

Project Junior: Accelerating Chemical Informatics with Windows Azure

Paul Watson, David Leahy, Jacek Cala, Hugo Hiden, Dominic Searson, Vladimir Sykora, Simon Woodman,
Newcastle University, UK

Summary

In “Project Junior” we used the Windows Azure cloud to reduce the time taken to generate models that chemists use to predict the behaviour of molecules from 5 years to 3 weeks.

The Chemists in the project have unique software - the Discovery Bus (Figure 1) - that automatically builds Quantitative Structure-Activity Relationship (QSAR) models from chemical activity datasets. These models can then be used to design better, safer drugs, as well more environmentally benign products. Chemists use these models as it is much faster and cheaper than the alternative of carrying out experiments in the lab.

Recently, there has been a dramatic increase in the availability of activity data, creating the opportunity to generate new and improved models. Unfortunately, the competitive workflow algorithm used by the Discovery Bus requires large computational resources to process data. However, this is potentially an ideal Cloud application as large computational resources are required, but only when new datasets become available. Therefore, in the “Junior” project, we designed and built a scalable, infrastructure on the Windows Azure cloud in which the competitive model-building techniques are explored in parallel on 100 workers in order to massively accelerate model generation (Figure 2).

Professor David Leahy who is the Chemistry lead on the Junior project wrote:

“QSAR models are important as they allow chemists to predict the activity of molecules without the need to test them in the lab (which is expensive and time-consuming). We have a system – the Discovery Bus – that automatically generates QSAR models from chemical activity data using a novel competitive workflow approach. Using the Windows Azure cloud through e-Science Central has allowed us to generate 750,000 new QSAR models, which we have now been made freely available for anyone to use (at www.openqsar.com). The nearest comparison is 50 times smaller and has taken nearly 20 years to collate manually. Before project Junior, we thought this would be impossible - we estimated that it would take 5 years for our existing server to process the vast amounts of newly-available chemical activity data and generate new models. But by spreading the work over 100 Azure workers we were able to complete the run in less than 3 weeks.”

e-Science Central

Cloud computing can give scientists the computational resources they need, when they need them. To assist them, we developed “e-Science Central” (<http://www.esciencecentral.co.uk/>) – a Science-as-a-Service platform that combines three emerging technologies – Software as a Service (so users only need a web browser to do their science), Social Networking (to support collaboration) and Cloud Computing (to provide storage and computational power). Using only a browser, users can upload data, share it in a controlled way with colleagues, and analyse the data using either a set of pre-defined services, or their own, which they can upload for execution and sharing. Workflow editing and enactment through the browser is provided to allow automation of analysis.

The initial version of e-Science Central ran on our own internal servers, but in Project Junior we have ported it to use the Microsoft Azure Cloud. This has many advantages: we can now support a greater number of users, running more complex scientific analyses. And, from the point of view of users, it gives them the power of cloud computing, without them actually having to manage the complexity of developing cloud-specific software. We therefore used e-Science Central in Junior to simplify the process of moving the Discovery Bus onto Azure.

Future Developments

In the European Union funded Venus-C project we are further developing the system to allow greater scalability (if we can scale to 1000 Azure workers then the 3 week run would take 2 days) and flexibility. The scalable, competitive workflow pattern implemented on Windows Azure for the Discovery Bus is a good match to cloud computing as it requires large resources, but only irregularly (i.e. when new datasets or model-building algorithms become available). We are therefore working to apply it in other areas, including automated software checking.

Acknowledgements

The authors would like to thank Microsoft External Research for their financial and intellectual support to the project. We would particularly like to thank Christophe Poulain for the assistance he has given us, and to Savas Parastatidis for his original efforts to champion the project.

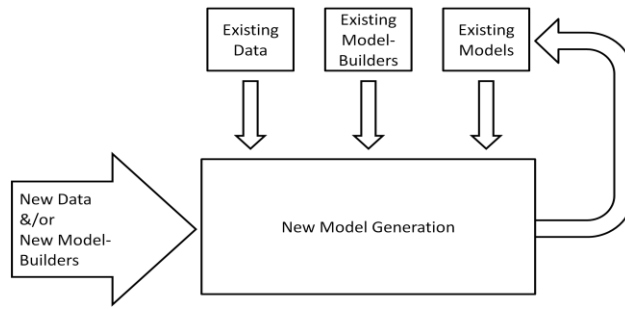


Figure 1. The Discovery Bus – Overview

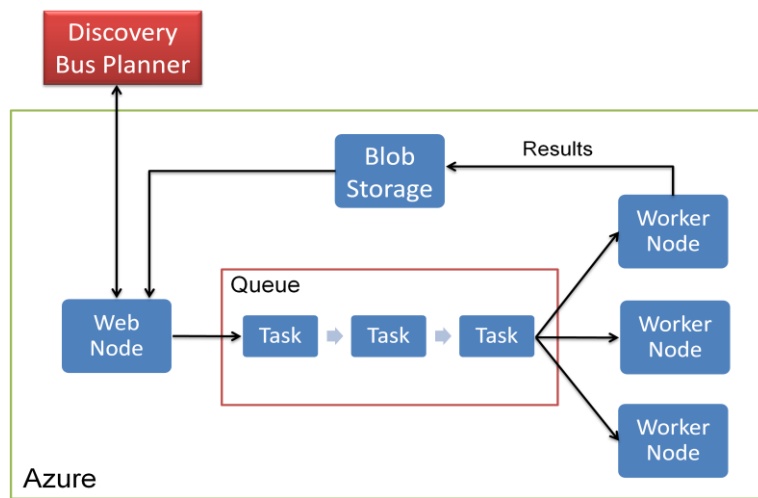


Figure 2 Processing Tasks in Azure