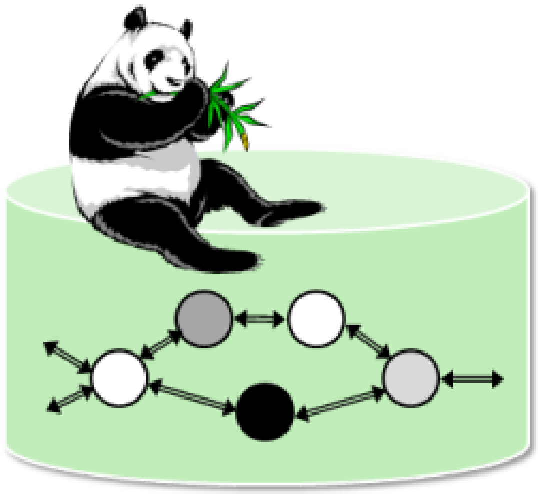

PANDA

A System for Provenance and Data

Overview

(Excerpted from December 2009 short overview paper)

In its most general form, *provenance* (also sometimes called *lineage*) captures where data came from, how it was derived, manipulated, and combined, and how it has been updated over time. Provenance can serve a number of important functions:

- **Explanation.** Users may be particularly interested in or wary of specific portions of a derived data set. Provenance supports "drilling down" to examine the sources and evolution of data elements of interest, enabling a deeper understanding of the data.
- **Verification.** Derived data may appear suspect -- due to possible bugs in data processing and manipulation, because the data may be stale, or even due to maliciousness. Provenance enables auditing how data was produced, either for verifying its correctness, or for identifying the erroneous or outdated source data or processing nodes that are responsible for erroneous or outdated output data.
- **Recomputation.** Having found outdated or incorrect source data, or buggy processing nodes, we may want to correct the errors and propagate the corrections forward to all "downstream" data that are affected. Provenance helps us recompute only those data elements that are affected by the corrections.

There has been a large body of very interesting work in lineage and provenance over the past two decades. Nevertheless, we believe there are still many limitations and open areas. Specifically:

1. Most work has been either: *data-based*, in which fine-grained provenance of data elements is tracked based on well-defined, transparent properties of data models and query languages; or *process-based*, in which coarse-grained provenance is tracked, typically involving workflows and data at the schema level.
2. Often the primary focus is on *modeling* and *capturing* provenance: How is provenance information

represented? How is it generated? There has been considerably less work on *querying* provenance: What can we do with provenance information once we've captured it?

3. Many projects have focused on specific functions or application domains, rather than developing a general provenance system that can be used for different purposes and across domains.

Our goal is to fill these gaps. Specifically, we want to:

1. Seamlessly merge data-based and process-based provenance, so that the two types of provenance can be combined (e.g., workflows that combine "opaque" processing nodes with well-understood relational queries and transformations). We also want to develop a model and system that offers users a full range from fine-grained to coarse-grained provenance.

2. Define a set of useful operators for taking advantage of provenance after it has been captured, as well as a general-purpose language for querying and analyzing provenance, and for combining provenance with relevant data.

3. Develop a general-purpose open-source system that is flexible and configurable enough to be used for a wide variety of applications. The system will support its own mechanisms for provenance capture, storage, operators, and queries, while also offering interfaces for coupling with outside data sources, processes, and systems.

The Panda project is supported by the National Science Foundation (grant IIS-0904497), the Boeing Corporation, KAUST, and an Amazon Web Services Research Grant.

Panda project Wiki

Papers

- R. Ikeda, J. Cho, C. Fang, S. Salihoglu, S. Torikai, and J. Widom Provenance-Based Debugging and Drill-Down in Data-Oriented Workflows. To appear in Proceedings of the 28th International Conference on Data Engineering, Washington, DC, April 2012. Demonstration description.

- R. Ikeda, S. Salihoglu, J. Widom. Provenance-Based Refresh in Data-Oriented Workflows. To appear in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM '11), Glasgow, Scotland, October 2011.

- H. Park, R. Ikeda, and J. Widom. RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows. To appear in Proceedings of the 37th International Conference on Very Large Data Bases, Seattle, Washington, August 2011. Demonstration description. (Also presented in Hadoop Summit 2011, Santa Clara, California, June 2011.)

- R. Ikeda, H. Park, and J. Widom. Provenance for Generalized Map and Reduce Workflows. Proceedings of the Fifth Biennial Conference on Innovative Data Systems Research (CIDR '11), Pacific Grove, California, January 2011.

- R. Ikeda and J. Widom. Panda: A System for Provenance and Data. IEEE Data Engineering Bulletin, Special Issue on Data Provenance, 33(3):42-49, September 2010. Note: An earlier, shorter version of this paper appeared in TaPP '10, cited below.

- R. Ikeda and J. Widom. Panda: A System for Provenance and Data. Proceedings of the 2nd

USENIX Workshop on the Theory and Practice of Provenance (TaPP '10), San Jose, California, February 2010. Note: A newer, extended version of this paper appeared in IEEE Data Engineering Bulletin, cited above.

Talks

- *Panda: A System for Provenance and Data* -- slides in PowerPoint (recommended for many animations) and pdf. Medium-length overview talk covering the state of the project as of Spring 2010.

People

- **Faculty**
  - Jennifer Widom
- **Students**
  - Robert Ikeda
  - Hyunjung Park
- **Alums**
  - Junsang Cho
  - Akash Das Sarma (IIT Kanpur)
  - Charlie Fang
  - Semih Salihoglu
  - Satoshi Torikai (Hitachi)

Last edited by J. Widom, July 2011