Policy-Based Integration of Provenance Metadata

Ashish Gehani Dawood Tariq Basim Baig
SRI International

Tanu Malik
University of Chicago

Background

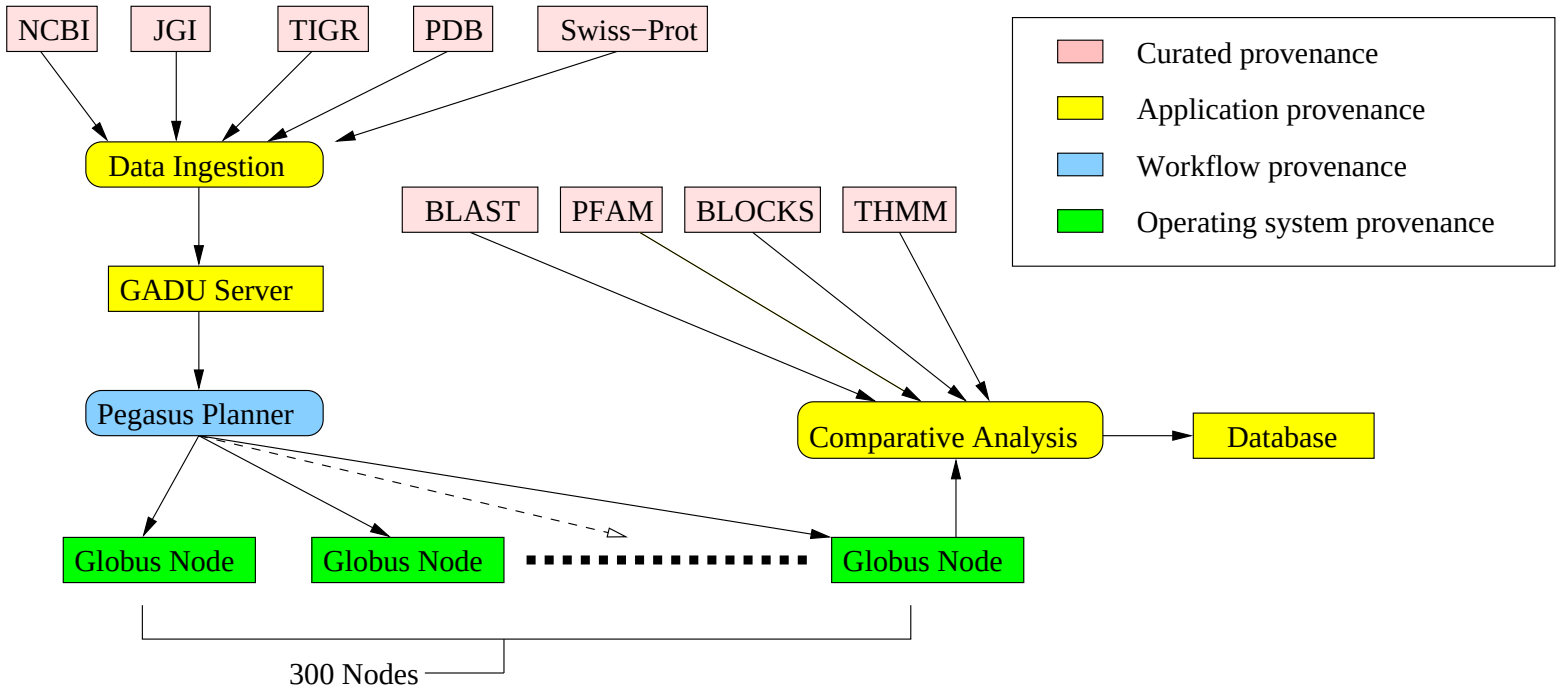
- Traditional science:
  - Human hypotheses
  - Experimental, theoretical verification
- Computational paradigm:
  - Automated exploration
  - Machine verification
- Provenance tracking need:
  - Reproducibility (execution context)
  - Data sharing (dependencies)
  - Scenario analysis (application profiles)

Open Provenance Model

- Vertex types
  - Artifact
  - Process
  - Agent
- Edge types
  - Used
  - WasGeneratedBy
  - WasDerivedFrom
  - WasControlledBy
  - WasTriggeredBy
- Domain semantics in annotations

Integration Need

- Genome analysis provenance record

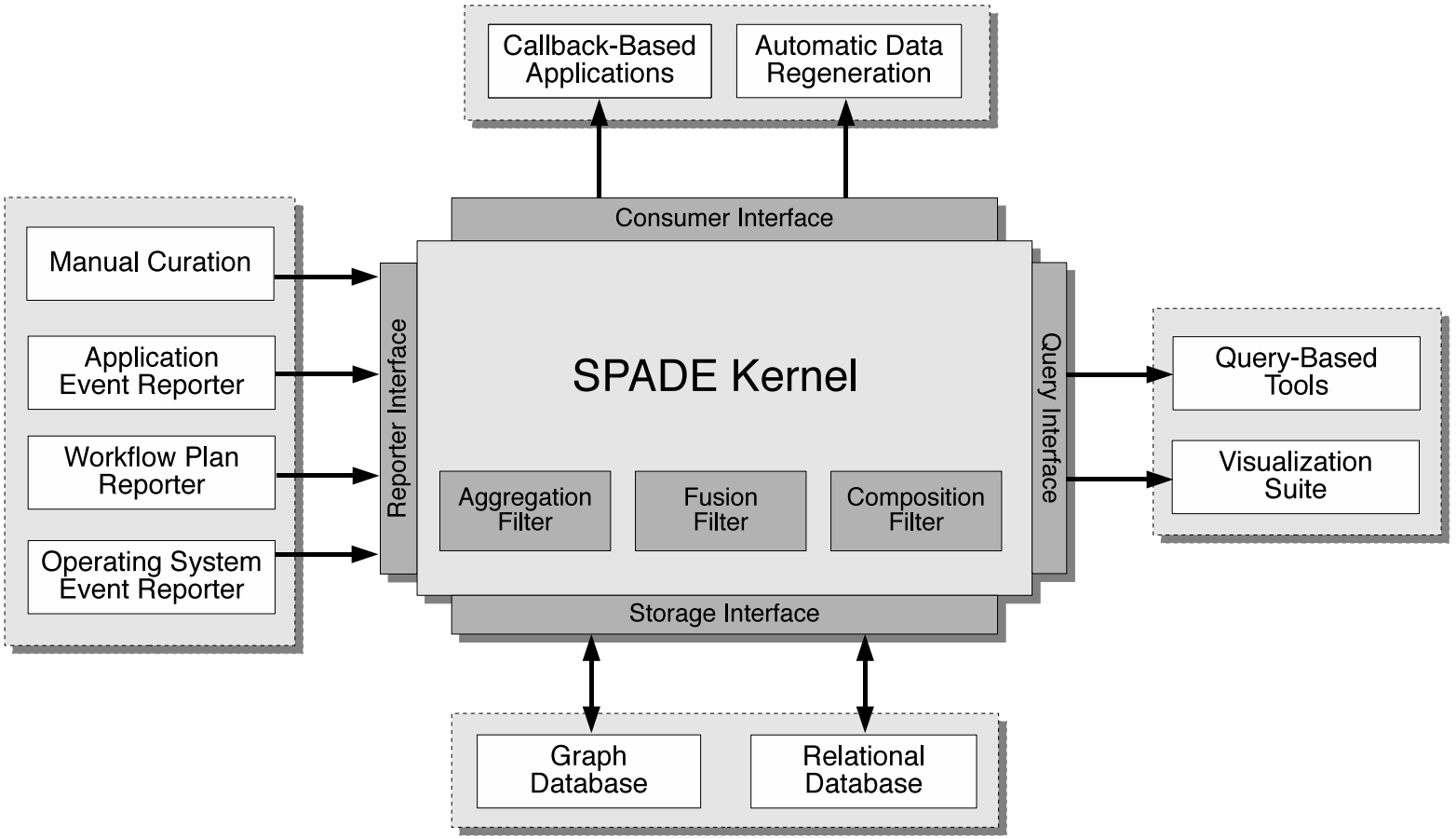


Integration Issues

- Metadata variation:
  - Abstraction levels
  - Completeness
  - Identifiers
  - Semantics
- Querying requires:
  - Record assembly
  - Reconciling syntax
  - Mapping semantics

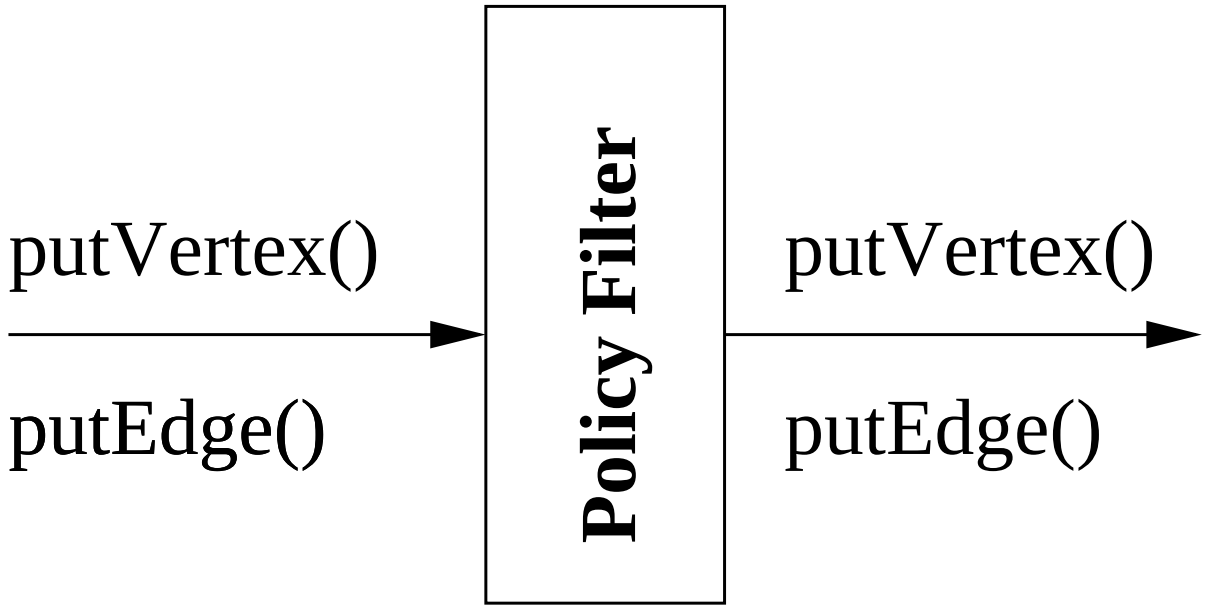
Provenance Middleware

- Support for Provenance Auditing in Distributed Environments



Implementing Policy

- Annotations are key-value pairs
- Filters operate on vertex / edge stream
- Arbitrary transformations possible



Sample Workload

- BLAST (sequence alignment)
- Influenza data

  `ftp://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/influenza.faa`

- Database construction

```
makeblastdb -in influenza.faa
-parse_seqids -hash_index -out outputdb
```

Aggregation Policies

Policy	Version created after:
ALL	Every write
SEQ	Each sequence of *read()* or *write()*
CA	Cycle avoidance (Harvard)
GF	Graph finesse (Harvard)
OC	*open() - close()* pair

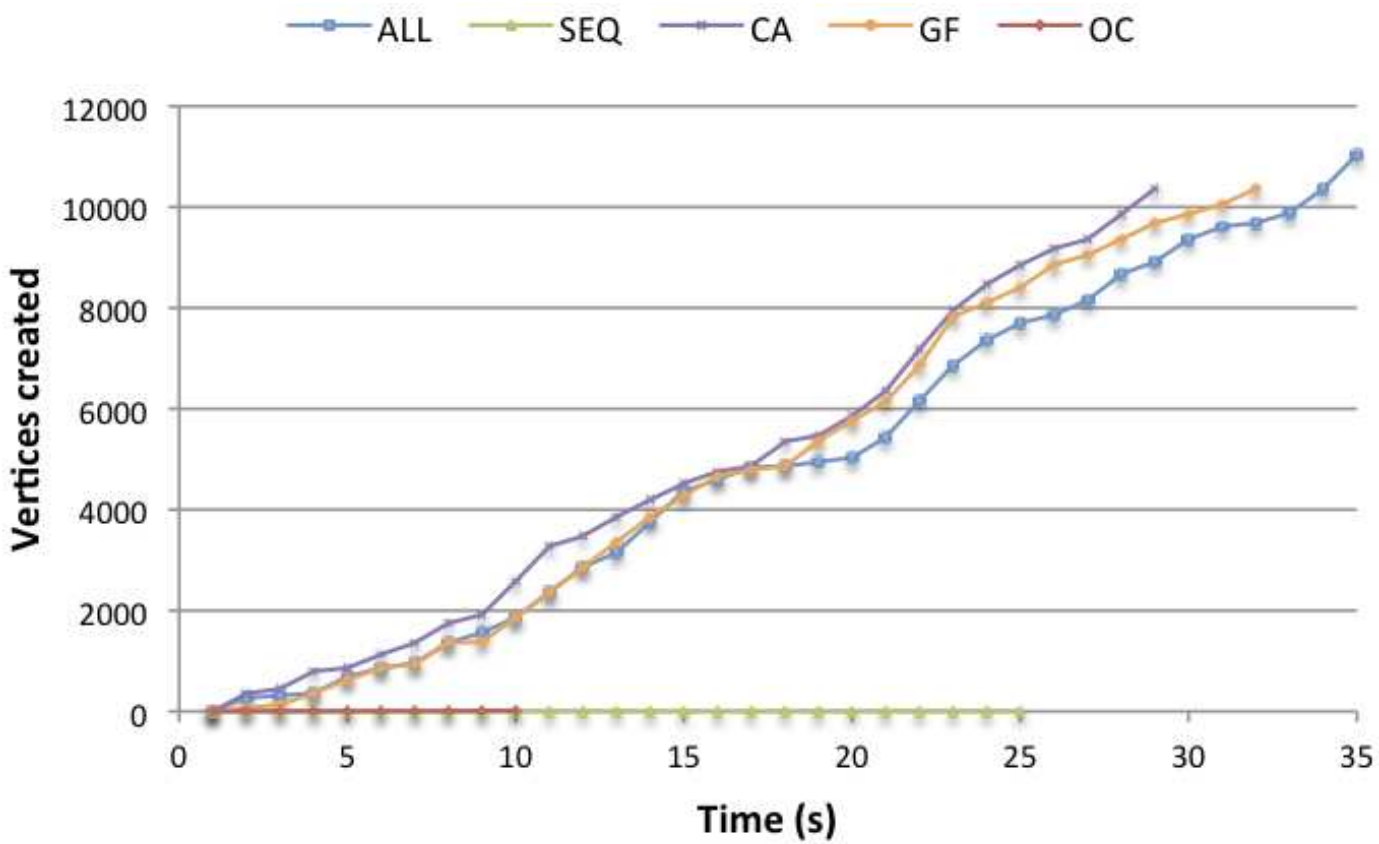
Policy Size

- Size of different *aggregation policy* filters.

Policy	Lines of code
ALL	38
SEQ	73
CA	63
GF	80
OC	6

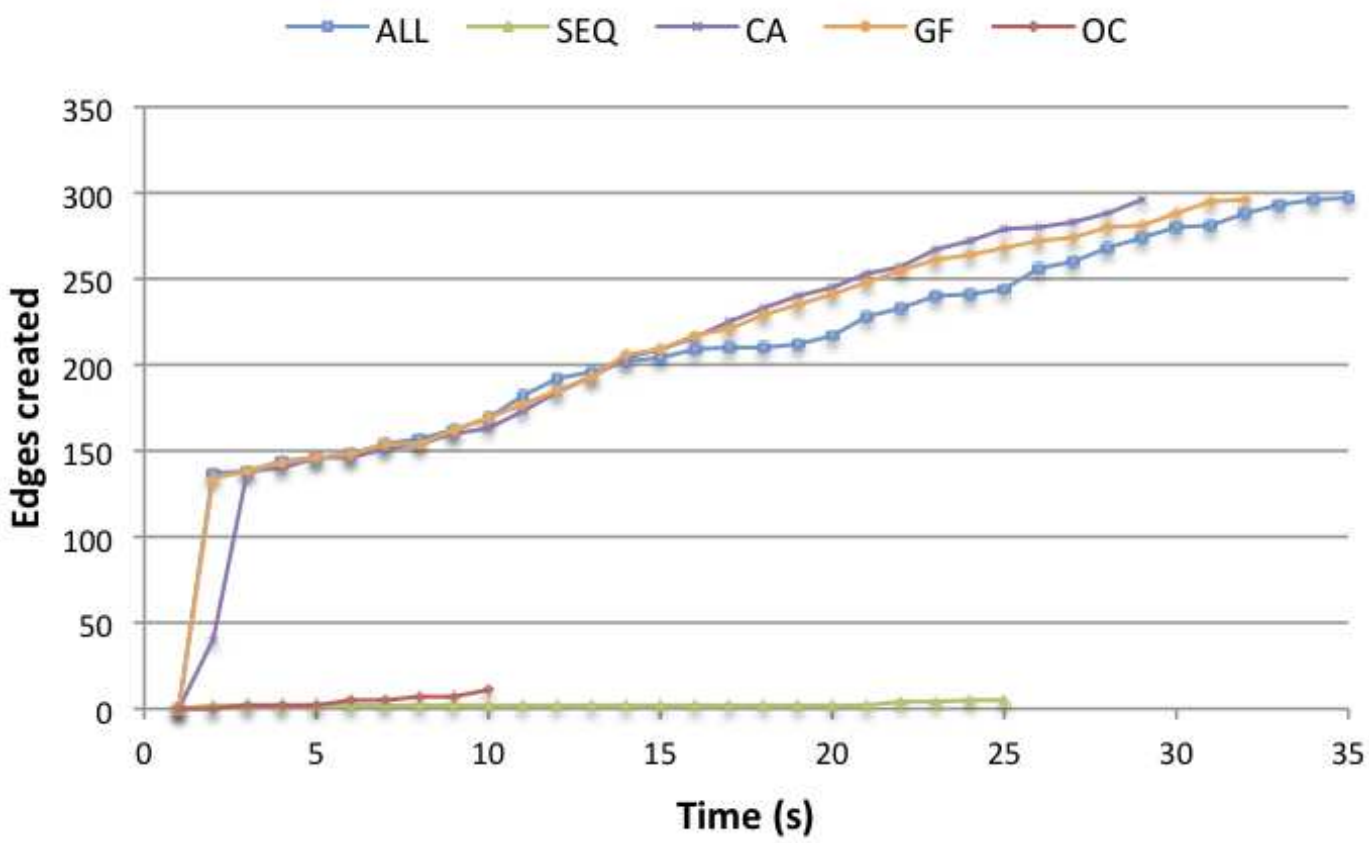
Artifact Vertices

- Effect of *aggregation policies* over time.



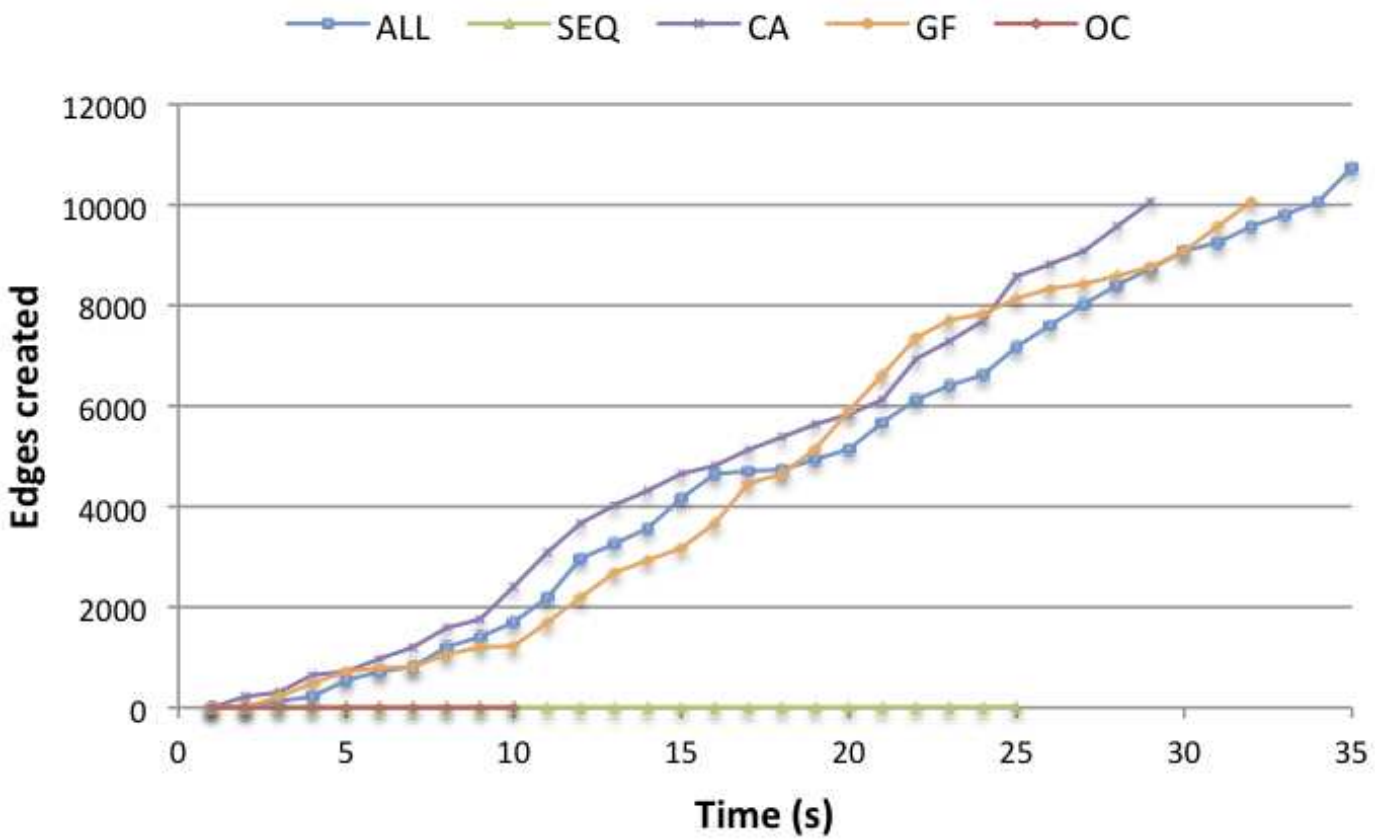
Used Edges

- Effect of *aggregation policies* over time.



WasGeneratedBy Edges

- Effect of *aggregation policies* over time.



Conclusion

- Provenance integration policy matters
- Substantial impact:
  - Runtime - CPU, memory
  - Persistent - storage, querying
- Acknowledgement
  - US NSF Grant OCI-0722068
- URL: http://spade.csl.sri.com
- Code license: GPLv3
- Email: ashish.gehani@sri.com
- Questions?