**Project Leapfrog**

Alignment with Reactor Sprint:

“RSE’s background in both computing and academic research puts them in the unique position of being to be able to **align the needs of the research community\* with the services and resources provided by The Cloud**.”

**Streams considered: Data Management, Data Lifecycle**

**\*Research community:**

Within the research community, there is a group of established academic researchers that have worked essentially without computing resources. These researchers are often focused on the basic sciences. They have significant amounts of data in personal storage and have struggled to analyse it on their local computers using fairly basic software. They worked in this way through the years and expressed no need for On-site HPC because they could not imagine scaling up to that level. They did not use research computing because they were trained in and comfortable with older research techniques. This resulted in them focusing on what they could do with local resources, often collecting far more data than they could analyse. They have been reluctant to mine this data because of the extensive and laborious effort involved in capturing, cleaning and reworking the data. The prospect of working with HPC, made them think that they needed to buy servers and or IT tech support. We think that if they are aware of the possibilities for research that The Cloud would offer them, they would be eager to use the resources directly.

**Objective:** Leapfrog these researchers over the HPC hurdle straight into The Cloud.

**How:** Tailor a learning pathway syllabus that takes them through the process to achieve the following goals:

1. The absolute basics:
   1. What is The Cloud?
   2. How does the internet-based computing resources support research?
   3. How is working with data different on The Cloud?
   4. How do I access The Cloud from my laptop?
   5. What software can I use in The Cloud that I cannot use on my desktop or can use differently with The Cloud resources?
   6. What are the advantages of these software?
   7. Presenting a survey of The Cloud based research methodologies
2. Intermediate:
   1. What are all The Cloud products offered in the Calculators (e.g. VM, Storage, App server, etc.)?
   2. How do I start designing a Cloud service that meets my needs by looking at what can be done with my research data?
   3. What support and current methodologies are there that are related to my field?
   4. How do I know how much of each product do I need?
   5. Understanding the costs.
   6. How do I install my data analysis software on The Cloud?
   7. Are there software options if my software does not work on The Cloud?
3. Advanced:
   1. How and where do I clean data?
4. Formatting data from obsolete formats
5. Formatting messy data into accessible formats
6. Uploading data into The Cloud
7. Automation of basic data management tasks
8. Using high speed and power computing in data analysis
   1. What is data management on The Cloud (e.g. Open Access, Open Science, etc.) to work in my discipline (e.g. basic sciences, advanced sciences)?
   2. What data management tools help me achieve the above?
   3. How do I use these tools?
   4. How do I link my data with the world’s data sets?
9. Creating data in compatible standards
10. Merging smaller or more specific data into generalized or larger datasets
11. Creating analysis at greater scale.

This would take us through the Data lifecycle of:

1. Creation/acquisition
2. Data management and research
3. Archiving, sharing and preservation

*Data Linking is a powerful research data share discipline. This fits in with the new product introduced at the Summit: “publicdataset.azurewebsite.net”*

**Challenges with the current course on the learning pathway:**

These courses do not link to the environments the researcher use currently or show them how they could do faster research operations on The Cloud. They sound very technical and can be intimidating to Researchers. The researchers cannot immediately see that the investment would enable them to work faster or better. It would be a matter of delivering the same content, in a way Researchers are comfortable but it would also be to focus on the software and the possibilities of the software.

Based on the “HOW” section above, there may be gaps in the literature that need to be filled.

**Proposal:**

1. Establish if the syllabus presented in the “HOW” is complete.
2. Review the current course that would fit the syllabus.
3. Identify gaps:
   1. “missing” courses
   2. Technical course that need to be repackaged for the research community
4. Compile the syllabus into a learning pathway.
5. Produce a case study that focuses on data cleaning and data mining.