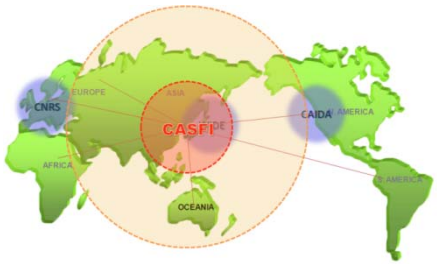
CASFI Data-sharing Platform

Seoyeon Kang
Haewoon Kwak
Keon Jang
Sue Moon

August 15th, 2008
1st CAIDA-WIDE-CASFI Workshop



Outline

- Review of existing data-sharing platforms
- Our requirements for data sharing
- Design choices
- CASFI data-sharing platform design

Review of existing data-sharing platforms

- DatCat
- CRAWDAD
- UMASS trace repository
- WITS

DatCat *http://datcat.org/*

- Serviced by CAIDA
  - Internet measurement data catalog
- Organizes real-world objects
  - Collection, package, data, publication, …
- Functions as a mediator between owner & user
  - Store metadata only

CRAWDAD *http://crawdad.cs.dartmouth.edu/*

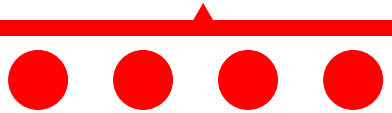
- Serviced by Dartmouth
  - Community resource for archiving wireless data
- Four categories of resources
  - Data(dataset, traceset, trace), tools, authors, papers
- Access rights in upload / download
  - Upload: Contact CRAWDAD
  - Download: Anyone

UMass trace repository *http://traces.cs.umass.edu/*

UMassTraceRepository

- Serviced by UMass
  - Network, storage, DTN traces, etc.
- Access rights in upload / download
  - Upload: Submit files w/ brief info
  - Download: Anyone

WITS *http://www.wand.net.nz/wits/*



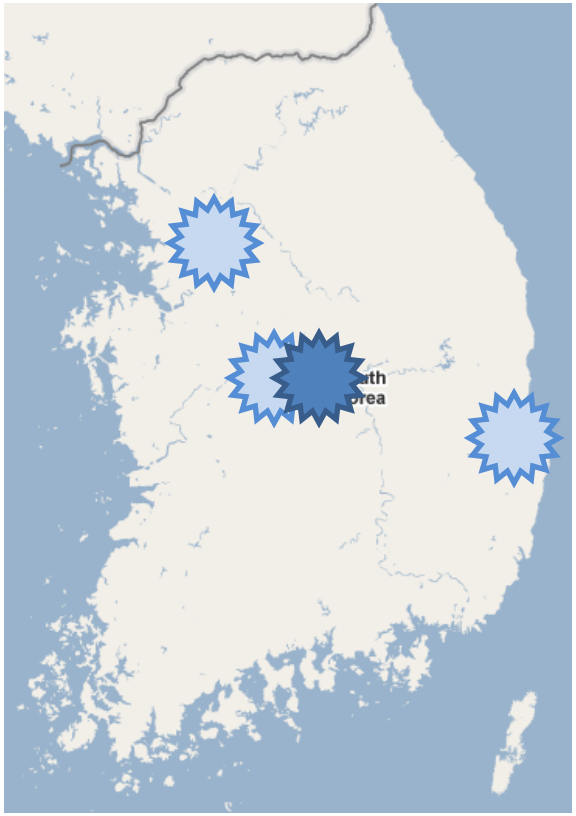
- Waikato Internet Traffic Storage project
  - All the Internet traces of WAND
- Access rights
  - Will be made public in a near future

Requirements of our platform

- Physical constraints
- Korean Law
- Types of data sources
- Our goal

Physical Constraints

- Geographically distributed
  - (10TB NAS x 2) x
    (4 physically distributed areas)
  - KAIST, Chungnam National U,
    POSTECH, Kyunghee U



Korean Law

- Regulation to capture and distribute data
  - Do NOT capture full packet traces
  - Do NOT distribute even header-only data
- But we can publish statistics
  - Reveal no privacy information

Types of Data

- Not only packet traces but application data
  - Standard format: packet traces (.erf, .pcap)
  - Free format: Traceroute, BGP snapshots,
    Crawled data (Cha07 YouTube data),
    Cyworld data (Ahn07 & Chun08 data)

Our Goal

- Build a data-sharing platform that …
  - Is open to researchers
  - Operates in a legally responsible manner

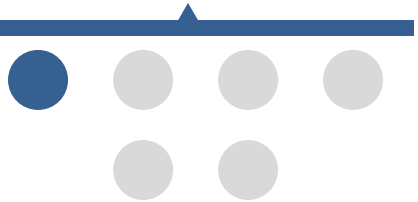
Design Considerations

- Make user interface simple and hide system details
- Retain legal administrative responsibility
- Offer processing capability
- Support consistent metadata mgmt of free-format data
- Allow fine-grained access control

CASFI Data-Sharing Platform Design

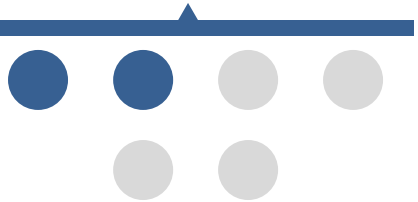
- Present a “Facade“
- Allow script execution
- Maintain flexibility in data description
- Use Role-Based Access Control (RBAC)

Present a “Facade"



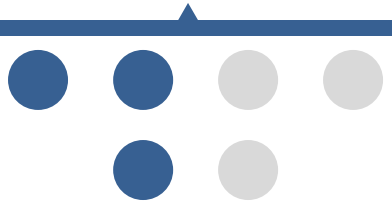
- Users contact only front-end servers
- Front-end servers communicate with back-ends
- Front-end servers provide web-based tools for
  - Search, upload, and even download data

Allow script execution



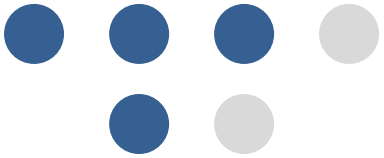
- Broad research areas to access secure data
  - NEON, HiStar, Raksha, Secure information flow, …
- We provide ‘primitives’ to access secure data

Development plan for primitives



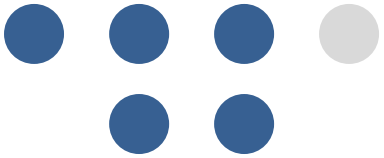
- Use daemon on front-end servers
  - A user send a request (method, params) to Daemon
  - Daemon passes a request to back-end servers
  - Daemon waits for a response from back-end servers and returns it to the user
- Primitive roll-out plan
  - Ver 1.0: SUM, AVG, MAX, MIN, MEDIAN
  - Ver 1.1: GROUP BY
  - Ver 2.0: CDF & CCDF

Maintain flexibility in data description



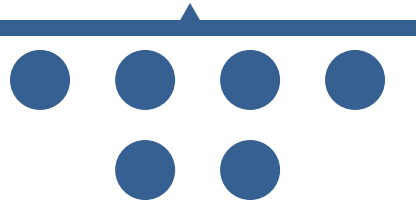
- Many tools to extract metadata from standard format data
  - No extra work needed
- Few tools for free format data
  - PADS (Fisher05)
    - Generate PADS desc. & C library for parsing data
    - Not human readable descriptions
  - Our goal is to assist in metadata input & description

Web-assisted data description generation



- Semi-automate data description generation
  1. Infer delimiters and data type of each column: date, IP, int, string, …
  2. Users set secure level of each column
  3. Users can add miscellaneous metadata
- Keep data description in XML
  - Flexible representation, manipulation, and expansion of data description
  - Easy backward compatibility

Role-based access control



- A role maps to permissions for a set of operations
- A user can have multiple roles
- Roles are hierarchical
- Multiple operations are defined per resource
  - Create, modify, delete, …

Summary

- Overview of CASFI data-sharing platform
  - Background
  - Our goal
  - Our solutions
- We plan to implement the design step-by-step

References

- Fisher05
  http://www.padsproj.org/papers/pldi.pdf
- Role-based access control
  http://csrc.nist.gov/groups/SNS/rbac
- Raksha
  http://csl.stanford.edu/~christos/publications/2007.raksha.isca.pdf
- Mycha07
  http://an.kaist.ac.kr/traces/papers/imc131-cha.pdf
- Ahn07
  http://an.kaist.ac.kr/~sbmoon/paper/intl-conf/2007-www-social-networking.pdf