Scientific Workbench and Workflow management for GIS workflow

Michiel Johan Baird
Department of Computer Science
University of Cape Town

May 14, 2012

Abstract

Automated workflow systems have been successfully implemented across various disciplines, including scientific- and business workflows. This is an overview of what has has been done in the field. Special attention will be given to the building of a Scientific Workbench. It focusses on previous efforts, highlighting some the methods used as well as the lessons learnt during the implementations.

It also looks at how these principles could be applied specifically to GIS workflow by giving an overview of the structure of the field. GIS workflow is then shown to be an appropriate match for an automated workflow system.

This solution is shown to be highly applicable to GIS workflow, provided the necessary middle-ware can be built to facilitate integration.

1 Introduction

Workflow management systems define a complex process in into well formed tasks and coordinates the process completion (Qiu et al., 2003). Automated workflow management has been in wide use across various disciplines since the concept was formalised in 1996(Carstensen & Srensen, 1996). Successful systems have been implemented across various fields including banking, pharmaceuticals and various others (Brahe & Schmidt, 2007; Johnson et al., 2009).

It has been shown to be very successful the sciences as the same scientific process can easily be repeated on a different set of data(De Roure & Goble, 2009). This not only aids in reproducibility but also saves time. This is done by efficiently abstracting the operations in the flow, allowing it to be automatically handled.

Geographic Information Systems(GIS) is the field that concerns itself with the organisation, representation and processing of geographic data, for the purpose of querying it and making decissions off of the data (Di Martino et al., 2007). The workflow in GIS is very distributed and the set of data that is operated on is large and diverse. Workflow management within GIS has been considered and solutions have been proposed, but not implemented or evaluated(Migliorini et al., 2011).

2 Overview

A workflow management system consists of definitions on how a set of tasks should be executed (Carstensen & Srensen, 1996; van der Aalst & Basten, 2002). The overall procedure is defined by the following components: (i) actors, (ii) roles, (iii) responsibilities and obligations, (iv) tasks, (v) activities,(vi) conceptual structures and (vii) resources.

A real life problem or task can then be broken up into these components in such a way that the tasks represent a flow network. These tasks then connect to the actors and resources via the other components(Taylor et al., 2006, p. 4). This allows task to be executed efficiently in a distributed manor.

The initial implementations of a workflow system, however, almost immediately failed. The system was too rigid and was unable to accommodate the high levels of change that was required by the users (Suchman, 1983).

These changes come from a number of sources, including: ill-specification of initial problems, change in actors or resources, exceptions that occurred and new requirements. Adaptive workflow systems were proposed to solve this problem by providing a mechanism for allowing change in the system(van der Aalst & Basten, 2002). This allows processes to be extended, replaced or re-ordered. It also adds the ability to change already running tasks by providing restart, transfer

and proceed options.

Scientific workflow management has also been very successful with how experiments are defined, and more importantly, reused. Another benefit that was quickly discovered was that it also allowed researchers to trade workflows, making the replication of results much easier than they were previously(De Roure & Goble, 2009). Keys to this success were: that the workflow systems were made to fit the researchers, quick responses to adding required features when needed, listening to user input and making sharing of workflows as easy as possible.

Such a system has also been applied in fields that operate on large data sets, as would be the case if applied to GIS problems(Aragon & Runge, 2009). Workflow systems were found to work well in the management of getting this data processed. Applying the concept to Observational astrophysics, it revealed that it could be used to identify bottlenecks that could be optimised. Further it was used to automatically ensure local access of large files that needed to be processed.

3 Geographic Data

GIS concerns itself with the collection, organisation and query of geographic data (Di Martino et al., 2007). This data includes but is not limited to landscapes, coordinate data, building models, statistics, pictures, textures and routes. This is a very broad set of data, varying from very large to very small. That variation however, means that there exists no uniform method to efficiently deal with the data.

The processing of this data can vary from human to software processing (Di Martino et al., 2007). Various Web applications have been written to facilitate the tasks that need to be accomplished. This software is known as WebGIS and is becoming more popular with scientists; it also means that even within the field there is a strong shift toward Web based services.

A key realisation with the usage of this data is that the same data is used across various applications, to create various amounts of abstractions(El Adnani et al., 2001). The core data is seldom changed. Instead a new abstraction layer is added on top of it. The data can be thought of as a graph, where the nodes represent either a data or abstraction element, and the edges represents the functions/tasks required to create the particular abstraction as a set of topological relationships. This can be effectively used to provide high levels of GIS interoperability.

4 Implementations

There are various products available that can compose scientific workflows. *The Trident workbench* (Simmhan et al., 2009) is an open source workflow management system developed by Microsoft Research that also adds middleware services and a graphical composition interface. Trident builds workflows of control and data flows, off of built-in, user defined activities and nested subflows.

The flows are represented using XOML, an XML Specification, while the activities are stored as a set of sub routines(Simmhan & Barga, 2011). Trident can be used on a local system, remote systems and even clusters. Queries on the system can be performed using LINQ.

Kepler is another scientific workflow management system that provides workflow design and execution. Actors are designed to perform independent tasks that can either be atomic or composite (Wang et al., 2009). Composite actors(subflows) consist of multiple atomic actors bundled together. Actors can consume data and produce output, called tokens. Actors communicate tokens with each other via links. The order of execution and the links are defined by an independent entity called the director. As a consequence, the workflow can either be executed in a sequential or parallel manner. Kepler effectively separates the workflow from its execution, allowing for easy batch execution. Actors can easily be exported and shared. Kepler is very popular due to its adaptability and easy integration.

Taverna is a scientific workbench that supports application-level workflow and does not focus on scheduling as much others(De Roure & Goble, 2009). Taverna has a strong focus on workflow sharing. Taverna is quite popular, since there exists a social network, designed to facilitate workflow sharing among scientists(*myExperiment*). Services are linked to the model to execute the various tasks. Taverna can be used in such a way that it can utilize all the services a client has to facilitate the flow by easily adding services. The Taverna language is a simple data-flow language called the Simple Conceptual Unified Language(SCUFL), that can be encoded to XML.

In order for these workbenches to be successful, there needs to exists a high level of interoperability between the workflow management and the services that are

required (Shegalov et al., 2001). However, due to the fact that there is a relatively high chance of failure when building this interoperability into the services as a core component. It is an extremely high risk and therefore is not typically done. A Cheaper way of doing this is providing middleware that can wrap around the service to provide the required interfaces.

This need for interoperability has led to the popularisation SOA(Service Orientated Architecture) (Sanders et al., 2008). It should be noted that SOA is *not* an implementation, but rather an *Architectural Model* SOA refers to a collection of loosely coupled services, that individually carry out a particular process. Each service should have a well defined interface with self-contained functionality. It should allow other applications or services to use this functionality without knowing the underlying technical details. These services should be hidden from the end-user and its usage should preferably be platform-independent.

Although the concept has been around since the 1970s, it has only recently gained favour due to Web services. Web services are software that run on the internet through XML standards-based interfaces(Tai et al., 2004). Each service provides a fuctional description using the *Web Services Description Language*(WSDL). This description provides the supported operations, as well as the definition of the input and output messages.

By using these concepts, a workflow system can be built that automatically uses these Web Services to facilitate both the data and control flow using well defined interfaces in standards such as XML/JSON. (Shegalov et al., 2001). With the advancement of WebGIS, a lot of Web Services that facilitates GIS processing already exists.

5 Case Studies

The next section will look at two instances where workflow management systems were implemented and used. These case studies will look at both a business and a scientific application.

Danske Bank

The workflow management system at *Danske bank* was incrementally implemented as their system moved from a manual system(Brahe & Schmidt, 2007).

This system was developed as an in-house solution when the manual system could not cope anymore. Several lessons were learnt that are applicable to other work flow systems. When work was divided purely from an efficiency point of view, the workers became complacent as they felt that they did not understand the overall mechanism and felt that they were not involved. They discovered that the system did not handle change very well. This change was expensive and inevitable. Their system had to be adapted to handle this change. The success of the system is mainly attributed to the interoperability and close relationship between the users and the developers

OrthoSearch

OrophoSearch is a workflow, built on *Kepler*, that is designed to work on work on data in the field of Bio Informatics. (da Cruz et al., 2008)

A workflow system was implemented in *Kepler* as it addressed the requirements they had, including: (i) Workflow definition and Design; (ii) workflow execution control; (iii) fault tolerance; (iv) intermediate data management; and (v) data provenance support.

Although the system was not without its hiccups and changes, the integration with Kepler provided the workflow increased overall productivity.

6 Conclusion

The field of GIS concerns itself with a vast amount of geographic data. This data comes in various sizes and as such different methods of handling and transferring would need to be used to facilitate dataflows within the system.

The work however is done in a very distributed manner, which allows for a very effective mapping onto a grid-based computing solution, provided middleware can be developed to support the systems that are used(Montella et al., 2007). This would allow for an effective Content Delivery Network that provides data on demand where it is needed on the grid.

GIS workflow, due to its distributed nature, would map well onto a automated workflow system

(Withana et al., 2010). The nature of the science is supported well. It would allow for effective automatisation of some of the functions are available.

References

Aragon C R & Runge K J 2009 *in Proceedings of the 2009 ACM symposium on Applied Computing*

SAC '09 ACM New York, NY, USA pp. 949–955.
URL: *http://doi.acm.org/10.1145/1529282.1529491*

Brahe S & Schmidt K 2007 *in Proceedings of the 2007 international ACM conference on Supporting group work* GROUP '07 ACM New York, NY, USA pp. 249–258.
URL: *http://doi.acm.org/10.1145/1316624.1316661*

Carstensen P H & Srensen C 1996 *Computer Supported Cooperative Work (CSCW)* **5**, 387–413. 10.1007/BF00136712.
URL: *http://dx.doi.org/10.1007/BF00136712*

da Cruz S M S, Batista V, Dávila A M R, Silva E, Tosta F, Vilela C, Campos M L M, Cuadrat R, Tschoeke D & Mattoso M 2008 *in Proceedings of the 2008 ACM symposium on Applied computing* SAC '08 ACM New York, NY, USA pp. 1282–1286.
URL: *http://doi.acm.org/10.1145/1363686.1363983*

De Roure D & Goble C 2009 *Software, IEEE* **26**(1), 88 –95.

Di Martino S, Ferrucci F, Paolino L, Sebillo M, Tortora G, Vitiello G & Avagliano G 2007 *in Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* GIS '07 ACM New York, NY, USA pp. 57:1–57:4.
URL: *http://doi.acm.org/10.1145/1341012.1341081*

El Adnani M, Yétongnon K & Benslimane D 2001 *in Proceedings of the 9th ACM international symposium on Advances in geographic information systems* GIS '01 ACM New York, NY, USA pp. 70–75.
URL: *http://doi.acm.org/10.1145/512161.512177*

Johnson D, Meacham K & Kornmayer H 2009 *in E-Science Workshops, 2009 5th IEEE International Conference on* pp. 86 –91.

Migliorini S, Gambini M, Belussi A, Negri M & Pelagatti G 2011 *in Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications* COM.Geo '11 ACM New York, NY, USA pp. 36:1–36:6.
URL: *http://doi.acm.org/10.1145/1999320.1999356*

Montella R, Giunta G & Riccio A 2007 *in Proceedings of the second workshop on Use of P2P, GRID and agents for the development of content networks* UPGRADE '07 ACM New York, NY, USA pp. 81–86.
URL: *http://doi.acm.org/10.1145/1272980.1272995*

Qiu X, Du Y & Zheng F 2003 *in Systems, Man and Cybernetics, 2003. IEEE International Conference on* Vol. 5 pp. 5016 – 5021 vol.5.

Sanders D T, Hamilton, Jr. J A & MacDonald R A 2008 *in Proceedings of the 2008 Spring simulation multiconference* SpringSim '08 Society for Computer Simulation International San Diego, CA, USA pp. 325–334.
URL: *http://dl.acm.org/citation.cfm?id=1400549.1400595*

Shegalov G, Gillmann M & Weikum G 2001 *The VLDB Journal* **10**(1), 91–103.
URL: *http://dl.acm.org/citation.cfm?id=767132.767139*

Simmhan Y & Barga R 2011 *Future Generation Computer Systems* **27**(6), 790 – 796.
URL: *http://www.sciencedirect.com/science/article/pii/S0167739X10001986*

Simmhan Y, Barga R, Ingen C v, Lazowska E & Szalay A 2009 *in Proceedings of the 2009 Third International Conference on Advanced Engineering Computing and Applications in Sciences* ADVCOMP '09 IEEE Computer Society Washington, DC, USA pp. 41–50.
URL: *http://dx.doi.org/10.1109/ADVCOMP.2009.14*

Suchman L A 1983 *ACM Trans. Inf. Syst.* **1**(4), 320–328.
URL: *http://doi.acm.org/10.1145/357442.357445*

Tai S, Khalaf R & Mikalsen T 2004 *in Proceedings of the 5th ACM/IFIP/USENIX international conference on Middleware* Middleware '04 Springer-Verlag New York, Inc. New York, NY, USA pp. 294–310.
URL: *http://dl.acm.org/citation.cfm?id=1045658.1045680*

Taylor I J, Deelman E, Gannon D B & Shields M 2006 *Workflows for e-Science: Scientific Workflows for Grids* Springer-Verlag New York, Inc. Secaucus, NJ, USA.

van der Aalst W & Basten T 2002 *Theoretical Computer Science* **270**(12), 125 – 203.
URL: *http://www.sciencedirect.com/science/article/pii/S0304397500003212*

Wang J, Crawl D & Altintas I 2009 *in Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science* WORKS '09 ACM New York, NY, USA pp. 12:1–12:8.
URL: *http://doi.acm.org/10.1145/1645164.1645176*

Withana E C, Plale B, Barga R & Araujo N 2010 *in Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing* HPDC '10 ACM New York, NY, USA pp. 756–765.
URL: *http://doi.acm.org/10.1145/1851476.1851586*