Cloud-Centric Assured Information Sharing, for Secure Social Networking

Dr. Bhavani Thuraisingham
The University of Texas at Dallas (UTD)
bhavani.thuraisingham@utdallas.edu
@CyberUTD

August 21, 2015

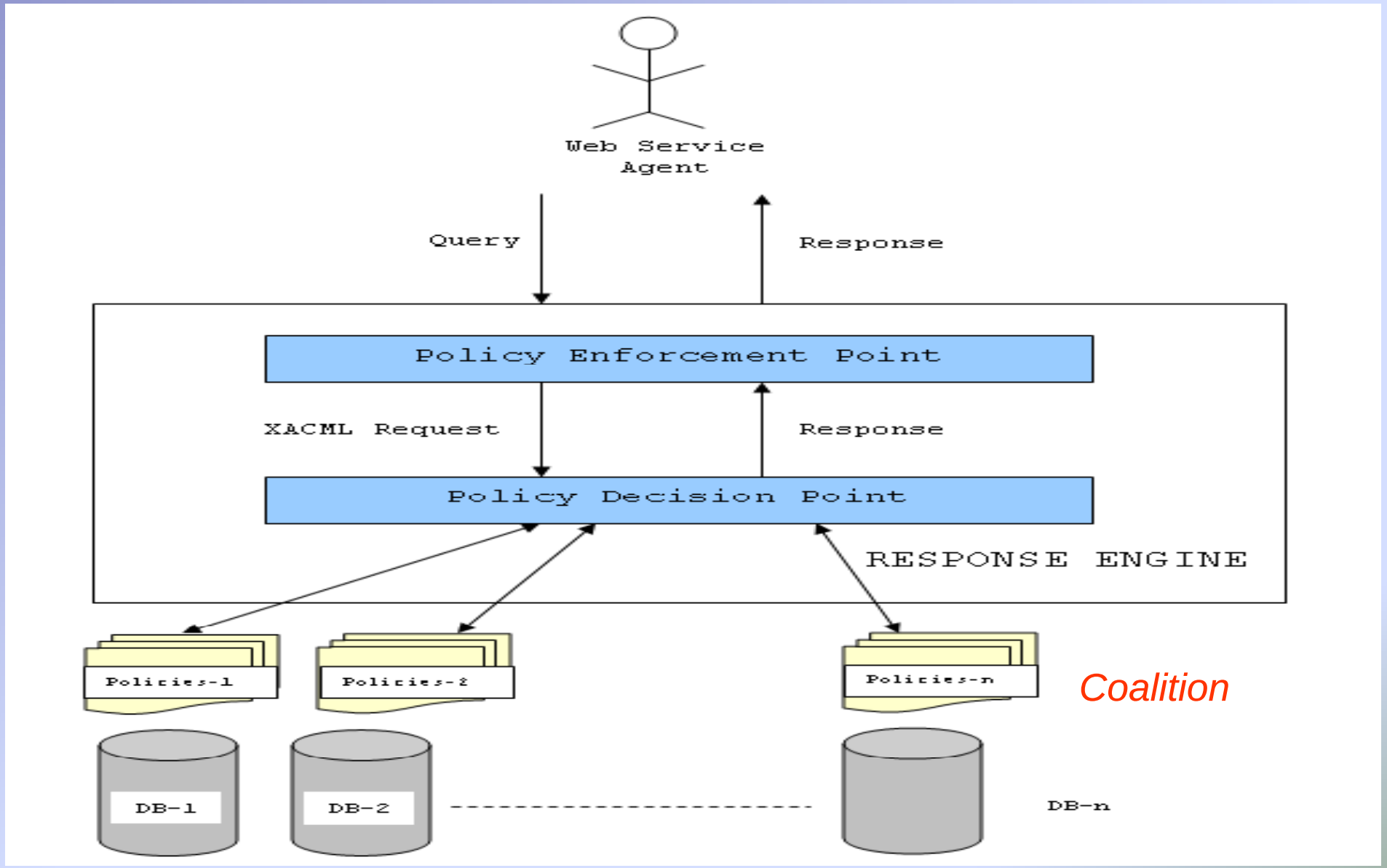
Outline and Acknowledgements

- Assured (Secure) Information Sharing
- Secure Cloud Computing
- Assured Information Sharing in the Cloud
- Analyzing Social Networks and Privacy Implications
- Securing Social Networks
- Directions
- **Acknowledgements: Air Force Office of Scientific Research (subcontract to Purdue University)**

Assured Information Sharing Approach

- Policy and Incentive-based Information Sharing
  - Integrate the Medicaid claims data and mine the data;
  - Enforce policies and determine how much information has been lost (Trustworthy partners);
  - Determine incentives and risks for information sharing
- Apply game theory and probing to extract information from semi-trustworthy partners
- Conduct Active Defence and determine the actions of an untrustworthy partner
  - Defend ourselves from our partners using data analytics techniques
  - Conduct active defence – find our what our partners are doing by monitoring them so that we can defend our selves from dynamic situations

Policy Enforcement Prototype



Layered Framework for Assured Cloud Computing

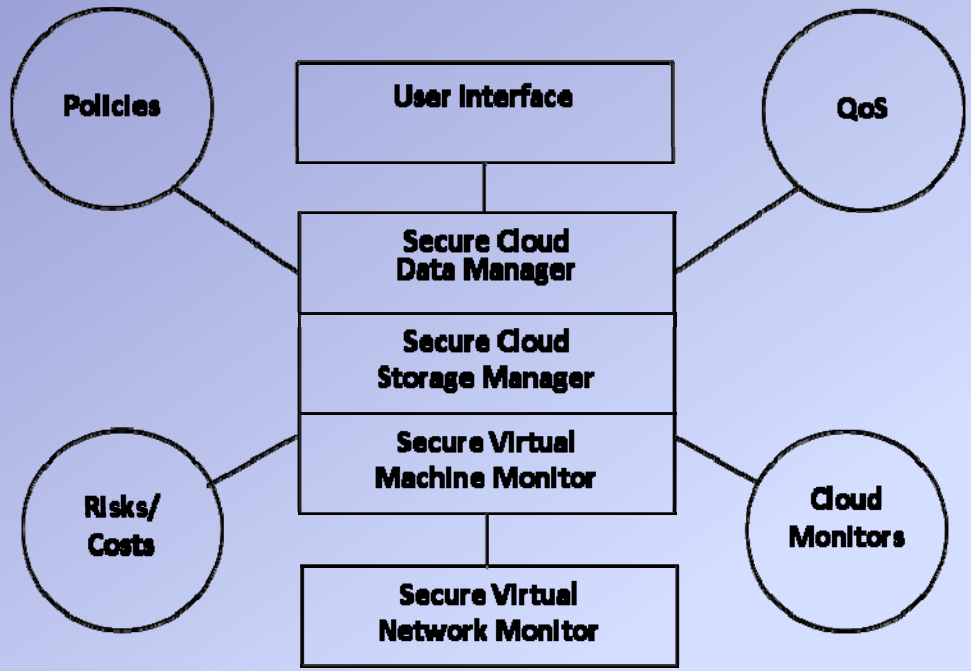


Figure 1. Layered Framework for Assured Cloud

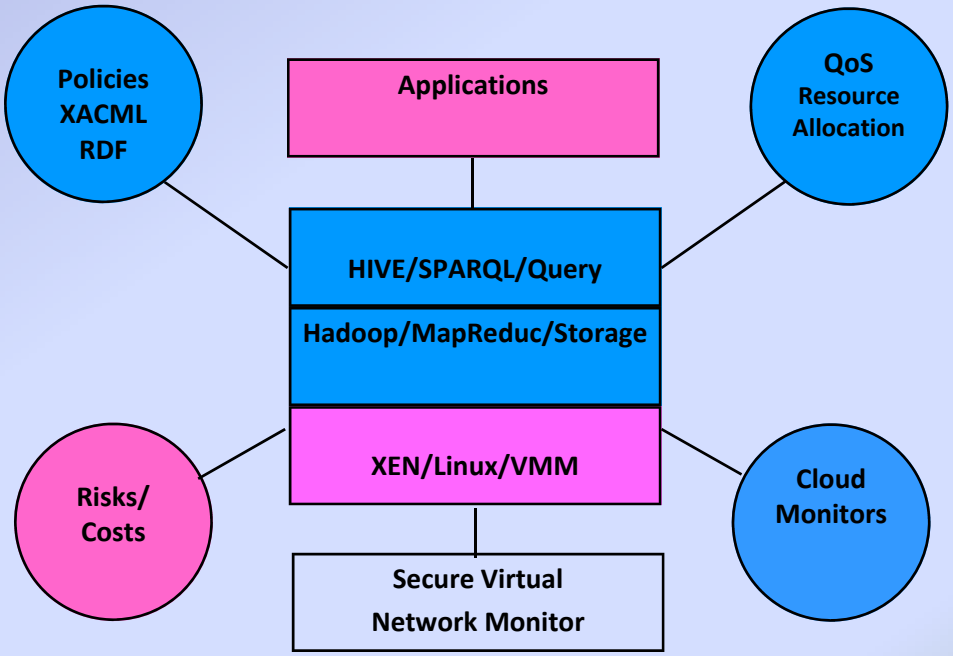


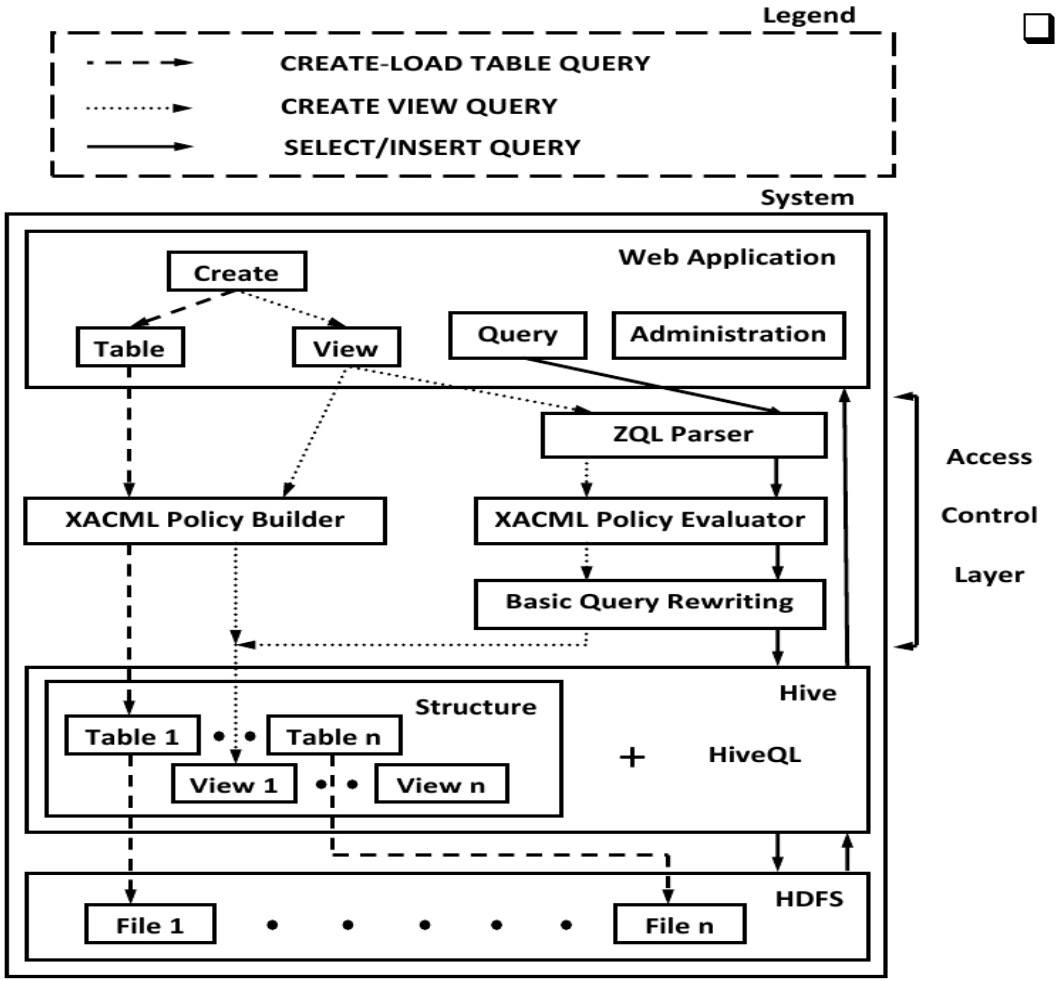
Figure2. Layered Framework for Assured Cloud

Secure Query Processing with Hadoop/MapReduce

- We have studied clouds based on Hadoop
- Query rewriting and optimization techniques designed and implemented for two types of data
- (i) Relational data: Secure query processing with HIVE
- (ii) RDF data: Secure query processing with SPARQL
- Demonstrated with XACML policies
- Joint demonstration with Kings College and University of Insubria
  - First demo (2011): Each party submits their data and policies
  - Our cloud will manage the data and policies
  - Second demo (2012): Multiple clouds

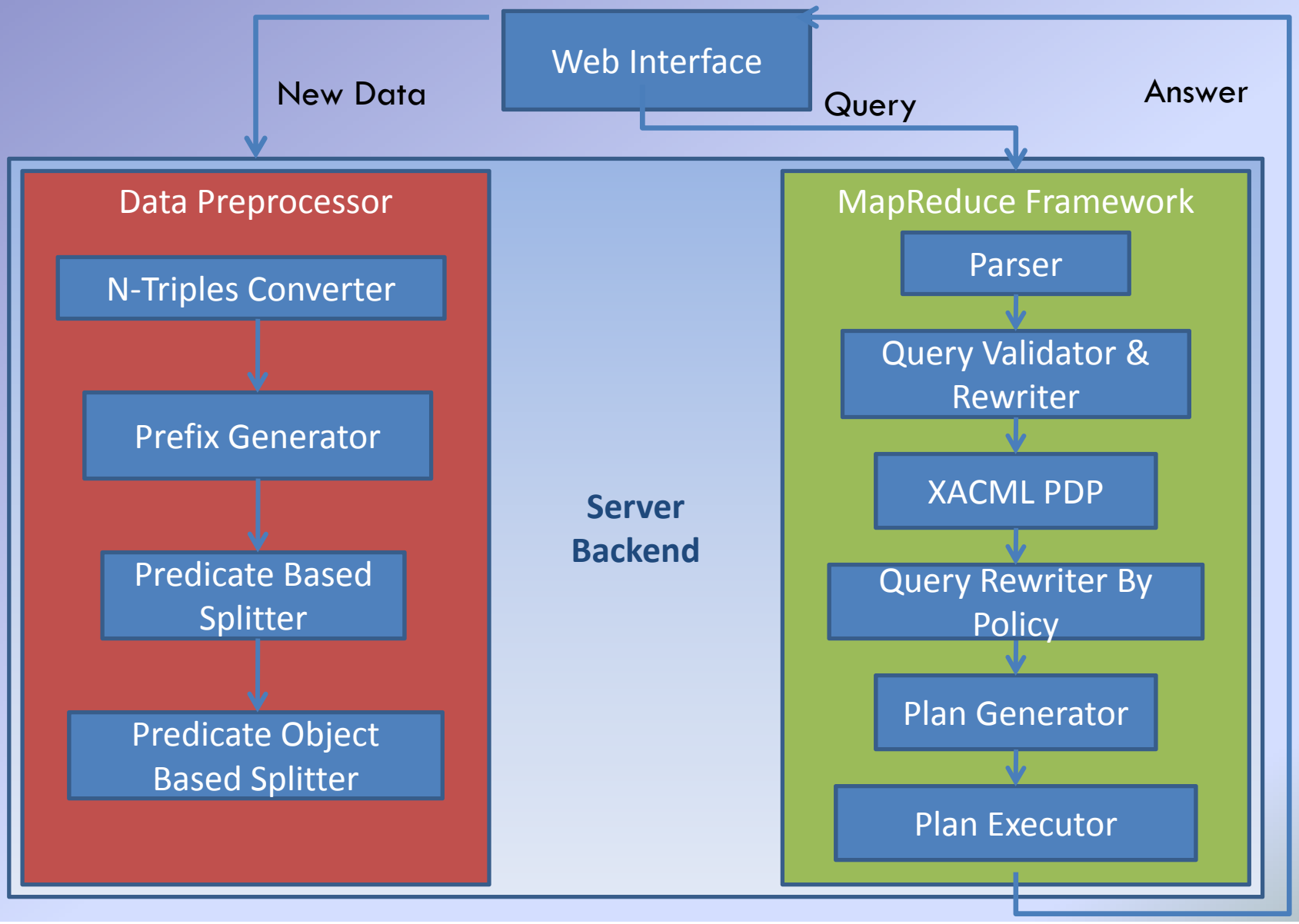
Fine-grained Access Control with Hive

System Architecture



- Table/View definition and loading,
  - Users can create tables as well as load data into tables. Further, they can also upload XACML policies for the table they are creating. Users can also create XACML policies for tables/views.
  - Users can define views only if they have permissions for all tables specified in the query used to create the view. They can also either specify or create XACML policies for the views they are defining.
  - CollaborateCom 2010

SPARQL Query Optimizer for Secure RDF Data Processing

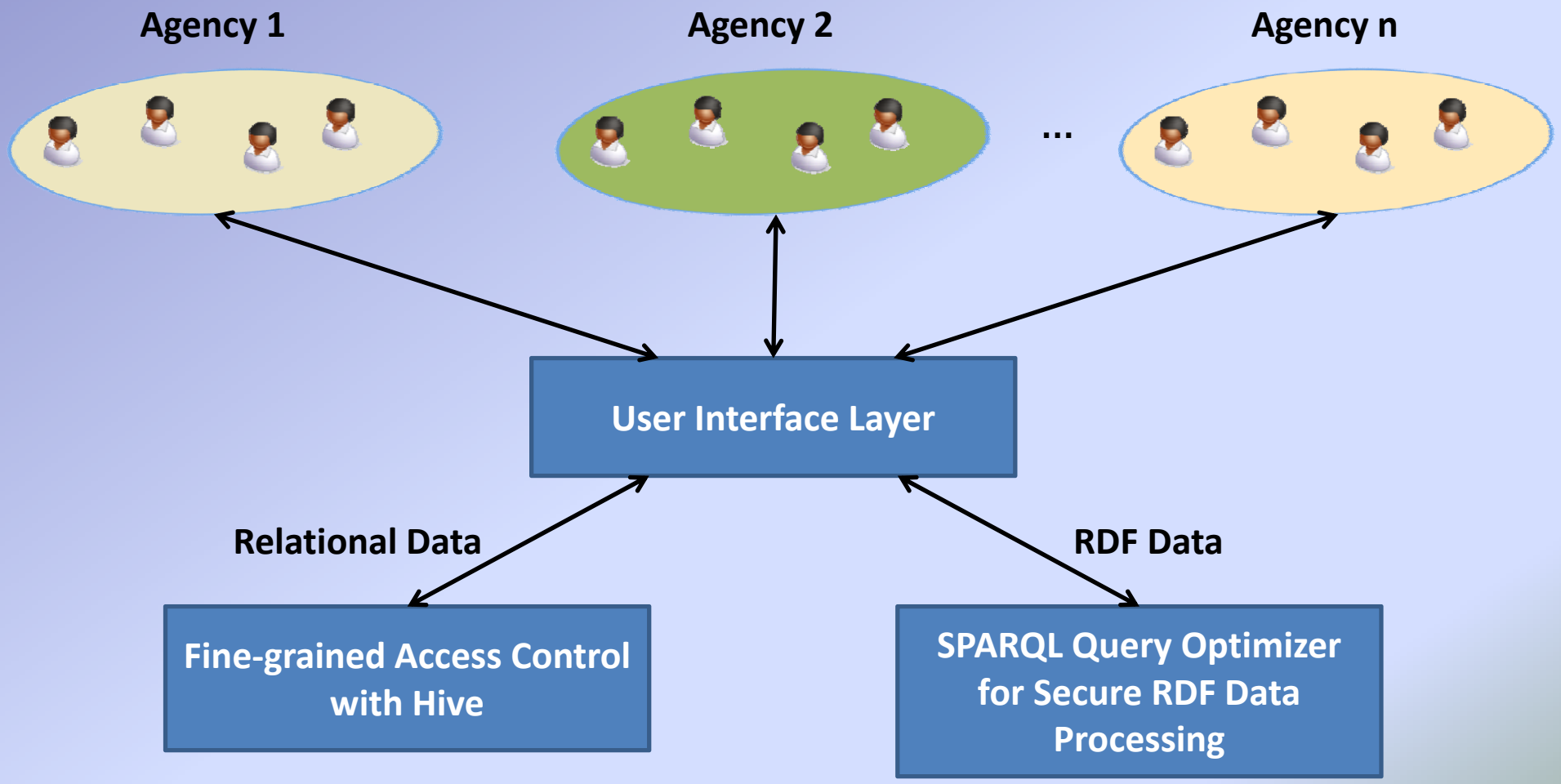


To build an efficient storage mechanism using Hadoop for large amounts of data (e.g. a billion triples); build an efficient query mechanism for data stored in Hadoop; Integrate with Jena

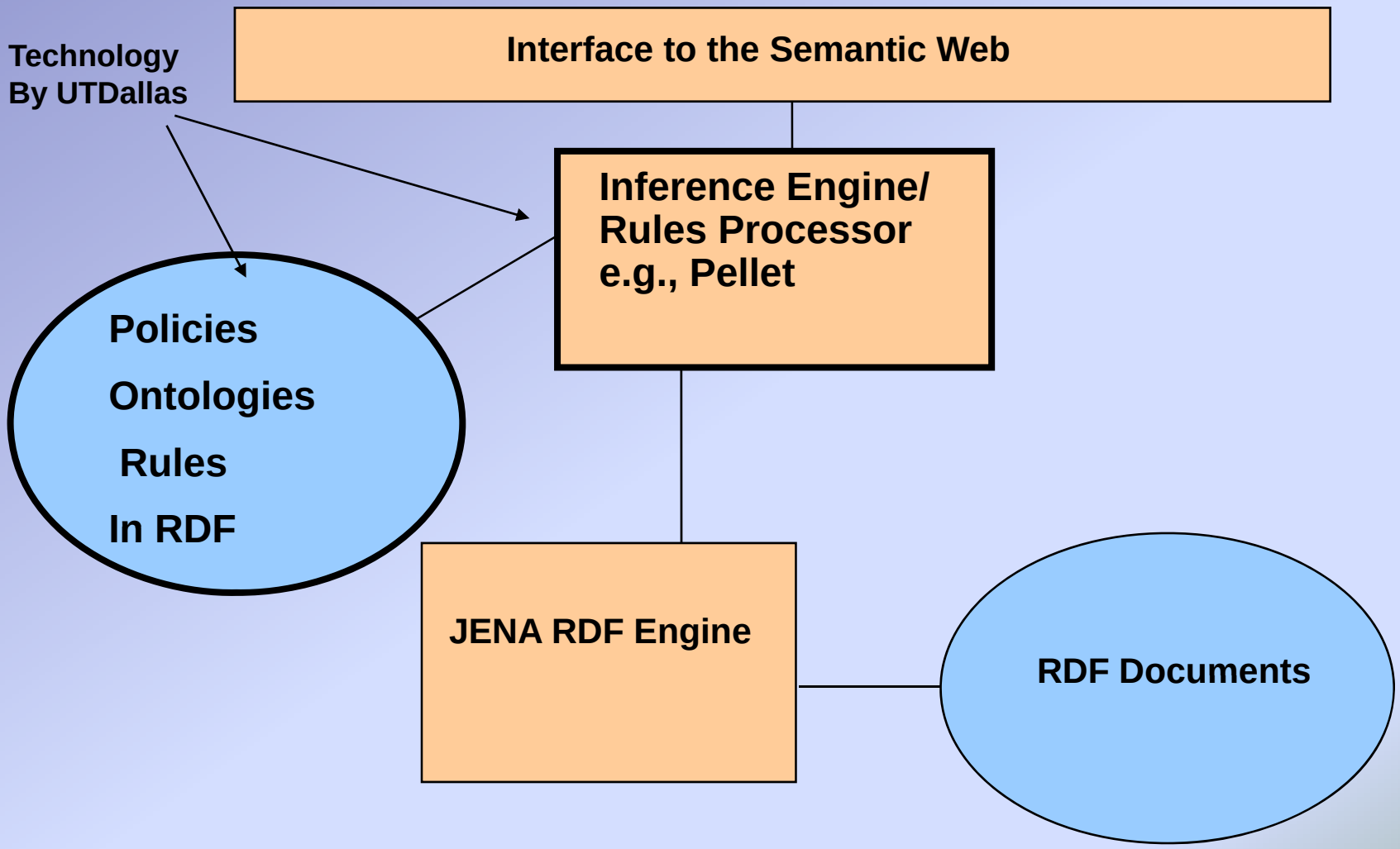
Developed a query optimizer and query rewriting techniques for RDF Data with XACML policies and implemented on top of JENA

IEEE Transactions on Knowledge and Data Engineering, 2011

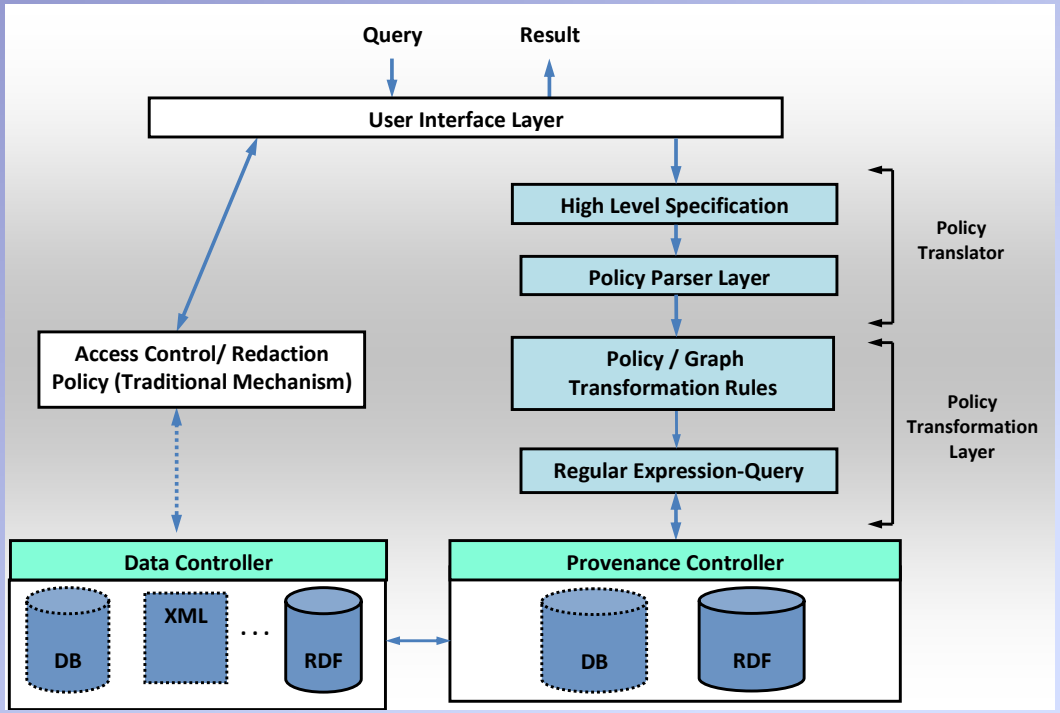
Demonstration: Concept of Operation



RDF-Based Policy Engine



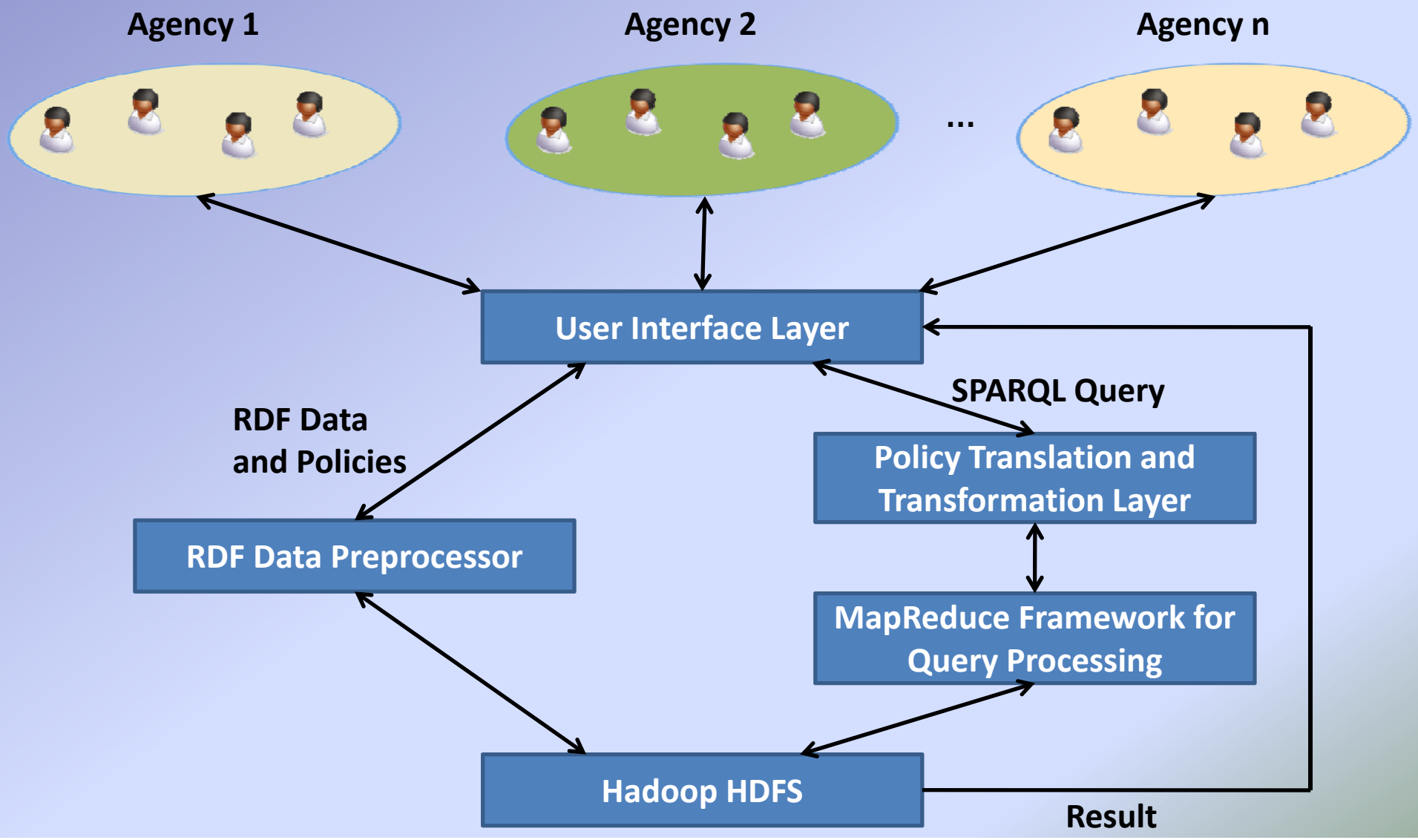
RDF-based Policy Engine on the Cloud



A testbed for evaluating different policy sets over different data representation. Also supporting provenance as directed graph and viewing policy outcomes graphically

- Determine how access is granted to a resource as well as how a document is shared
- User specify policy: e.g., Access Control, Redaction, Released Policy
- Parse a high-level policy to a low-level representation
- Support Graph operations and visualization. Policy executed as graph operations
- Execute policies as SPARQL queries over large RDF graphs on Hadoop
- Support for policies over Traditional data and its provenance
- IFIP Data and Applications Security, 2010, ACM SACMAT 2011

Integration with Assured Information Sharing:



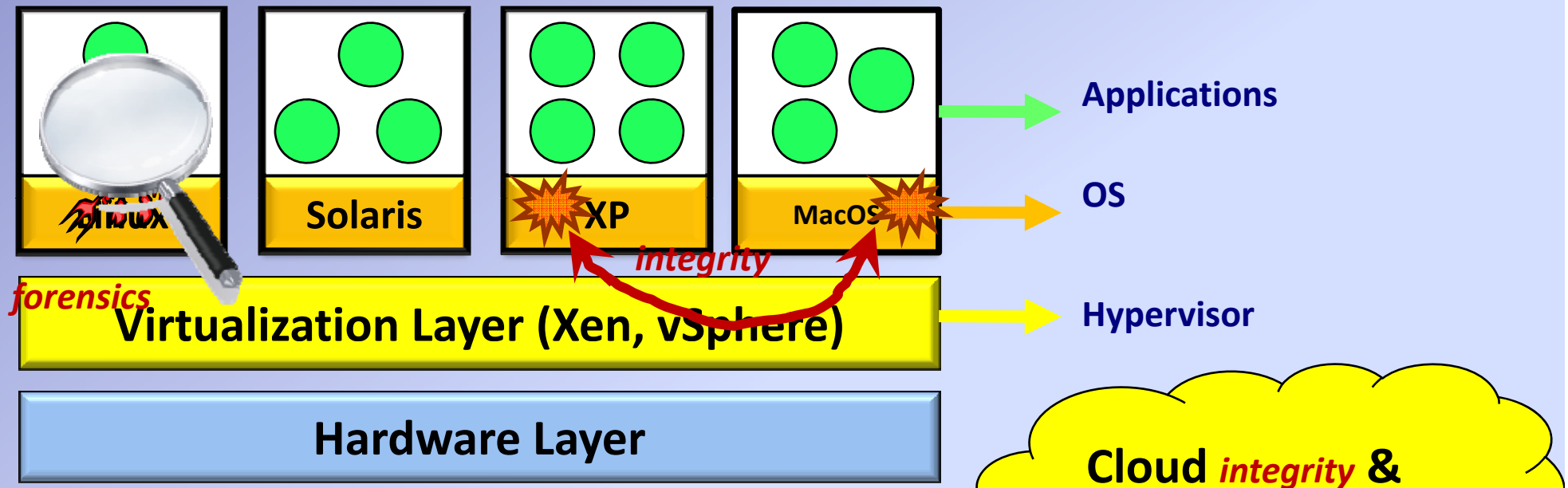
Types of Policies

- Agency 1 wishes to share its resources if Agency 2 also shares its resources with it
- Agency 1 asks Agency 2 for a justification of resource R2
- Agency 1 shares a resource with Agency 2 provided Agency 2 does not share with Agency 3
- Agency 1 shares a resource with Agency 2 depending on the content of the resource or until a certain time
- Agency 1 shares a resource R with agency 2 provided Agency 2 does not infer sensitive data S from R (inference problem)
- Agency 1 shares a resource with Agency 2 provided Agency 2 shares the resource only with those in its organizational (or social) network

Secure Storage and Query Processing in a Hybrid Cloud

- The use of hybrid clouds is an emerging trend in cloud computing
  - Ability to exploit public resources for high throughput
  - Yet, better able to control costs and data privacy
- Several key challenges
  - Data Design: how to store data in a hybrid cloud?
    - Solution must account for data representation used (unencrypted/encrypted), public cloud monetary costs and query workload characteristics
  - Query Processing: how to execute a query over a hybrid cloud?
    - Solution must provide query rewrite rules that ensure the correctness of a generated query plan over the hybrid cloud

Hypervisor integrity and forensics in the Cloud

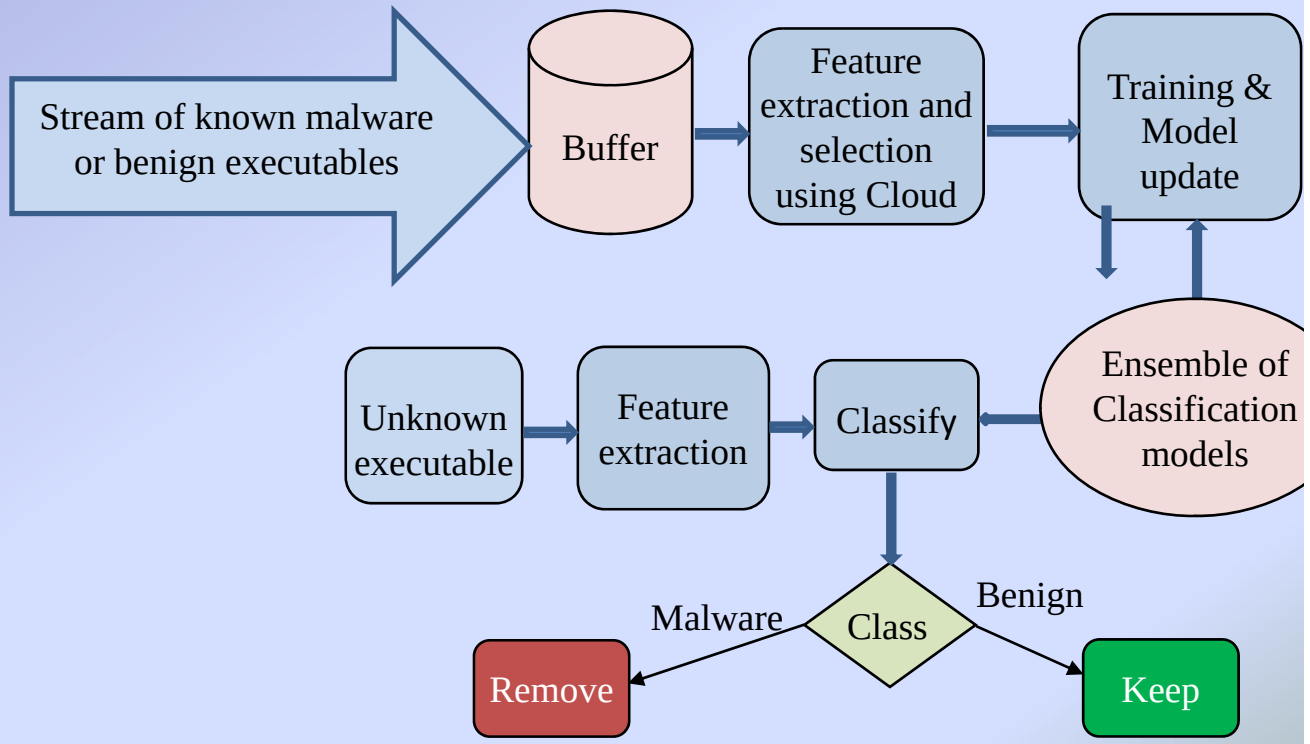


- Secure control flow of hypervisor code
  - Integrity via in-lined reference monitor
- Forensics data extraction in the cloud
  - Multiple VMs
  - De-mapping (isolate) each VM memory from physical memory

Cloud/Big Data for Malware Detection

Binary feature extraction involves
- Enumerating binary *n*-grams from the binaries and selecting the best *n*-grams based on information gain
- For a training data with 3,500 executables, number of distinct 6-grams can exceed 200 millions
- In a single machine, this may take hours, depending on available computing resources – not acceptable for training from a stream of binaries
- We use Cloud to overcome this bottleneck

A Cloud Map-reduce framework is used to extract and select features from each chunk
- A 10-node cloud cluster is 10 times faster than a single node
- Very effective in a dynamic framework, where malware characteristics change rapidly



Identity Management Considerations in a Cloud

- Trust model that handles
  - (i) Various trust relationships, (ii) access control policies based on roles and attributes, iii) real-time provisioning, (iv) authorization, and (v) auditing and accountability.
- Several technologies are being examined to develop the trust model
  - Service-oriented technologies; standards such as SAML and XACML; and identity management technologies such as OpenID.
- Does one size fit all?
  - Can we develop a trust model that will be applicable to all types of clouds such as private clouds, public clouds and hybrid clouds Identity architecture has to be integrated into the cloud architecture.

Directions

- Secure VMM and VNM
  - Designing Secure XEN VMM
  - Developing automated techniques for VMM introspection
  - Determine a secure network infrastructure for the cloud
- Integrate Secure Storage Algorithms into Hadoop
- Identity Management in the Cloud
- Secure cloud-based Social Networking / Big Data Management

Secure Social Networking in the Cloud

Part I: Location Mining from Online Social Networks

Part IIA: Preventing the Inference of Private Attributes

Part IIB: Access Control in Social Networks

Analyzing Social Networks

- Social networks are analyzed for several applications
  - Determining the strength of a friendship
  - Predicting future friendships
  - Clustering groups with similar interests
  - Determining hidden associations
  - Predicting demographics information
- Location is important for the following reasons
  - Privacy and Security
  - Trustworthiness
  - Location Driven Mining for Business
  - Location-Based Social Networking to generate US $21.14 billion by 2015[1]
  - But only ~14.3% provide it explicitly[2]

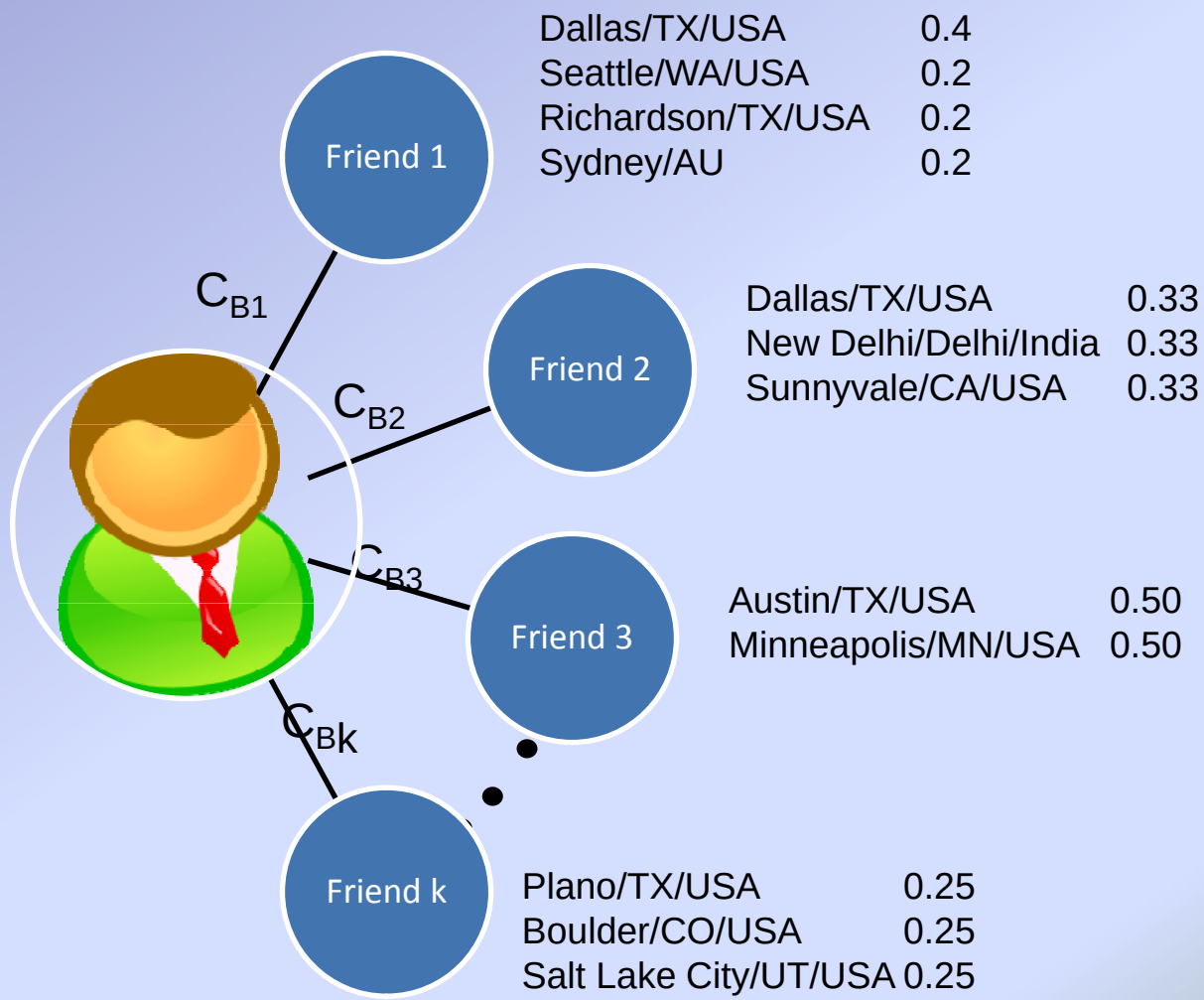
1 According to New Report by Global Industry Analysts, Inc., (GIA) (http://www.strategyR.com/)
2 According to an experiment performed by us on 1 million users

Tweethood: Fuzzy k-Closest Friends with Variable Depth

- Choose k “closest” friends for the user
- If location is not found look further for the answer
- Each node is defined by a vector having locations with their respective probabilities
- Boost and Aggregate at each step

Satyen Abrol, Latifur Khan, “*TweetHood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining*”. In Proc. of the Second IEEE International Conference on Social Computing (SocialCom-2010), Minneapolis, USA, August 20-22, 2010

Location Vector for John Doe's friends



Privacy of Social Networks: Our Approach

- Graph Model
  - Graph represented by a set of homogenous vertices and a set of homogenous edges
  - Each node also has a set of Details, one of which is considered *private*.
- Analysis
  - Apply variety of data mining techniques to determine whether private attributes can be inferred

Experiments

- 167,000 profiles from the Facebook online social network
- Restricted to public profiles in the Dallas/Fort Worth network
- Over 3 million links
- Conducted on 35,000 nodes which recorded political affiliation
- Tests removing 0 details and 0 links, 10 details and 0 links, 0 details and 10 links, and 10 details and 10 links

General Data Properties

Diameter of the largest component	16
Number of nodes	167,390
Number of friendship links	3,342,009
Total number of listed traits	4,493,436
Total number of unique traits	110,407
Number of components	18
Probability Liberal	.45
Probability Conservative	.55

Inference Methods

- Details only: Uses Naïve Bayes classifier to predict attribute
- Links Only: Uses only the link structure to predict attribute
- Average: Classifies based on an average of the probabilities computed by Details and Links
- Future research will include additional inference methods

Most Liberal Traits

Trait Name	Trait Value	Weight Liberal
Group	legalize same sex marriage	46.16066789
Group	every time i find out a cute boy is conservative a little part of me dies	39.68599463
Group	equal rights for gays	33.83786875
Group	the democratic party	32.12011605
Group	not a bush fan	31.95260895
Group	people who cannot understand people who voted for bush	30.80812425
Group	government religion disaster	29.98977927

Most Conservative Traits

Trait Name	Trait Value	Weight Conservative
Group	george w bush is my homeboy	45.88831329
Group	college republicans	40.51122488
Group	texas conservatives	32.23171423
Group	bears for bush	30.86484689
Group	kerry is a fairy	28.50250433
Group	aggie republicans	27.64720818
Group	keep facebook clean	23.653477
Group	i voted for bush	23.43173116
Group	protect marriage one man one woman	21.60830487

Online Social Networks Access Control

- Current access control systems for online social networks are either too restrictive or too loose
  - "selected friends"
    - Bebo, Facebook, and Multiply.
  - "neighbors" (i.e., the set of users having musical preferences and tastes similar to mine)
    - Last.fm
  - "friends of friends"
    - (Facebook, Friendster, Orkut);
  - "contacts of my contacts" (2nd degree contacts), "3rd" and"4th degree contacts"
    - Xing

Our Approach

- We use semantic web technologies (e.g., OWL) to represent social network knowledge base and semantic web rule language (SWRL) to represent various security, admin and filter policies.
- Existing ontologies such as FoAF could be extended to capture user profiles.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix fb: <http://example.org/facebook> .
<http://www.facebook.com/profile.php?id=999999999>
foaf:name "John Smith" .
<http://www.facebook.com/profile.php?id=999999999>
fb:lookingFor "Friendship" .
```

- Relationship among resources could be captured by using OWL concepts
  - PhotoAlbum rdfs:subClassOf Resource
  - PhotoAlbum consistsOf Photos

Security Policies for On-Line Social Networks (OSN)

- Security Policies ate Expressed in SWRL (Semantic Web Rules Language) examples

	SWRL rule
(1)	$Video(?targetObject,) \wedge ParentOf(Bob,?controlled) \Longrightarrow PRead(?controlled,?targetObject)$
(2)	$Owner(Bob,?targetObject) \wedge Photo(?targetObject) \wedge Friend(Bob,?targetSubject) \Longrightarrow Read(?targetSubject,?targetObject)$
(3)	$Photo(?targetObject) \wedge photoOf(Alice,?targetObject) \wedge Friend(Alice,?targetSubject) \Longrightarrow Read(?targetSubject,?targetObject)$
(4)	$Photo(?targetObject) \wedge Owns(?owner, ?targetObject) \wedge Friend(?owner, ?targetSubject1) \wedge Friend(?targetSubject1, ?targetSubject2) \Longrightarrow Read(?targetSubject2, ?targetObject)$

Security Policy Enforcement

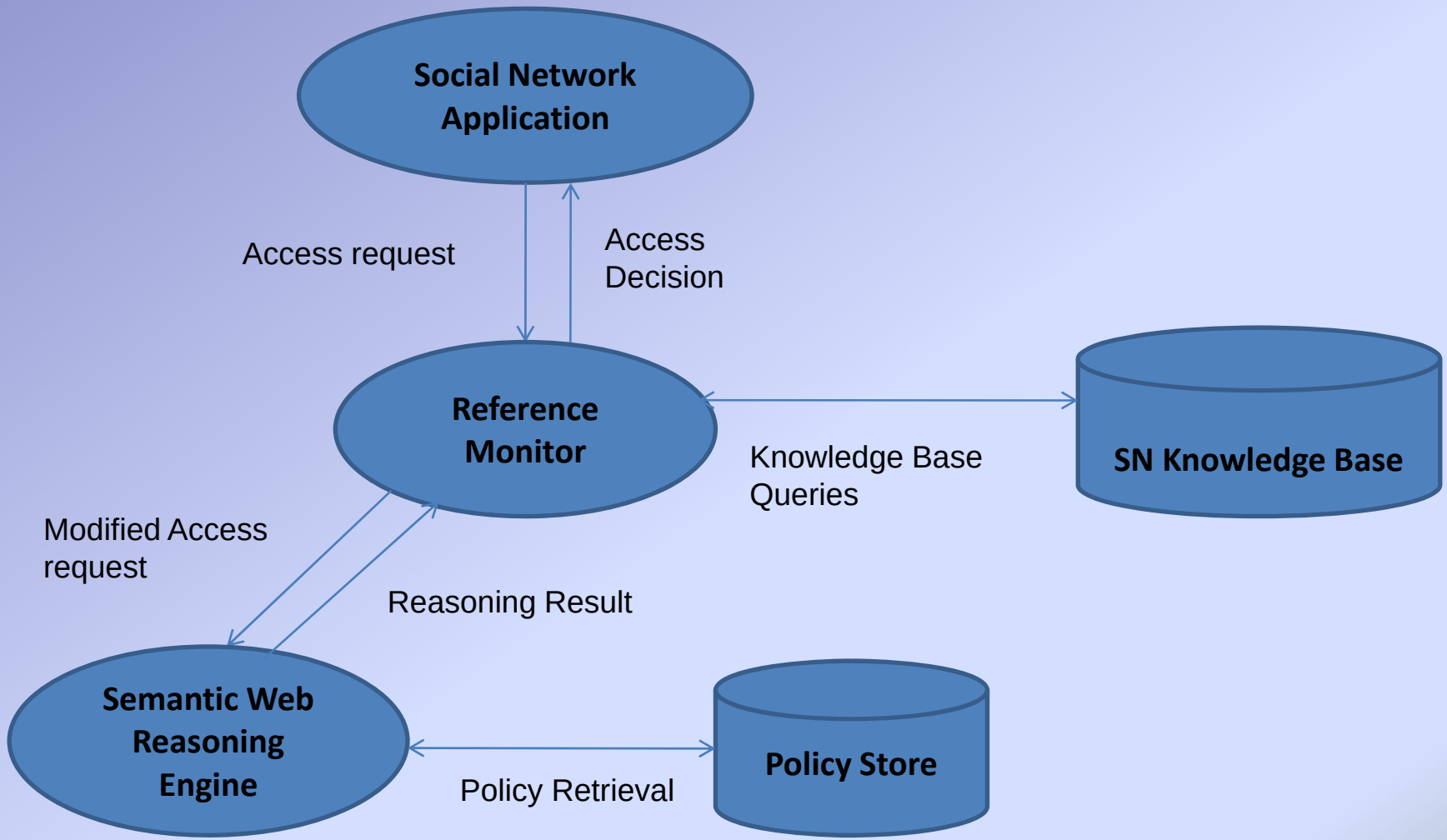
- A reference monitor evaluates the requests.
- Admin request for access control could be evaluated by rule rewriting
  - Example: Assume Bob submits the following admin request

$$SWRL_1: \ Owns(Bob,?targetObject) \wedge Photo(?targetObject) \wedge Friend(Bob,?targetSubject) \Longrightarrow Read(?targetSubject,?targetObject)$$

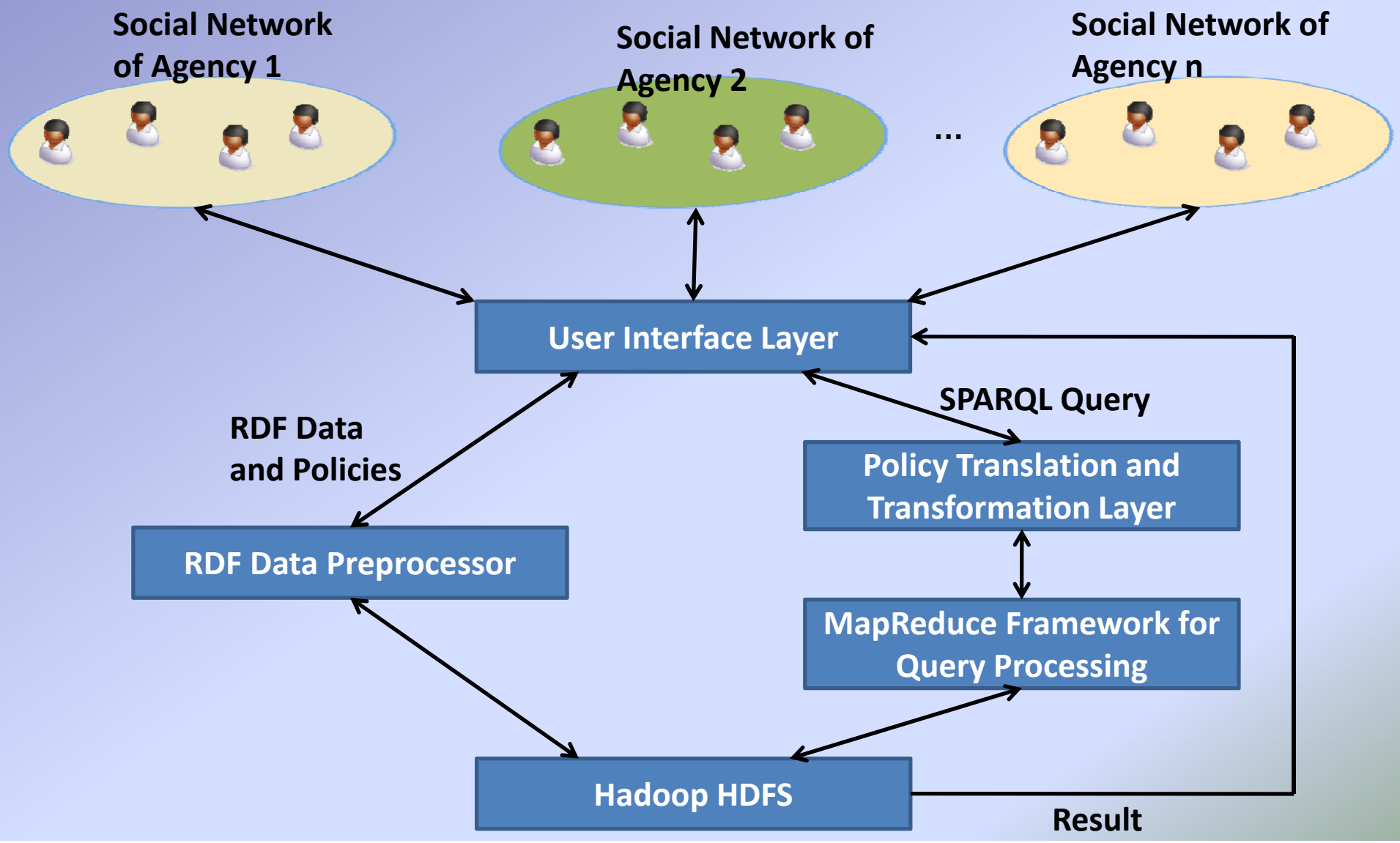
  - Rewrite as the following rule

$$New\_SWRL_1: \ AdminRead(Bob,?targetObject) \wedge Owns(Bob,?targetObject) \wedge Photo(?targetObject) \wedge Friend(Bob,?targetSubject) \Longrightarrow Read(?targetSubject, ?targetObject)$$

Framework Architecture



Social Networking in the Cloud



Other Directions

- Various attacks exist to
  - Identify nodes in anonymized data
  - Infer private details
- Recent attempts to increase social network access control to limit some of the attacks
- Balancing privacy, security and usability on online social networks will be an important challenge
- Directions
  - Scalability
    - We are currently implementing such system to test its scalability.
  - Usability
    - Create techniques to automatically learn rules
    - Create simple user interfaces so that users can easily specify these rules.
  - **Big Data Security and Privacy, Secure Cloud and Internet of Things**

Security and Privacy for Big Data

- NSF Workshop on Big Data Security and Privacy
- Secure Storage and Infrastructure
  - How can technologies such as Hadoop and MapReduce be Secured
- Secure Data Management
  - Techniques for Secure Query Processing
- Big Data for Security
  - Analysis of Security Data (e.g., Malware analysis)
- Regulations, Compliance Governance
  - What are the regulations for storing, retaining, managing, transferring and analyzing Big Data
  - Are the corporations compliance with the regulations
  - Privacy of the individuals have to be maintained not just for raw data but also for data integration and analytics
  - Roles and Responsibilities must be clearly defined
- **Secure Internet of Things (Next Steps)**

Security and Privacy for Internet of Things (Bertino, Kantarcioglu, Thuraisingham)

Mobile devices such as smart phones have rapidly become an extremely prevalent computing platform, with just over a billion smart phones

Ever increasing popularity in smartphone apps used in a multitude of applications including the monitoring of personal data such as health, food intake, exercise and sleep patterns, among others.

With the recent emergence of *Quantified Self* (*QS*) movement, such personal data collected by the various devices (e.g., wearable devices and smart phone apps) are being analyzed to give guidance to the user

Data collection and sharing are also being carried out often without the knowledge of the user.

To address this ever increasing challenge caused by the digitized world that we live in today, we are developing tools and techniques to enforce policies that are guided by the regulations to use such personal data in a privacy enhanced manner.

Contact

- Dr. Bhavani Thuraisingham
- bhavani.thuraisingham@utdallas.edu
- @CyberUTD

- Ms. Rhonda Walls
- rhonda.walls@utdallas.edu