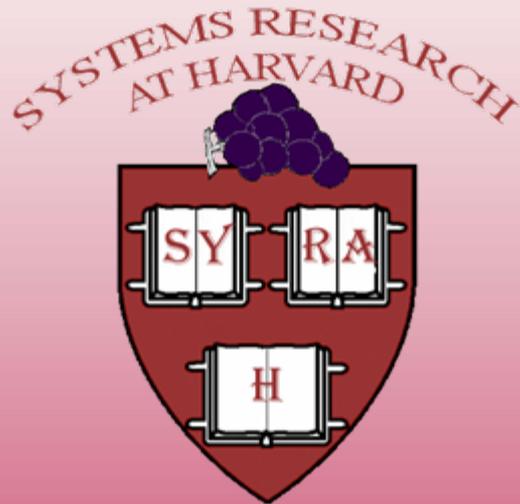


Provenance: Meta-data or Not?



August 4, 2008

Margo Seltzer

Harvard University

School of Engineering and Applied Sciences



Provenance

- From the French word for “source” or “origin”
- The complete history or lineage of a object
- In computer terms:
 - On what is this object based?
 - How was this object created?
 - How can it be re-created?
- Examples
 - Source code control
 - make



Applications of Provenance

- Homeland security
- Archival
- Science
- Business compliance
- Software development
- Uncertainty



Provenance and the Storage System

- Most provenance solutions are user-level.
- PASS takes a systems approach
 - Provenance is collected and stored by the file system.
 - Generated and maintained transparently.
 - Indexed and queried.
 - Maintained in the presence of deletes, copies, renames, etc.
- Why provenance in the storage system?
 - Can generate automatically.
 - Requires effort to subvert.
 - Can be bound tightly with data.
 - Subsumes other system data collection (e.g. logging).



Cool Applications

- Understand system dependencies (mtab)
- Intrusion detection (~-/garbage)
- Detecting system change (new libraries)
- Script generation (get me from a to b)
- Build debugging (missing dependency)
- ... lots more!



Provenance is Meta-data

- Some existing meta-data is provenance:
 - Owner/Group
 - Create time/mod time
- Characteristics
 - Stored in per-file (object) file system structures
 - Available via standard network protocols
 - Rarely searched/queried
 - Usually needs to be in-memory when file is accessed



Provenance as Extended Attributes

- Examples of EA-like provenance
 - Environment
 - Argument vector
 - User-supplied annotations
- Represented as key/data pairs.
- Unlimited vocabulary
- Not standardized
- Characteristics
 - Not necessary to cache while file accessed
 - Unstructured
 - Can be user/application-defined



Provenance as Search Indexes

- Used to answer searches or queries:
 - Find me files “like” this one.
 - Find me the file created by a particular process.
 - Tell me why these two objects differ.
- Characteristics
 - Requires query capabilities.
 - Accessed on a volume-wide basis (not per-file).
 - Requires indexes for good performance.



Provenance is Different

- Provenance forms a DAG.
- Search on provenance combines attribute lookup with graphical queries.
- Provenance is immutable.
- Provenance persists after the object it describes.
- Provenance has challenging security properties.



Meta-data Taxonomy

	Inode	EA	Search	Prov
Extensible	No	Yes	Yes	Yes
Access	F(directory)	F(file)	???	independent
Lifetime	File	File	File	Forever
Mutable	Not really	Yes	Yes	No
In-memory with data	Yes	No	No	No



Thank You!

- This project has been funded by:
 - Network Appliance
 - IBM
 - The National Science Foundation

Thank You!