Knowledge Transfer Network

ICT

Case Study: Kindura cloud storage

By Gillian Law, 21 July 2011

There's a data explosion going on in the world of academia, and increasingly in business, too. With many Research Councils demanding that research data be kept – and be accessible for decades - there's a real need for a reliable, secure and cost effective way of managing long term storage.

The JISC-funded Kindura project was established in August 2010, with team members from King's College London, the Science and Technology Facilities Council (STFC) and the DuraSpace organisation. The aim is to pilot the use of a "hybrid cloud, shared service and in house model for providing repository-focused services to researchers across the partner institutions".

"What we're aiming for is to combine internal storage with commercial cloud, into a single framework," says project manager Simon Waddington, who is based at the Centre for e-Research at King's College London.

"Our storage middleware lets you mix different providers, and a brokerage layer decides where to put the data, based on a cost calculation, how many copies should be retained, where it should be stored geographically – all that information is derived from metadata entered by the user, allowing the brokerage layer to make the right decisions," he says.

While cloud is the word of the day – "people are talking about 'cloud' at practically every meeting I go to," Waddington says – there are still a lot of practical issues involved with using it properly.

"The Kindura project looks at a more cautious approach. You get people saying that yes, they'd like to use the cloud, but, well, they'd put some of the academic data on it, but they want sensitive, personal data, such as medical records, to stay in house, for example. There's a lot of decisions to be made about where each piece of data should go."

There's also a need for a reliable long-term storage solution for research data, he says, as much of the current work is being stored on personal drives, stored under desks – it might, theoretically, still be around in ten years but the chances of anyone being able to find and use it are slim.

"It's all very ad hoc at the moment, and we don't even have a clear idea of how much data there is, stored away around the country."

To create an affordable, reliable storage product, the Kindura team is therefore creating an 'inelastic' cloud. Jens Jensen, a researcher from STFC, admits there's a bit of a clash there.

“It’s true that it’s technically not a cloud - more like a traditional data infrastructure – but it looks like a cloud! We have a migratory layer sitting on top and a user interface where you can move your data into the real cloud, where you can do your analysis or whatever you need to do with it, then move it back out for long term archive into the Kindura cloud.

“Elasticity is the reason cloud is relatively expensive. If you skip the elasticity you can do it much more cost effectively – and that’s where we come in. We say let’s allocate a quota, say 100 terabytes, and within this project we can move data into that and store it in long term storage in a more cost effective way,” Jensen says. “Of course, if you only need up to, say, a few hundred gigabytes, it will still be elastic enough. It is inelastic in the sense that you hit a ceiling at some point. And the cost model is different.”

“Cloud does help in enabling the institutions to rapidly provision storage” says Waddington. “Previously researchers might wait for months for new storage hardware to be purchased and set up. With the cloud, this can be done in a matter of hours.”

With ‘data explosions’ expected in every field, there will be a real need for affordable, reliable storage, Jensen says.

“Everyone’s talking about next-generation sequencing in genomics, for example, where the sequencers are the size of printers and you can have one on your kitchen table, if you like. Soon everyone will generate so much data they won’t know what to do with it. And that’s true for practically all areas of research – humanities are increasingly digital, they’re finding that they need to scan large volumes of information, large images and so on. And by going digital you open up for more research – you can simply do more, and we’re trying to support that, by helping you manage the data,” he says.

The Kindura pilot is being implemented using an instance of DuraSpace’s DuraCloud as a mechanism for interfacing to multiple cloud providers. The Fedora repository provides the front-end application for researchers to store and retrieve their data. In addition, part of Kindura can be used to run an institutional repository, so researchers can have the best of three worlds: lots of replicas, keeping it locally, and “shopping around” for the cheapest service.

“It will be interesting to see the effect of combining cloud cost models with those of the traditional data centres. Kindura’s data migrator will be smart: it will work for you to minimise overall costs based on how long you need it in a commercial cloud service and whether it can be deleted from there or has to be moved out,” Jensen says.

Curation is also an important issue in the long term storage of data, he says.

“In a previous project called ASPIS we built a metadata infrastructure for provenance data, which could be tied into this as well. Then the Kindura cloud would know where the data is coming from and who has done what to it, what sort of processing has been done,” he says.

“Provenance information also helps us understand the value of datasets. For example, source data such as environmental measurements can’t be replicated, so we would want to replicate this data across storage providers to prevent loss. Of course this increases the cost. Data that can be replicated easily doesn’t need high replication. The cost optimisation in Kindura helps us solve these problems” says Waddington.

Long term, Jensen is confident that the work will be useful: “Clouds are good for many things, but will not meet everybody’s needs, particularly for larger data volumes or long term storage. Being able to combine the cloud model

with other data storage models will be useful."

Waddington says: "There has been a great deal of interest in Kindura from researchers and administrators at King's College and it is helping us to define our storage strategy. I also get frequent requests for information from other institutions who are concerned about preserving their research data. Kindura provides an efficient and flexible solution."