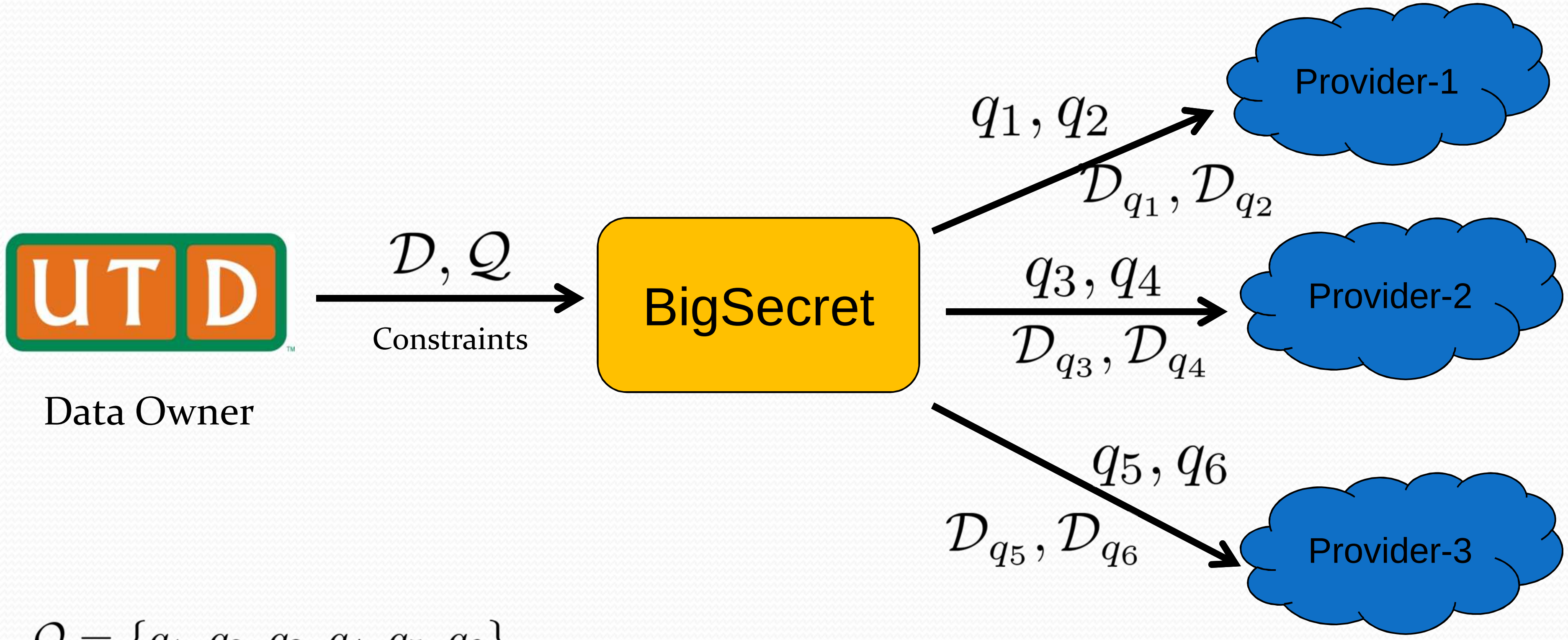
Key	Value
pattuk_erman:bank	1919381
pattuk_erman:ssn	1928319
ulusoy_huseyin:bank	4476861
ulusoy_huseyin:ssn	1148793

IEEE Cloud Conf. 2013

Proposed Framework: BigSecret

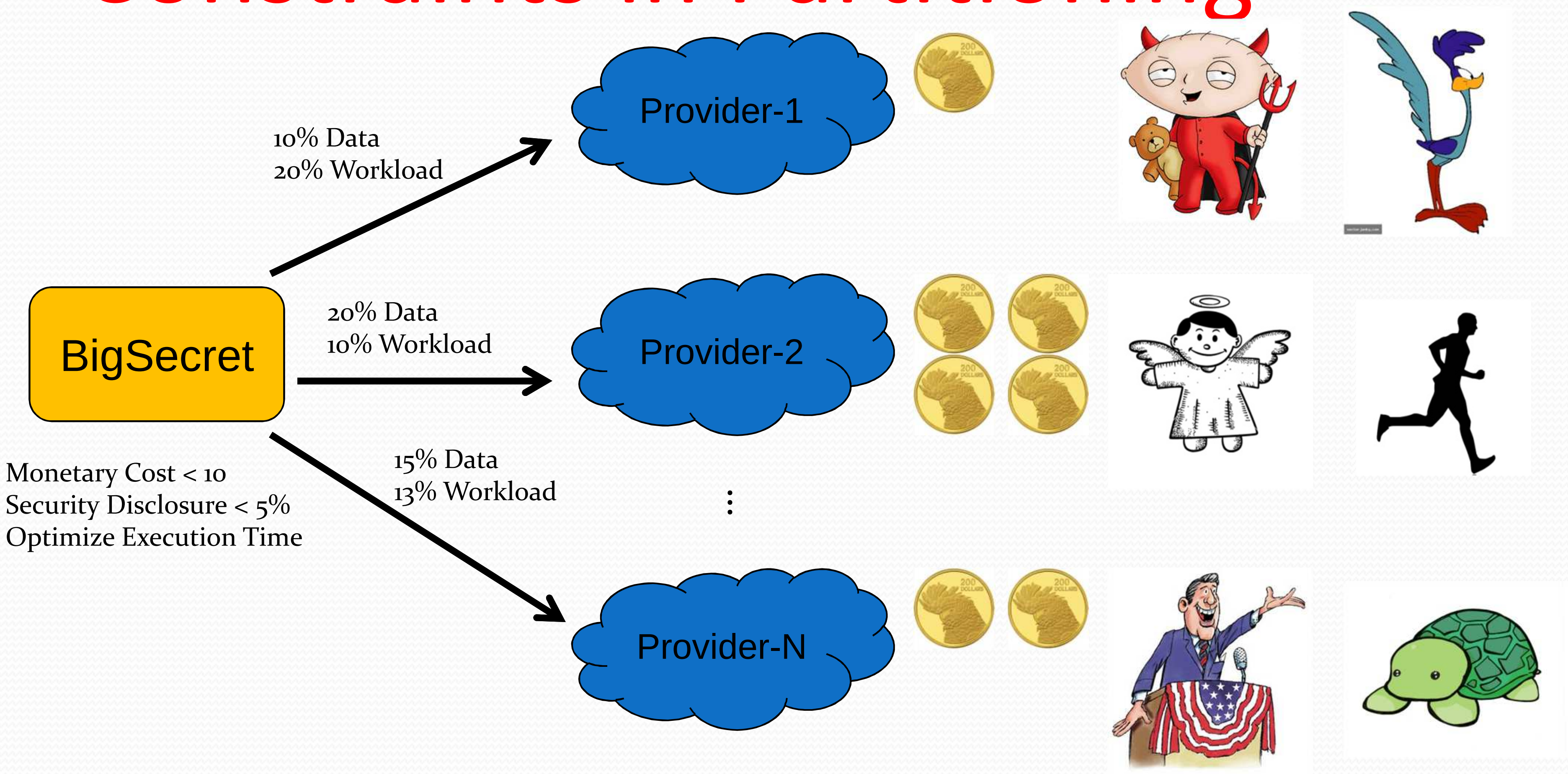


Data and Workload Sharing



$$\mathcal{Q} = \{q_1, q_2, q_3, q_4, q_5, q_6\}$$

Constraints in Partitioning

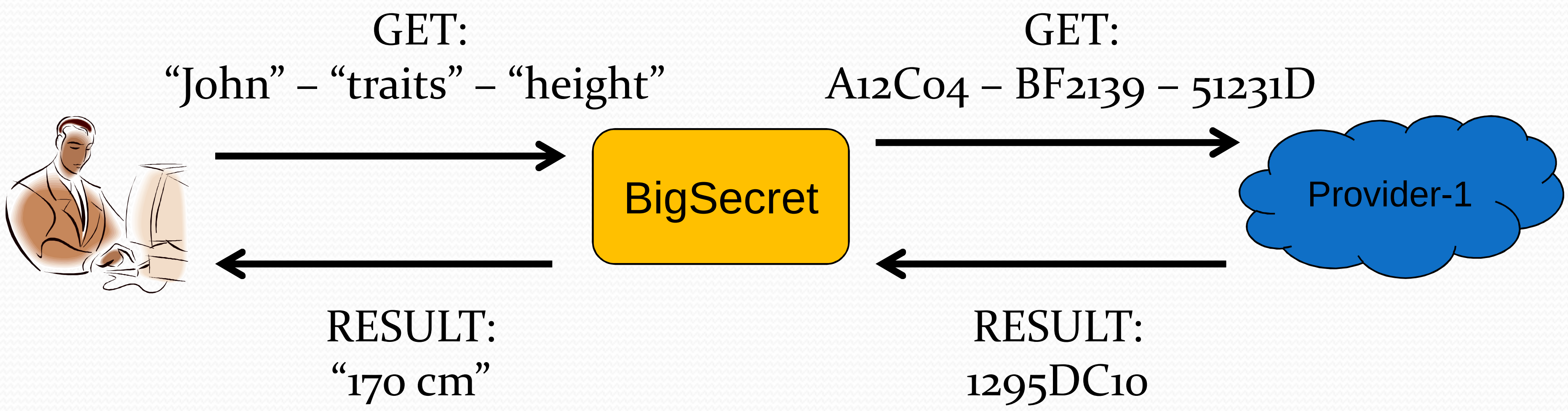


Storing Data in Secure Form

	Model-1	**Model-2**	**Model-3**
row	$Map(row)$	$H(row)$	$H(row)$
fam	$Map(fam)$	$H(fam)$	0
qua	$Map(qua) \| E(KEY)$	$H(qua) \| E(KEY)$	$E(KEY)$
ts	$Map(ts)$	$H(ts)$	1
val	$E(val)$	$E(val)$	$E(val)$

- Transform data using *Encryption Models*

Query Execution

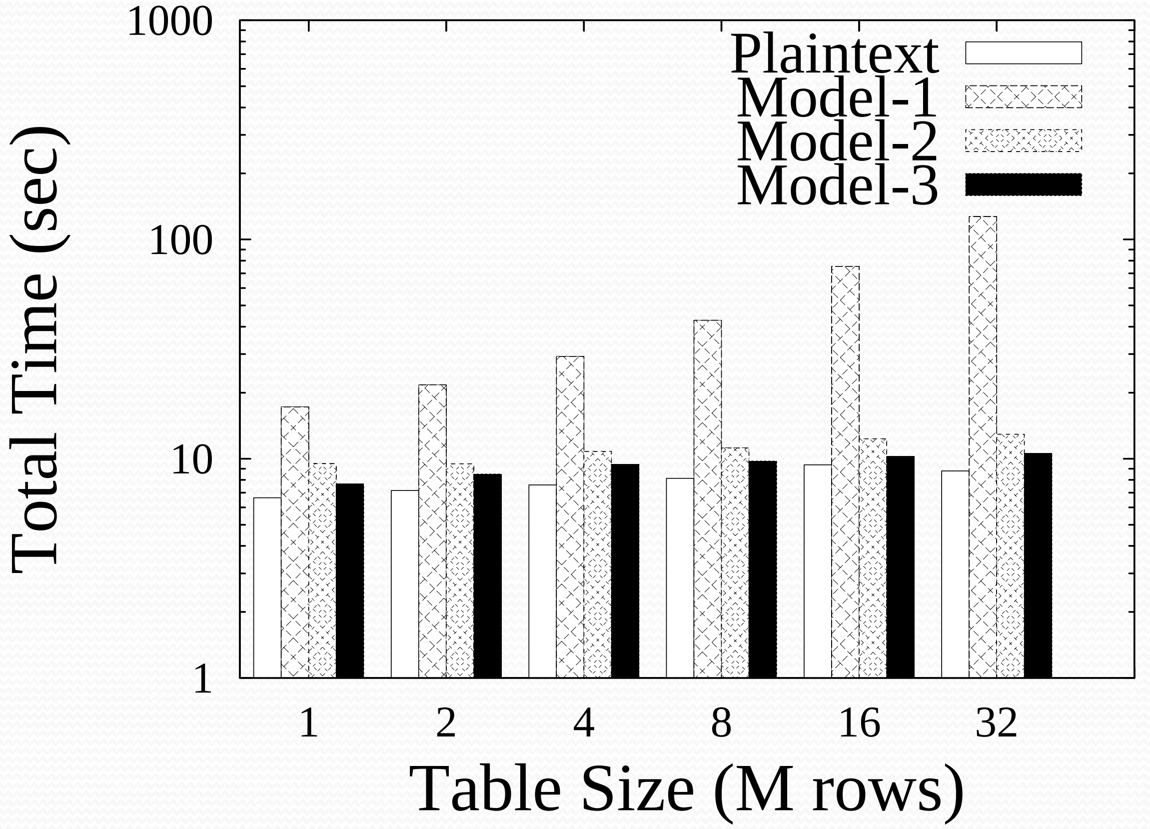


Experiments

- Performed experiments using Yahoo! Cloud Serving Benchmark
- Created tables consisting of 1,2,4,8,16, and 32 Millions of rows
  - Each row has 10 Key-Value entries of 100B
- Created 3 different workloads
  - 1K queries for single-cloud experiments
  - 100K queries for multi-cloud experiments

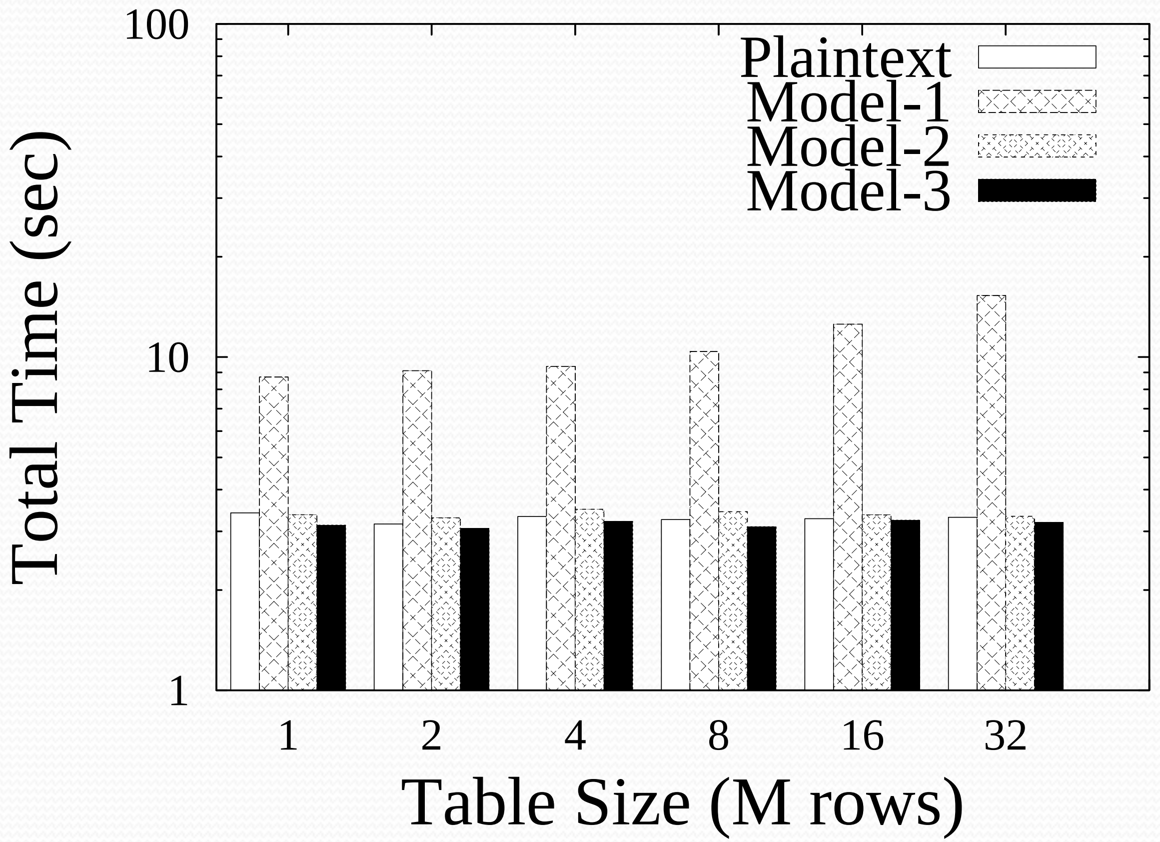
	Workload-1	Workload-2	Workload-3
Put (%)	5	95	25
Get (%)	95	5	25
Scan (%)	0	0	50

Single-Cloud Experiments



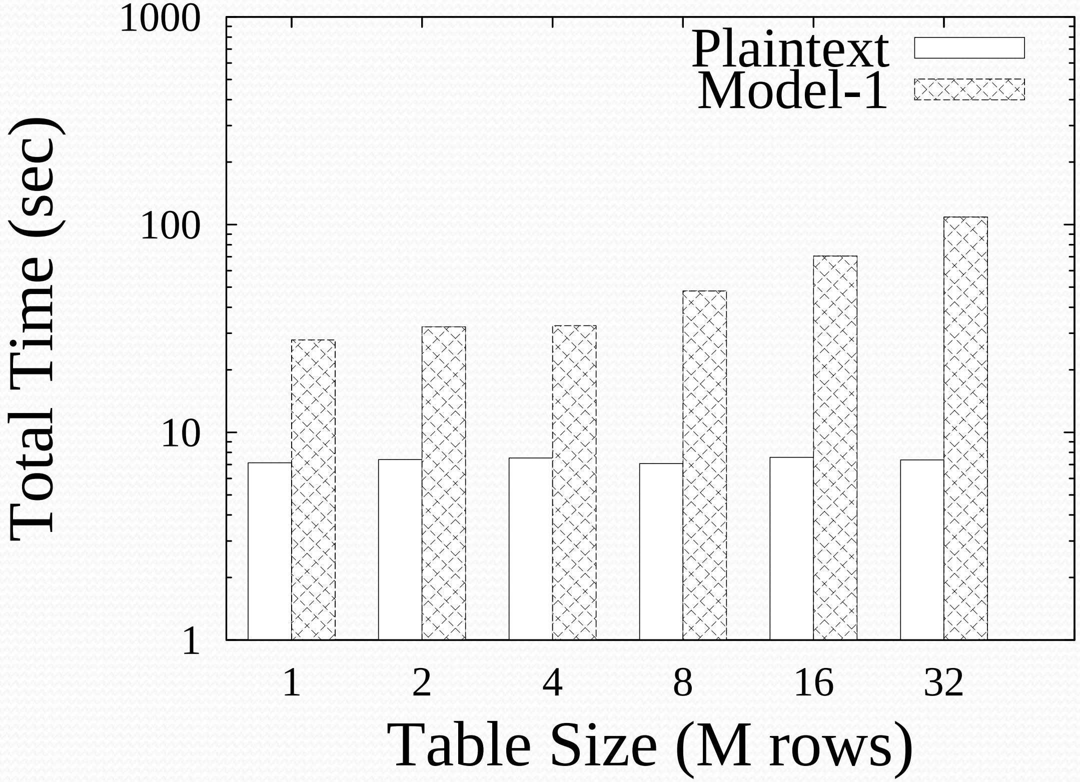
Workload – 1 (Get intensive)

Single-Cloud Experiments



Workload – 2 (Put intensive)

Single-Cloud Experiments

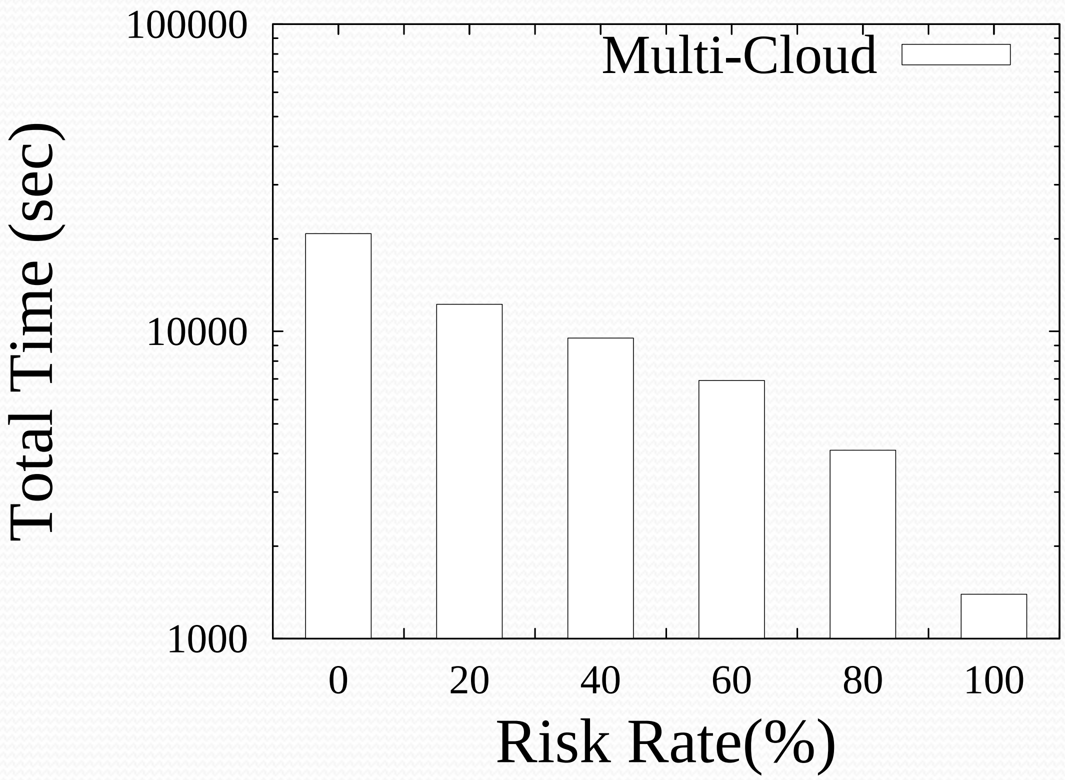


Workload – 3 (Scan intensive)

Multi-Cloud Experiments

Provider / Properties	Provider 1	Provider 2
Storage	Plaintext	Model-1
Risk weight	1	0.7
Speed	Fast	Slow
Monetary cost	$700	$3700
Sensitivity disclosure risk	100%	70%

Multi-Cloud Experiments



Workload – 3 (Scan intensive)

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.
- Encrypted query processing could be integrated to our approach

Acknowledgements

- This work was partially supported by:
  - Air Force Office of Scientific Research Grant FA9550-12-1-0082
  - National Institutes of Health Grants 1R01LM009989 and 1R01HG006844
  - National Science Foundation (NSF) Grants Career-CNS-0845803, CNS-0964350, CNS-1016343, CNS-1111529, CNS-1228198
  - Army Research Office Grant W911NF-12-1-0558.