NSF ACI-1443069

Collaborative Scientific Workflow Composition as a Service

-An Infrastructure Supporting Collaborative Data Analytics Workflow Design and Management

PI: Jia Zhang, Carnegie Mellon University -Silicon Valley, USA
CO-PI: Shiyong Lu, Wayne State University, USA
jia.zhang@sv.cmu.edu; shiyong@wayne.edu

The need for collaborative data analysis increases significantly when confronted with the challenges of big data. Although workflow tools offer a formal way to define, automate, and repeat multi-step computational procedures, designing complex data process workflow requires collaboration from multiple people with complementary expertise. Existing tools are not suitable to support collaborative design of comprehensive workflows.

To address these challenges, we have designed and developed a technique supporting *collaborative data-oriented workflow composition*, as a key component toward supporting big data collaboration through the Internet. As the first step, we address one major research challenge of collaborative provenance management. We focus on issues unique to collaborative composition provenance with regard to modeling, gathering, versioning, storing, and querying of workflows. Our contributions are three-fold. First, we have developed a collaborative provenance data model equipped with a graph-level provenance querying formalism. Second, we have developed hypergraph theory-based algorithms for provenance management and mining. Third, we have developed a novel software tool supporting (a)synchronous collaborative scientific workflow design, composition, reproduction, and visualization. Instead of reinventing the wheel, we have extended an existing workflow tool VisTrails as a proof of concept.

We argue that the details of such a collaboration process should be recorded as provenance. In contrast to normal workflow provenance capturing derivation history of data products at run time, we define *collaborative workflow composition* provenance to record knowledge allowing participants: 1) to validate a workflow by tracking how a workflow has become as it is from multiple collaborators; 2) to acknowledge credits by recording who has done what at what time; 3) to capture and retrieve collaboration knowledge (annotations and discussions); and 4) to form the basis for merging workflow changes from distributed multiple users.

In order to catch the collaborative workflow composition activities, we have extended the PROV-DM (PROV Data Model, https://www.w3.org/TR/prov-dm) and developed a Collaborative Provenance Model (CPM). CPM is equipped with a **graph-level** querying formalism and efficient query optimization techniques for managing the lifecycle of collaborative provenance. In contrast to existing approaches, our higher graph-level query formalization will not be tightly coupled to the underlying provenance storage strategies, while featuring the native support for query processing at the provenance graph level. After identifying a collection of graph patterns, we define a set of operators to manipulate and query a CPM graph. They enable composition of various graph patterns, allowing querying formulation for collaborative provenance.

Collected provenance data can be leveraged to support optimized workflow co-design. In order to realize automatic workflow derivation and composition from provenance mining, our strategy is to model workflow graphs, versioning, and derivation history as hypergraphs. In contrast to normal graph where each edge links between two nodes at two ends, hypergraph as a generalization of normal graph can have edges that connect together more than two nodes. With such hypergraph structure, we have applied and developed hypergraph-theory based workflow mining algorithms to support collaborative workflow composition.

Without reinventing the wheel, we have built our software as a plug-in to VisTrails, a widely used scientific workflow management system. Leveraging their software infrastructure in the last decade, we extended it into a collaborative version. Multiple people can use our collaborative-VisTrails to design a workflow together. Any change made by one scientist will be immediately reflected on all other collaborators' screens. A backend version tree is established to ensure concurrency control.

The resulting tools will explore the potential for using scientific workflows to accelerate scientific discoveries that require a collaborative effort on big data analytics. The design of the tools is targeted toward use cases in the civil engineering discipline, but has the potential to broadly impact other areas of science and engineering. Partnership with VisTrails enables usage and evaluation of the techniques in the VisTrails end user community.

In our future work, we plan to further study how CPM can answer a variety of types of queries. In addition, we will further explore hypergraph-based search algorithms. To improve the usability of graph database, which is challenging for regular users, we will explore a more high-level, user-friendly language for formulating provenance queries. Furthermore, we plan to move VisTrails online to develop an online, collaborative workflow development environment.