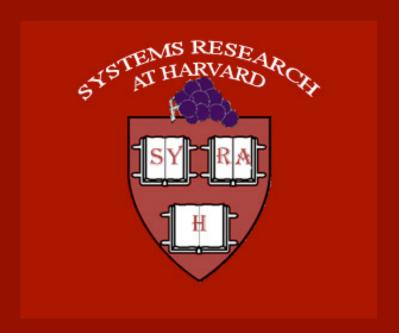# Provenance as First-Class Cloud Data
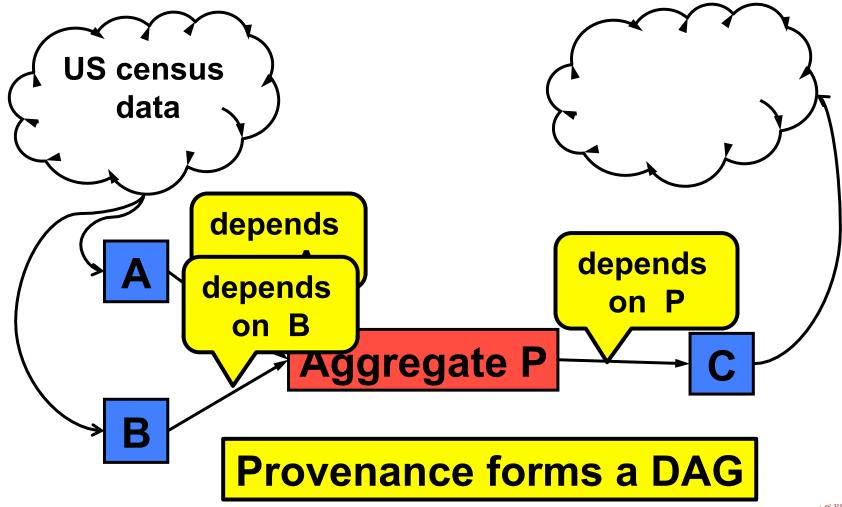
Kiran-Kumar Muniswamy-Reddy

and Margo Seltzer

# What is Provenance?

- ## Meta-data describing the history of an object
  - What objects does this object depend on?
  - What applications modified/generated this object?

# What is Provenance? (2)



US census data

depends

A

depends on B

Aggregate P

depends on P

C

B

Provenance forms a DAG

# What is Provenance? (3)

- **Useful in various domains**
  - Scientific reproducibility
  - Business compliance
  - Security

# Why support provenance in the cloud?

- ## Users want it

- ## Provides information regarding structure of data and applications

  - Search
  - Pre-fetching
  - Access Control
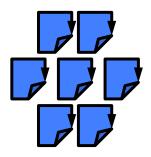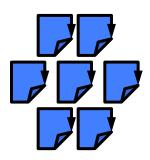  - Detect Application Anomalies

# Search

- As more data moves to the cloud, search will become an important problem
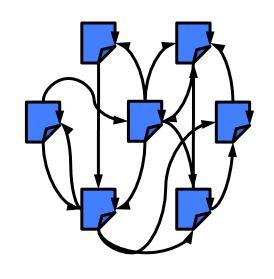- Some similarities to web-search

# Web Search vs. Cloud Search

# Web Search vs. Cloud Search



**Links to refine search**

**?**

**Provenance dependencies to refine search [shah-usenix'07]**

# Pre-fetching

- Lots of prior research on identifying related objects for pre-fetching

- Provenance DAG readily has this information
  - Aggregate provenance across multiple user data sets to build more accurate models

# Access Control

- **Provenance can provide good default policies**

- **Access control is usually based on ACLs**

- **Often process used to produce results provides hints about the data**

  - Data derived from "ultra-sound" process needs to be confidential

  - Data derived from "aggregation" can be public

# Detect Application Anomalies

- **User/application errors can overuse cloud resources**

- **Provider cloud alert users to overuse**

- **Provenance can help**
  - Provides a model of "normal" usage/behavior

# Requirements

- **Consistency**
- **Long-term persistence**
- **Queryable**
- **Security [Braun-hotsec'08]**
- **Coordinate storage and compute facilities**
  - Ex: Provenance-Aware Storage Systems (PASS) [Muniswamy-Reddy-usenix06]

# Related Work

- **IBM Haifa VISION project**
  - Use Object's meta-data for management
  - http://www.snia.org/cloud/JulyMiniSummit/VISION_for_SNIA_Cloud_minisummit_July_23_2009.PDF

# Conclusions

- **Provide native support for provenance**
  - Users want it
  - You want it

# Questions?

Contact:

kiran@eecs.harvard.edu

[www.eecs.harvard.edu/~kiran](www.eecs.harvard.edu/~kiran)