Leveraging the Open Provenance Model as a Multi-tier Model for Global Climate Research

Eric G. Stephan, Todd D. Halter, and Brian D. Ermold

Pacific Northwest National Laboratory
Richland, Wa
{Eric.Stephan,Todd.Halter,Brian.Ermold}@pnl.gov

Abstract. Global climate researchers rely upon many forms of sensor data and analytical methods to help profile subtle changes in climate conditions. The U.S. Department of Energy's Atmospheric Radiation Measurement (ARM) program provides researchers with a collection of curated Value Added Products (VAPs) resulting from continuous sensor data streams, data fusion, and modeling. We are leveraging the Open Provenance Model as a foundational construct that serves the needs of both the VAP producers and consumers. We are organizing the provenance in different tiers of granularity to model VAP lineage, causality at the component level within a VAP, and the causality for each time step as samples are being assembled within the VAP. This paper shares our implementation strategy and how the ARM operations staff and the climate research community can greatly benefit from this approach to more effectively assess and quantify VAP provenance.

Keywords: Provenance, Climate.

1 Introduction

In this paper we present how the Atmospheric Radiation Measurement (ARM) program is relying upon the Open Provenance Model [1] and its overlapping accounts feature to track provenance for data processing at different granularity levels.

The Pacific Northwest National Laboratory (PNNL) has been an integral part of the Department of Energy (DOE) ARM [2] program's infrastructure team since its inception in 1998. The ARM Data Management Facility manages data flow for over 300 sensors located around the world, ingests the data into an ARM standard format, performs quality control on the data through the ARM Data Quality Office, performs reprocessing on the data through the ARM Reprocessing Center, and transfers the resulting data sets to the ARM Archive. In addition, the facility is responsible for the development and deployment of Value Added Products (VAPs) [2] that provide derived data products through complicated processing pipelines. VAPs fuse information from sensors, models, algorithms, and other VAPs to derive information of interest that is either impractical or impossible to measure directly. The information of interest includes (but is not limited to) cloud microphysics, aerosol properties, atmospheric state, and radiometric properties. VAPs can also be used to improve the quality of existing sensor data, and when multiple sensors are producing the same

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 34–41, 2010.

type of data a "best estimate" VAP will identify the highest quality data. This experience has given us significant familiarity with a variety of climate data sets, as well as production-level experience handling streaming data, long-term data sets, and data reprocessing. ARM data is stored in a NetCDF file format that provide a structure that supports the storage of the data sets annotated with metadata. A significant need from the users is to directly disseminate provenance into the ARM NetCDF results, providing transparency to the user and greatly adding value to the analysis without requiring significant changes to the large body of existing analysis workflow. From an operational standpoint, it is foreseen that the number of sensors and ARM data products will increase significantly, dwarfing today's complexity of algorithm interdependency. We envision provenance as an overarching data-driven standard advancing many of the day-to-day tasks relating to data processing and reprocessing, error detection and troubleshooting in analytical methods. This is in contrast today where there is no standardization and all tasks are managed from scripts, legacy codes, and developer defined log files.

For the past twenty years most questions could be answered either through web reports or by relying on knowledgeable operations staff to determine the source of any problem by examining log and configuration files and performing database queries. However, with the deployment of new instruments, an order of magnitude increase in data throughput, and new advanced data products, a need for formalized methodologies was identified to automate analysis of the data products to efficiently continue to maintain strict quality assurance and quality control measures. In addition to this, data consumers began wanting a clearer understanding of the sources (instrument, data ingest processes, and higher order climate algorithms) relied upon for data, the confidence scores associated with the data, and other relevant information for each sample point.

In our early assessment of provenance needs using the Open Provenance Model (OPM) XML schemata, provenance exceeded sample datum by a storage space ratio of 1:23,000. In the past, we were accustomed to thinking of causality being considerably smaller than the resulting data, and that one causality graph consisting of hundreds of artifacts, processes, and relationships could represent the entire workflow history [3][4]. Because of diverse provenance-related questions being asked at different granularities, we also needed a flexible model that could be dissected in a variety of ways to support a number of analytical mash-ups. Based on all these factors a multi-scale conceptual model resulted [5].

2 Relevant Research

From a modeling perspective the provenance community has explored multi-tier approaches in the past. A visual analytic tool [6] developed for provenance exploration relied upon a multi-tier model approach for depicting multiple levels of provenance granularity. OPM uses accounts to depict overlapping provenance models, refinements, and hierarchies.

To describe the organization and structure of provenance for VAP-run instances the mathematical construct hedge is borrowed from the tree automaton community. A hedge is defined as a tree with an ordered sequence of unranked subtrees[7][8].

While we expect there to be mathematical or statistical properties of trees (and by extension hedges) that can be leveraged to derive additional provenance information or insights into either usual or anomalous states of the data analysis workflow, it is still too early to determine what can be leveraged until we can perform a deep dive into the provenance actually being generated by the VAP runs. This is needed before mathematical tools can be applied to augment the analysis.

3 Multi-tier Provenance Model

ARM provenance is being used as a means to more uniformly describe a complete history of VAP sample generation, VAP runs, and VAP interdependency. The rationale behind using a multi-tier model as opposed to a monolithic model is that each tier (or component) has a unique purpose, different characteristics, and distinct levels of granularity. The term multi-tier could be thought of as multi-part, with each tier being a separate component representation, but sharing overlapping parts. The model is broken into three distinct levels of granularity (Table 1) that are interconnected. To help visualize the model in its entirety we think of the three tiers as a landscaped park with directed paths running from hedge to hedge, and branches ordered along the hedge trunk.

Table 1. Granularity and purpose for each provenance tier

Granularity	Purpose	Tier	Account
High	Sample	Branch	
Middle	Run	Hedge	
Low	Interdependency	Path	

Branch Tier: As each VAP sample is processed the execution history will also be captured. This means that for a VAP that provides temperature measurements at a one second interval over a twelve month period, 500,000 samples will result along with a separate execution history (branch) for each data point occupying approximately 15GB of uncompressed storage (1 GB compressed). Because each sample is produced autonomously an ordered set of acyclic spanning trees are formed that can be analyzed to find missing samples, periodic anomalies in the workflow, or sporadic exceptions that may occur.

The Branch tier will rely on the following components in OPM: accounts, entities, and relationships. Each branch will be identified by its own account using an identifier that pertains to a corresponding sample time step. Because the time step interval is consistent, time will be inferred by sample order within the hedge tier. Each branch account will share with its parent hedge account a common process that initiates the sample analysis process as shown in Figure 1.

Processes will include references to sample analysis algorithms, workflow control logic, and data processing. Artifacts will include references to informational messages collected during processing along with warnings, errors, fatal messages, QC codes,

and sample origin. These entities rely on the following statements to depict relationships and on directionality to depict the execution history: Artifact Used By Process, Process Triggered By Process, Artifact Generated By Process. The definition of artifacts will be extended to include the Dublin Core Element Set [9].

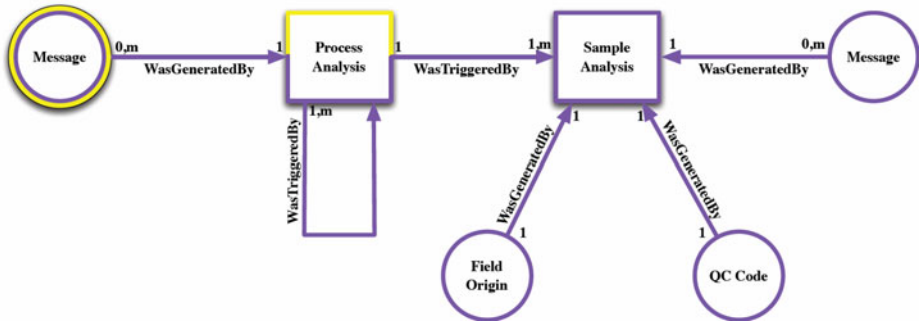

Fig. 1. Branch Tier

Hedge Tier: All sample points within a VAP rely upon the same overall workflow control logic and configuration parameters to process all samples. From a workflow perspective the VAP is a simple workflow that prepares data for analysis, performs analysis through a huge control loop that iterates over each sample, and then performs post-processing by storing the VAP in a NetCDF file. From a provenance perspective this forms an overarching graph of the VAP workflow history. Each VAP has only one hedge tier and one account that will be uniquely identified with a corresponding VAP identifier. As shown in Figure 2, the middle tier in the hedge account overlaps with each branch account (each iteration through the control loop), but leaves the detail of the sample analysis to the branch tier to separately provide these details. The hedge tier also interfaces with the Path Tier that will be described in the next section.

The Agent entity describes the person or control that initiated the VAP run. Artifact entities correspond to VAP parameter settings along with VAP or sensor data streams to build the product. The definition of artifacts will be extended to include the Dublin Core Element Set along with emerging standards from the Climate and Forecast Working Group.

Process entities include workflow controller logic, and data pre and post processing along with Dublin Core metadata describing the software identity and version number. These entities rely upon relationships and directionality to form the following statements: Process Controlled By Agent, Artifact Used by Process, Process Triggered by Process (branch), and Artifact Generated by Process.

Path Tier: Understanding lineage is extremely important because most of the VAPs are a composite of sensor streams disseminated from existing VAPs. Understanding this interdependence is vital to ARM operations staff that periodically must invalidate VAPs due to discoveries of error conditions. This creates a cascading effect, impacting VAPs reliant upon erroneous data. To track these conditions only one Account is used to track lineage. The Path account overlaps with each Hedge account. The VAP

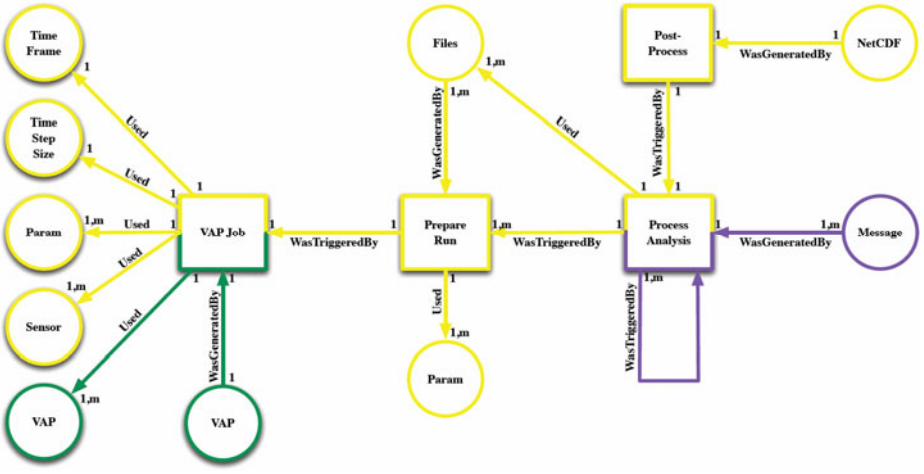

Fig. 2. Hedge Tier

Artifacts and the main VAP processes are shared between the Path and Hedge Tiers. The OPM entities included at this tier are: Artifacts, Processes, and Agents. Agents are considered a production batch process; otherwise the developer initiates a VAP run. The Process tracked is the overall VAP script used, and the Artifacts are the sensor streams and VAPs relied upon to create the VAP. These entities rely upon relationships and directionality to form the following statements: Process controlled by Agent, Artifact Generated by Process, and Artifact Used by Process.

4 Discussion

In this section we describe how provenance is automated, and how producers and consumers will use the model. Relying on the provenance model gives us the opportunity to automate and streamline many of the quality-related processes.

Automating Provenance: The data producers are responsible for the full life cycle of VAPs, algorithms used within the VAP, managing the interdependencies between the VAP runs, and maintaining the latest products. Our current plan is to apply a provenance listener that ties into the ARM error handler and message logging system. This is currently deployed in the ARM environment and has already been abstracted to track events from a developer's perspective.

For any given VAP run, provenance will be generated in the following top-down context: path, hedge, and branch. The script or workflow generating the VAP will be responsible for attaching any referential information and the associated event history along with any relevant event history.

Once the provenance is captured, it will be managed by maintaining synchronization between the different provenance tiers in the store. Provenance will be stored persistently as separate storage blocks. The path tier block provides references to the hedge block; the hedge tier will in turn provides references to each branch block. Because of the foreseen storage bloat issues we are remaining flexible on our storage

technical solution to determine the best overall approach and depending on the producer/consumer needs how long the raw provenance will be retained. Some alternatives under consideration are: distributed file stores with provenance formatted as XML or RDF N3 format, and relying upon the Open Source Array Database being developed by SciDB [10] which is expected to be released in 2010.

Analytical Methods – At least three automated analytical methods are now available that were either not previously available or easily attainable without provenance: automated quantitative analysis, interference, and discovery. We foresee many other types of analysis obtainable in the future. Quantitative analysis is extremely important at the sample level. Erroneous conditions typically are not self evident when looking at a single branch, rather, anomalies may occur due to a faulty instrument reading at a particular time of the day. Only by comparing multiple branches side by side are the conditions apparent. Squarified treemap [11] display tree structures as flattened nested boxes and arrange similar looking boxes together. In our tests we relied on VAPs that collected cloud cover data. Through the squarified treemap we were able to visualize conditions corresponding to clear skies where no samples were collected and processed. By rendering branches into summary level views over selected time periods (day, hour minute) of distinct entity (artifacts or processes) counts in a relational database schemata we were able to detect times when provenance branches contained atypical number of nodes this we were not only able to detect and rendered branches. OPM provides transitive closure is relied upon to determine is workflows successfully completed, and we explored uses of discovery either by providing a web-based search engine based on selected indexed provenance and by using a tree structure visualization tool Prefuse [12] to browse selected portions of the path, hedge, and branch tiers.

In addition to the flexibility offered to analytical methods is also scalability by means of being able to discretely dissect provenance horizontally along one tier, or vertically between one or more tiers by selecting accounts based on VAP interdependency, VAPs themselves, or specific branches representing time slices of different VAPs. The branches represent a wealth of knowledge needed for comparative analysis. It may become quite common to detect various anomalies from different branch time slices to differentiate between sensor mechanical problems and natural phenomena. In this example to thoroughly conduct this analysis the time branches for the last ten years over a given month might be necessary to confirm the origin of a sensor's behavior. Because each branch is a non overlapping acyclic graph, groups of branches can be analyzed one group at a time or potentially distributed and analyzed over multiple compute nodes to increase efficiency.

Uses – It was important to tie the analytical methods and provenance to practical examples to demonstrate how provenance might be used by scientists and operations. We split out a special category, developers, which are part of operations and who have a special need for provenance related to debugging activities [5].

From the consumer perspective the scientist's primary motivation in examining historical evidence has more to do with establishing confidence in their VAP of interest. In most cases we anticipate their interest at the hedge and branch tier of a given VAP run. At the hedge level the input deck used, the specific algorithms (including version) relied upon, along with the time span time step intervals used in the VAP run are fundamental to helping the researcher establish the context of how the VAP was created.

From our preliminary research we found that while the data producers are generally interested in the resulting graph representation of the branch, the scientists prefer a flattened view derived from each branch. This view currently consists of at least two fields: field origin and quality control codes. Each will correspond to a sample datum within the NetCDF file. This will provide scientists with a detailed knowledge of how the product was assembled at an incremental level and will allow them to understand possible reasons for trends or anomalies themselves.

While we do not have implementation details at this time we are determining how best to disseminate this knowledge to the user community. Simply showing the raw OPM data does not seem effective for end users. Our current plan is to disseminate provenance as part of the VAP as an encoded bit pack describing instruments used, parameters used, and quality/confidence level based on the provenance results. This encoded information may result from the branch tier for sample level data, or may be associated from the hedge tier associated with an overall VAP run. The key is that provenance for a given VAP instance will reside within the archive and the provenance encoding or will be distributed in the NetCDF file for each sample.

5 Conclusion

The Open Provenance Model has provided ARM a strong foundation for supporting overlapping provenance models that represent different processing refinements during the creation of the VAP. While we have identified some initial ways to exploit the provenance information there are many more to explore. We also believe that with the abundance of provenance, workflow query, and analysis, that exposing the hedges will make it far easier for future applications to transform, query, and analyze VAP results.

Acknowledgments. This research is supported by the Laboratory Directed Research and Development Program at the Pacific Northwest National Laboratory operated by Battelle for the U.S. Department of Energy under contract DE-AC05-76Rl0 1839.

We also acknowledge the collaborative efforts of U.S. Department of Energy as part of the Atmospheric Radiation Measurement Program.

References

1. Moreau, L., Clifford, B., Freire, J., Gil, Y., Groth, P., Futrelle, J., Miles, S., Myers, J., Simmhan, Y.L., Stephan, E.G.: The Open Provenance Model Core Specification, v1.1 (2010)
2. Atmospheric Radiation Measurement Climate Research Facility. ARM Annual Report. Technical report available from U.S. Department of Energy as DOE/SC-ARM-0706
3. Gibson, T.D., Stephan, E.G.: Application of Provenance for Automated and Research Driven Workflows. In: Tara at Second International Provenance and Anotation Workshop, June 17, Salt Lake City, UT (2008)
4. Gibson, T.D., Schuchardt, K.L., Stephan, E.G.: Application of Named Graphs Towards Custom Provenance Views. In: 1st Workshop on the Theory and Practice of Provenance (TaPP 2009), USENIX, Berkeley, CA (2009)

5. Stephan, E.G., Halter, T.D., Gibson, T.D., Beagley, N., Schuchardt, K.L.: A Multi-Tier Provenance Model for Global Climate Research. In: International Conference on Network-Based Information Systems (NBIS 2009), Indianapolis, Indiana, August 19-21, pp. 481–486. IEEE, Piscataway (2009), doi:10.1109/NBiS.2009.16
6. David, G., Michelle, Z.: Characterizing Users' Visual Analytic Activity for Insight Provenance (2008)
7. Comon, H., Dauchet, M., et al.: Tree automata techniques and applications (1997), http://www.grappa.univ-lille3.fr/ [11] 10
8. Courcelle, B.: On recognizable sets and tree automata. In: Nivat, M., Ait-Kaci, H. (eds.) Resolution of Equations in Alegebraic Structures,
9. Dublin Core metadata semantics: An analysis of the perspectives of information professionals Park and Childress. Journal of Information Science (2009); 0165551509337871v16
10. Stonebraker, M., Becla, J., Dewitt, D., Lim, K., Maier, D., Ratzesberger, O., Zdonik, S.: Requirements for Science Data Bases and SciDB. In: Conference on Innovative Data Systems Research, CIDR (2009)
11. Bruls, D.M., Huizing, C., van Wijk, J.J.: Squarified Treemaps. In: de Leeuw, W., van Liere, R. (eds.) Data Visualization 2000, Proceedings of the joint Eurographics and IEEE TCVG Symposium on Visualization, pp. 33–42 (2000)
12. Heer, J., Card, S.K., Landay, J.A.: Prefuse: A Toolkit for Interactive Information Visualization. In: CHI 2005, Portland, OR, April 2-7 (2005)