

A Novel Update Propagation Module for the Data Provenance Problem: A Contemplating Vision on Realizing Data Provenance from Models to Storage *

Abed Elhamid Lawabni¹ Changjin Hong¹ David H.C. Du² Ahmed H. Tewfik¹

Digital Technology Center, Intelligent Storage Consortium (DISC) and

¹ *Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455
{lawabni, hongcj92, tewfik}@ece.umn.edu*

² *Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN 55455
du@cs.umn.edu*

Abstract

To date, the systems approach to science, which emphasizes the connections among phenomena studied at different scales and by different disciplines, is causing dramatic changes in how scientific results are communicated. These changes drive a shift on how to propagate data with certain properties so that it can be used intelligently by others. In this work we elaborate on three major factors governing the propagation module of data provenance. The proposed propagation module provides an efficient solution for many critical problems in the management and provenance of scientific data. Unlike previous work, our work aims at realizing data provenance from models to storage. The natural representation of data as objects and its utility for capturing provenance has led us to consider a new storage architecture based on the object-based storage (OSD) technology. An outline of this framework is discussed.

1. Introduction

Provenance is a well-established concept in the art world where the lineage, pedigree or origins of a painting are critical to determining its authenticity and value. It is of equal importance in present and future data-rich environments ranging from computational biology, intelligence information gathering, to high energy physics.

Scientific research is based on exchanging data and conclusions. Data collected by a research group, or conclusions reached by the group, build on prior data and results produced by the group and the entire community. They in turn contribute to other derivative innovations, corrections and data. The integrity of scientific knowledge (its accuracy and reproducibility), the rate at which any scientific community can extend it, and the time elapsed between a new discovery and its widespread use for the greater good of society, all depend on the ability to track the propagation and complex interdependencies of the underlying representations and embodiments of knowledge.

There are numerous forms of provenance. In particular and of importance to this work, is the derivation path of

information. The derivation path records the process by which results are generated from input data. This could include a workflow that orchestrates a number of processes and their parameters, or services. Accurately tracking the lineage (origin and subsequent processing history) and the propagation of data in the derivation path is essential for effective management and decision making when an experiment needs to be re-run in light of new or modified information.

The following two motivating use cases illustrate the importance of the issues described above. In molecular biology, where data is repeatedly copied, corrected, and transformed as it passes through numerous genomic databases, understanding where data has come from, its original confidence level and how it arrived in the user's database is of crucial to the trust a scientist will put in that data. Furthermore, if one or more of the data inputs get slightly modified or changed, knowing exactly the contribution of each input on the output variability can have a tremendous saving in terms of computations.

Consider a second case involves analysis of remote sensing data from satellites. Remote sensing data may be processed and reprocessed in many different ways. Examples include need to correct for distortion caused by the atmosphere and to interpolate measured data values onto geolocations. Very large datasets consisting of several years of data are periodically reprocessed to contribute to other derivative data products. Assessing the uncertainty and the influences or relative importance of each input parameters on the output variability, can correct subtle errors and expedite lengthy and complex procedures. The chain or pipeline of processing steps that generate standard "levels" of NASA remote sensing data products provides one common example.

Aligned with this vision, in this work, we elaborate on three major factors governing the propagation module of data in the derivation path; namely:

- (1) Sensitivity analysis (SA): to ascertain how much a model (numerical or otherwise) depends on each or some of its input parameters
- (2) Confidence level and uncertainty: how much the data can be trusted, and

* This work was supported by StorageTek, Veritas, Engenio, and Sun Micro through the sponsorships of DISC.

