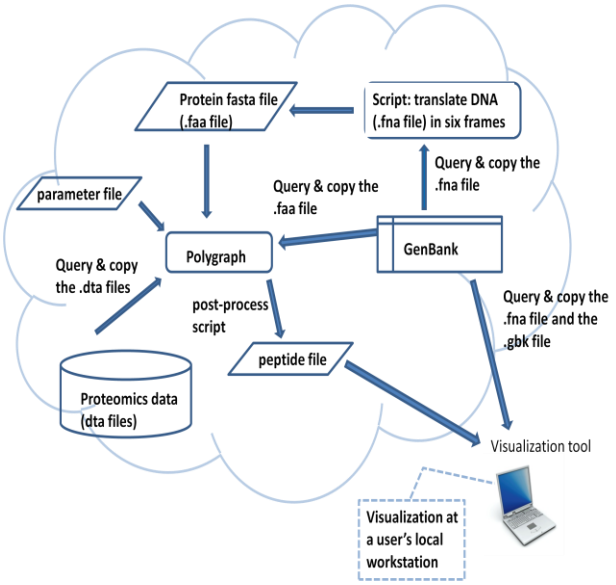
Pilot Project

Exploring Architecture Options for Workflows in a Federated, Cloud-based Systems Biology Knowledgebase

Ian Gorton, Yan Liu, Jian Yin, Leeann McCue, Bill Cannon, Gordon Anderson
Pacific Northwest National Laboratory
Richland, WA

Systems biology is characterized by a large community of scientists who use a wide variety of fragmented and competing data sets and computational tools of all scales to support their research. In order to provide a more coherent computational environment for systems biology, we are working as part of the Department of Energy Systems Biology Knowledgebase (Kbase) project to define a federated cloud-based system architecture. The Kbase will eventually host massive amounts of biological data, provide high performance and scalable computational resources, and support a large user community with tools and services to enable them to utilize the Kbase resources. We investigated the design of a workflow infrastructure suitable for use in the Kbase. The approach utilizes standards-based workflow description and open source integration technologies, and incorporates a data aware workflow execution layer for exploiting data locality in the federated architecture.



Use care scenario: Identify genome annotation

An overview of our use case for the Kbase depicting data sets and computations is shown in the Figure. A biologist retrieves data relevant to the organism under study from GenBank (a public database) and this is input into a script to translate the data into a format needed for the next step in the workflow. An Hadoop-based computation, Polygraph, is then invoked with the translated GenBank data and proteomics data. This produces a list or outputs (peptides) that can be fed into a desktop-based visualization tool that takes inputs data from the analysis and produces the spectrum of the genome where both published annotations and orphan peptides can be discovered.

We have implemented this workflow utilizing a federated cloud-based architecture consistent with the KBase architecture. In this prototype, we explored the necessary software architectures and technologies to enable data-location driven workflow for the Kbase. We coordinate the workflow by means of routing the workflow tasks to distributed Web services on a Cloud using REST APIs. The actual computation to be executed is determined by MeDICi pipelines that are driven by the data location and computation demands. In our experience, separating the workflow definition from its coordination is crucial to enable the extensible integration of the workflow with federated resources that are distributed across Cloud-based platforms.

The jBPM workflow definition tool was used as it allows customized workflow coordination to be easily deployed to its workflow engine, and interacts with REST APIs. Other scientific workflow tools such as Taverna**Error! Reference source not found.** and myExperiment**Error! Reference source not found.** are also candidate technologies for use as workflow tools in the UAL of the KBase architecture. Taverna provides a workbench to launch workflows published and shared on the myExperiment site. Taverna is efficient in accessing Web services defined with the WSDL and SOAP protocol. For generic REST APIs using HTTP protocol, some wrappers using Taverna's specific language need to be developed for Taverna to invoke REST APIs.

In our current work, we are investigating the necessary mechanisms and software tools required for integrating the infrastructure layer with UAL with an open specification. We are also extending our adaptive MeDICi prototype framework so that the abstract workflow definitions created by the users can be dynamically mapped to the underlying federated Cloud resources in a data-driven manner.