2 Developing the cyber-infrastructure

A key objective of the pilot Environmental Virtual Observatory (EVOp) project was to provide a cyber-infrastructure capable of demonstrating the potentials of a virtual observatory that utilises the powers of the Internet to support uniform and open access to a series of scientific resources, and to therefore support and encourage online experimentation. This aspect of the project aimed to show how the application of new Internet-based technologies, specifically cloud computing, and web portals can facilitate this vision and in particular achieve the integration of a wide variety of information sources (including disparate data sets, sensor data and models applicable at different temporal and spatial scales), together with associated information tools and services, to provide answers to big environmental science questions. From a technological perspective, the challenge was to define the architecture and associated architectural principles underpinning the EVOp, supporting multi-scale experimentation through an open extensible infrastructure, and harnessing existing resources (data, models, etc.). Core to this task was the definition of an overall architectural approach as a refinement of Web 2.0 standards, incorporating mechanisms to deal with meta-data, and the population of the architecture with exemplar services to support the project's other work packages.

2.1 What is cloud computing?

Cloud computing is core to the EVOp project, providing the underlying technology to implement the required cyber-infrastructure. Cloud computing has emerged as one of they key areas of digital innovation in recent years and the associated technology is having major impact in a variety of areas such as eCommerce, eGovernment and smart cities. The goal of EVOp was to investigate the potential impact of cloud computing on the Environmental Sciences and indeed on science more generally.

In general terms, cloud computing is a shift from resources being on individual computers towards having these services available in the greater Internet. The classic definition from the National Institute of Science and Technology (NIST) defines cloud computing as follows (??):

> "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

Coulouris et al (Coulouris et al., 2011) provide an alternative definition:

> "A cloud is defined as a set of Internet-based application, storage and computing services sufficient to support most users' needs, thus enabling them to largely or totally dispense with local data storage and application software".

They then go on to refine this vision saying:

> "[Cloud computing] also promotes the view of everything as a service, from physical or virtual infrastructure through to software, often paid for on a per-usage basis rather than purchased".

Cloud computing is better understood as not a single approach but actually a family of approaches and part of the EVOp story has been to work through the different options and the implications for the Environmental Sciences community. In particular, a key defining characteristic is who provides and manages a given cloud infrastructure:

- The first option is to have a private cloud implemented within an organisation for that organisation. This option requires significant up-front investment and the responsibility for managing the cloud rests with that organisation. The organisation though retains control and ownership of the resultant infrastructure and associated cloud services.
- The second option is to opt for a public cloud where the cloud is implemented and managed by a third party provider such as Amazon, Microsoft, or Google and offered through the Internet as services to the greater public (including organisations or private individuals). The advantages here are that the upfront investment and resultant management is delegated to a third party (for a given cost), but with a loss of control over the infrastructure (for example, for data it may not be possible to specify or know where the data is stored).

Hybrid solutions are also common where organisations will adopt a mix of public and private cloud provision. More generally, there is a tendency towards multi-cloud environments where organisations use multiple highlights the fact that, in cloud computing, everything is a service and this is the key unifying principle underpinning cloud computing. This encompasses though a variety of approaches, often distinguished as:

- Infrastructure as a Service (IaaS) - IaaS is the lowest level of abstraction and refers to access to basic underlying computing resources as a service. The most common example here is to request and be given access to a computer as a service, accessed across the Internet. More specifically, when coupled with virtualisation, you will be given a virtual machine which you can then use as if it is a real, physical machine for your purposes (e.g. to act as a web server or as a hosting environment for models).
- Software as a Service (SaaS) - at the opposite end of the spectrum, SaaS refers to being given access to application-level software. This is a very open ended category and includes the availability of a suite of applications that offer classic functionality such as word processing, e-mail, calendars, etc (e.g. Google Apps). This category also includes domain specific software, e.g. an environmental model that you would like to make available as a cloud service.
- Platform as a Service (PaaS) - PaaS sits in the middle and refers to services that reside above operating systems and which are useful in the construction of application software. Key examples include software frameworks to support web servers and databases.

A sub-goal of the EVOp project was to understand the full range of approaches under the 'cloud computing' umbrella and the implications for the Environmental Sciences community.

2.1 Why cloud computing for Environmental Services?

The general benefits of cloud computing are well document, for example see (Zhang et al., 2010). In this section, we focus on the specific benefits in the context of supporting an Environmental Virtual Observatory.

The most profound impact is in terms of supporting a new kind of science:

- An *open* science - whereby open access is provided to a range of environmental assets including data sets, models and supporting assets such as visualisations;
- A *shared and collaborative* science - whereby assets are accessible and shareable from anywhere via the cloud by different stakeholder groups, encouraging collaboration and opening up new avenues such as citizen science;
- An *integrative* science - whereby problems can be studied by bringing together data and models from different disciplinary perspectives, over different geographical regions and at different scales.

In addition to this, we gain a number of key benefits emanating from the cloud computing approach:

- *Everything as a service* - All models and data assets follow a common service model. This offers a level of transparency, by which details of where and how the data are held are hidden from users. This allows for data to be used in models and simulations without necessarily giving it to the users, avoiding some of the delicate issues of ownership.
- *On-demand cloud elasticity* - In cloud computing, it is possible to request resources as and when they are needed and then return them to the cloud provider when no longer needed. Through this approach, resources can grow and shrink to meet current demand (a property known as elasticity). For example, if a user is running a complex climate change model that requires extra virtual machines, these can be requested and returned when the model completes execution.
- *Delegated management* - Managing a cyber-infrastructure is complex, especially when it comes to ensuring key properties such as secure and reliable access to resources. For example, it is important to ensure that data is continually available in spite of the inevitable failure of underlying computing infrastructure, typically achieved through replication of data and the use of protocols to ensure consistency of replicas. By adopting a cloud approach, such management is delegated to the cloud provider meaning users do not need to worry about these key issues.

2.2 The EVOp approach

2.2.1 Overall approach

The first key decision was to adopt a multi-cloud approach spanning private and public cloud provision. The private cloud is hosted in Lancaster University and is operated by us using OpenStack, an open source virtual infrastructure management solution. The public cloud resources are provided by Amazon Web Services (AWS). The pairing of OpenStack and AWS is a common one in the cloud computing world: AWS is arguably the most mature and feature rich public IaaS provider, and OpenStack is backed by many (including large organisations like NASA, IBM, and many others) as the de facto open source alternative to the core AWS products, i.e. EC2 (utility computing) and S3 (storage service). This makes it possible, at least in theory, to use the same virtual machine images to start instances in either cloud. In order to promote portability and to avoid being tied in to one provider (vendor lock-in), we adopted cross-cloud library jclouds. This open source software provides abstractions across many of the widely used cloud solutions.

Where possible, we adopt standards to enable interoperability within the architecture. Both AWS and OpenStack adopt web service standards and hence, in EVOp, all services (data, models and other supportive services) offer standard web service interfaces. Again where possible, we adopt a RESTful approach to service APIs, an approach that promotes loose coupling and consequently major improvements in scalability and manageability (Fielding, 2000). The environmental models are implemented using the OGC (Open Geospatial Consortium) WPS (Web Processing Service) standard that specifies how geospatial inputs and outputs should be.

2.2.2 The system architecture

The overall architecture for the system is shown in Figure 2.1.

The Model Library is populated by domain specialists (e.g. hydrologists) in liaison with data providers. The process starts with online calibration and testing of a model against a certain dataset (e.g. TOP-MODEL on the rainfall data of the Eden catchment in the North West of England). The outcome of this process is a virtual machine image optimised to run a fine-tuned set of models that are exposed as web services and equipped with all required data. This streamlined execution bundle is then stored in the library to be instantiated upon demand.

Once a user navigates to one of the models, a connection is created with the Resource Broker module of the Infrastructure Manager. This broker responds with an address of a cloud instance that is suitable for the type of computation required, along with some session information. This communication is done using HTML5 WebSockets, which reduces network overhead and browser memory usage.

The Load Balancer monitors the status of running instances with two objectives: minimise costs and maintain instance responsiveness:

- For the former, user requests are served by default using private instances. Upon saturation of private cloud resources, the load Balancer initiates cloudbursting mode where public cloud instances are used beside private ones. This is reversed upon detecting underuse, migrating users back to use private instances.
- For the latter objective, performance metrics are collected and any notable degradation triggers the Load Balancer to start a new instance, redirecting users that were being served by the seemingly malfunctioning instance to the newly created one.

A number of models were migrated to the EVOp infrastructure. For each model, a bespoke visualisation was developed to suit the particular factors in question. In general terms, models generate one of two types of output: geospatial and time series. Geospatial data is visualised using interactive layers superimposed over maps. Google Maps is used due to its wealth in data, features, and the familiarity of the general public with it. The interactive nature of the geospatial layers allows expansion of the visualisation to include time series graphs over specific map locations.

An emphasis in all models is placed on adjustability and flexibility in order to provide an interactive and configurable user experience. This is achieved via dynamic HTML and HTML5 web elements, AJAX asynchronous communications, and browser scripting using advanced open source JavaScript libraries such as jQuery, Flot, qTip, and Google Maps.

2.2.3 The use of agile methodology

The development process in EVOp relied heavily on an agile methodology based on a behaviour-driven design. Requirements were drawn from specific storyboards that were outlined by the domain

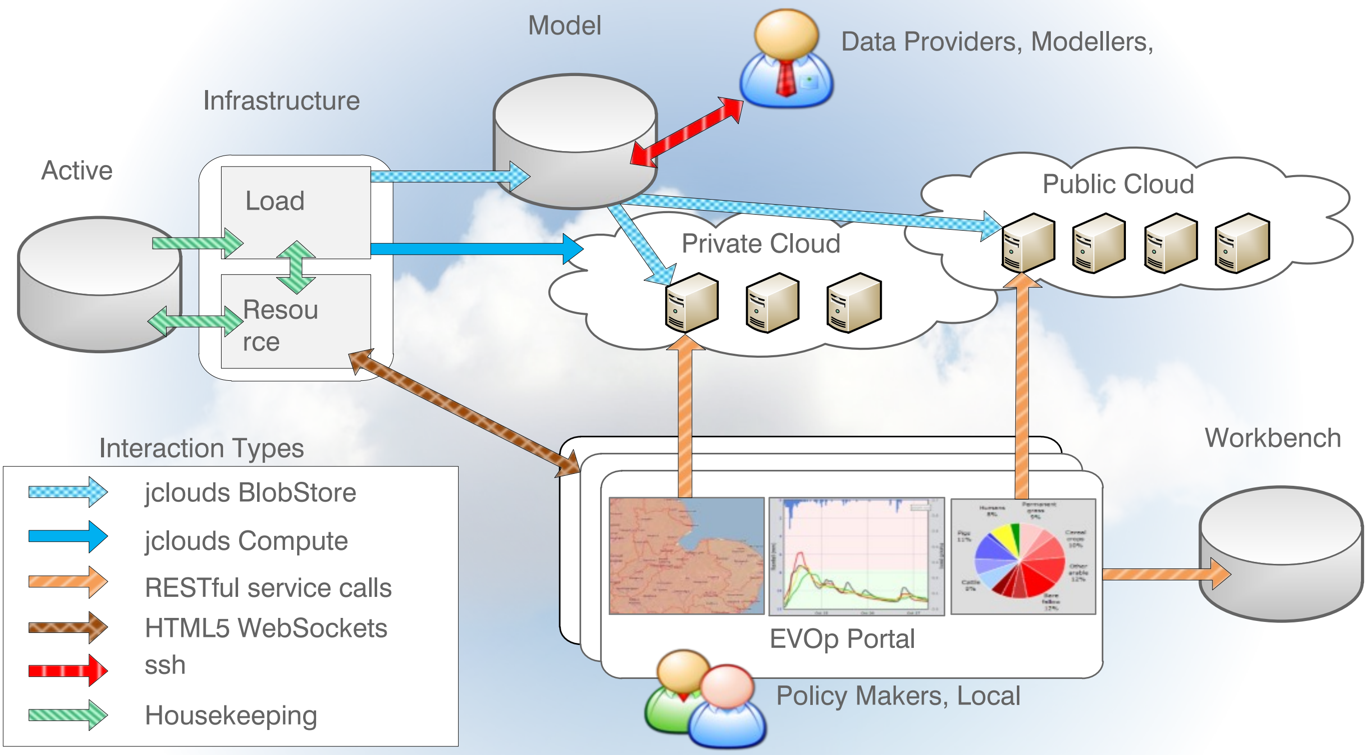


Figure 2.1 The overall architecture of the EVOp cyber-infrastructure.

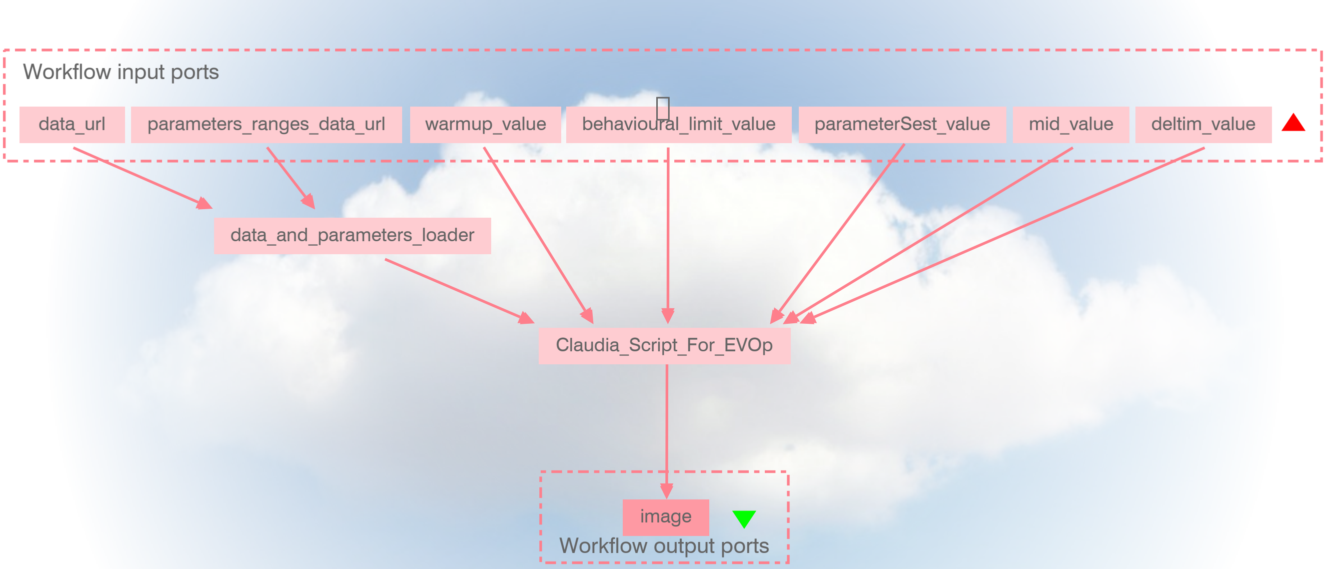
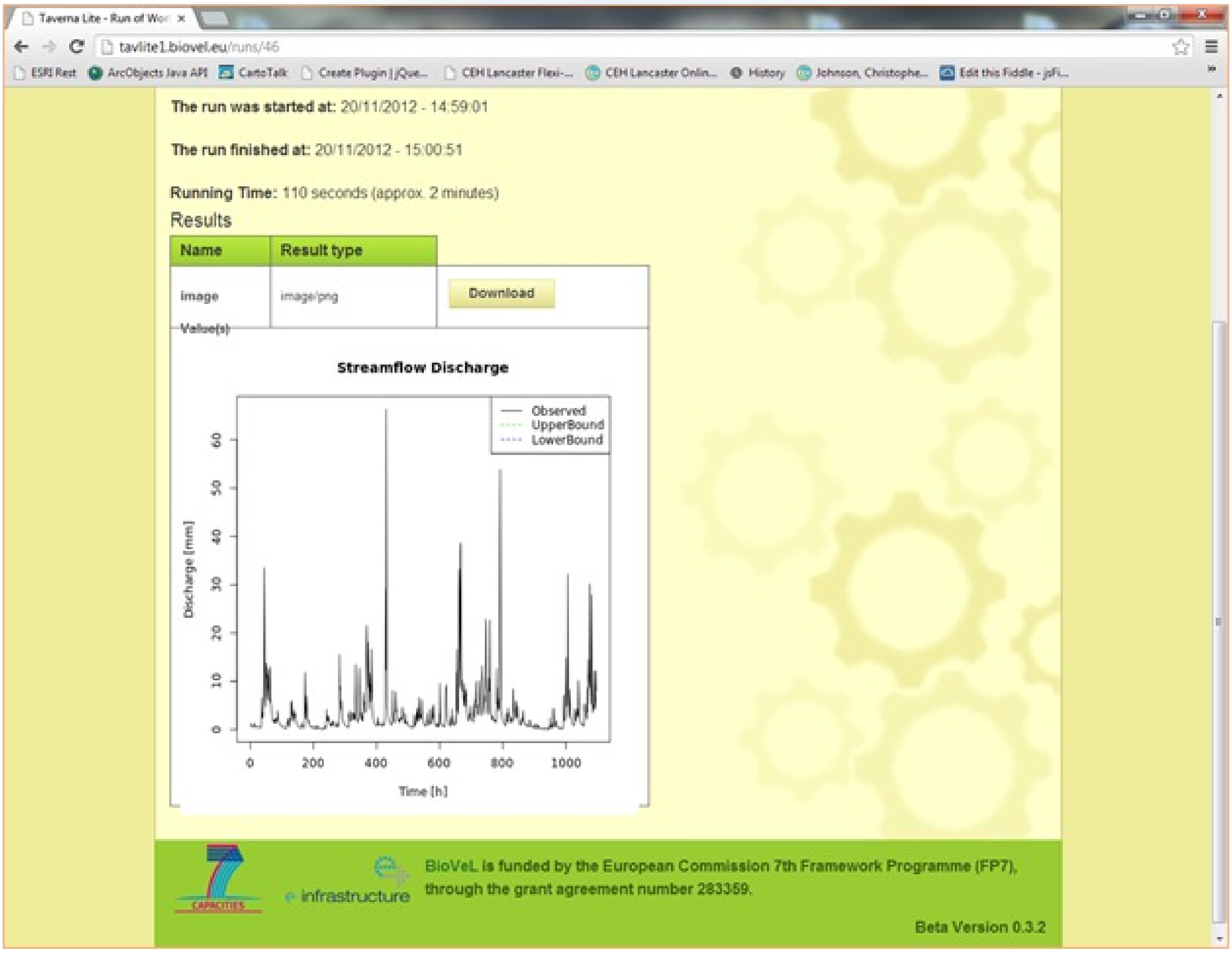


Figure 2.2 The Taverna workflow running the FUSE model as seen in the Taverna Workbench editor.



The run was started at: 20/11/2012 - 14:59:01

The run finished at: 20/11/2012 - 15:00:51

Running Time: 110 seconds (approx. 2 minutes)

Results

Name	Result type	
image	image/png	Download

Value(s)

BioVeL is funded by the European Commission 7th Framework Programme (FP7), through the grant agreement number 283359.

Beta Version 0.3.2

Figure 2.3 The resulting hydrograph output from a Taverna workflow run within the BioVel interface.

specialist within the consortium. These are referred to as the storyboard owners. Prototypes were developed based on these requirements, and are iteratively improved and built following verification (within the development team) and validation (with the storyboard owners, the wider consortium, and stakeholders) processes.

2.2.4 The portal

The EVO portal provides several means of finding information. The portal front page hosts a selection of paths to content and functionality. Furthermore, there are a myriad of info pages that introduce the models and data hosted on the portal along with associated links, FAQs and tutorials. Examples of the portable can be seen throughout this report (for example, see Figure 1.4 for the entry point to the portal).

The project team also investigated the role of workflows to enhance the capabilities of the portal, allowing scientists to create bespoke experiments connecting data and models together in a desired and repeatable pattern ready for execution in the cloud.

The experimentation focused on hydrological models to enable users to explore different management scenarios that might affect water resources in their location. A number of web services and a bespoke web page to run them were created. Building on this, a more flexible interface was developed using the Taverna Workflow Management System.

Figure 2.2 illustrates an example Taverna workflow that was created in the project. This workflow features a WPS service running an R version of the FUSE hydrological model for stream discharge in a catchment (within the Claudia script). The BioVel EC2 image was configured to run this workflow with the Taverna server to handle input parameters and the data loading. The workflow was then connected to a separate EVOp EC2 image to call the WPS running the hydrological modelling service.

Figure 2.3 illustrates the view for the user of the workflow when loaded into the BioVel Web interface. This hides the details of the connection of different cloud services to run the FUSE model. The resulting hydrograph of stream discharge is saved so that users can review and repeat previous workflow runs.

2.2.5 Exploring semantic links for EVOp data

The team investigated the use of GeoNames ontology and Web accessible database. The GeoNames database provides unique Web URIs for locations that can then be used to link to other web resources using Web APIs. The team entered 25 EVOp related sites in to GeoNames using the EVOP tag. These locations could then be linked through the GeoNames database to other resources such as Wikipedia entries for nearby locations (linked through DBPedia) and to Web services providing local weather readings (located through International Civil Aviation Organization (ICAO) sites in GeoNames). This facility to access related information to EVOp sites was developed as a Javascript tool but not included in the final EVOp portal partly due to the reliability of the GeoNames API during testing.

2.3 References

Coulouris, G., Dollimore, J., Kindberg, T., Blair, G.S. (2011). Distributed Systems: Concepts and Design. Pearson.

Zhang Q., Cheng, L., Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. Springer Journal of Internet Services and Applications (JISA). 1:1, 7-18.

Fielding, R.T. (2000). Architectural Styles and the Design of Network-based Software

Architectures. Ph.D. thesis, University of California, Irvine.

Related links

The NIST definition of cloud computing

http://csrc.nist.gov/publications/PubsSPs.html#800-145

Apache jclouds

http://jclouds.apache.org/

The Taverna Workflow Management System

http://www.taverna.org.uk/