Trust but verify: fisheries science in the cloud

eResearch NZ 2016

Dr Finlay Thompson
Queenstown, 9 February 2016

DRAGONFLY
Data Science

Outline

1. Problem — increase efficiency of fisheries research

2. Solution – trust, but verify

3. What we have learnt

Applied Bayesian Modelling
Matrix Population Models
R Graphics
Real World Haskell
Bird Distribution
Marine Mammals
Bayesian Inference

Problem – increase efficiency of fisheries research

Background

New Zealand fisheries

New Zealand’s fishing industry exports around $1.5 billion

The value of all quota is estimated at over $4 billion

The quote management system has created a valuable asset.

Background

New Zealand fisheries

Fisheries science is directly supported by fishing activity.

Often this means research is restricted to high - value fisheries.

Difficult to support science work for long tail of lower value fisheries.

Problem

“How to support fisheries science to become more ***efficient****”* ,

Problem

*"How to support fisheries science to become more **efficient**"*,

We also need to ensure that fisheries science is **trusted**.

Problem

*"How to support fisheries science to become more **efficient**"* ,

We also need to ensure that fisheries science is **trusted**.

And, ideally, encourage a community that fosters **collaboration**.

The problem of restricted data

Restricted data sets

A big source of delay comes from the manual handling of requests for restricted data.

The problem of restricted data

Restricted data sets

A big source of delay comes from the manual handling of requests for restricted data.

Fisheries data is commercially sensitive. Data security is very important.

The problem of restricted data

Restricted data sets

A big source of delay comes from the manual handling of requests for restricted data.

Fisheries data is commercially sensitive. Data security is very important.

However, existing processes of managing access are cumbersome, and costly.

Opportunities

Restricted data sets

Opportunities to increase efficiency of managing data access include:

- standardised data preparation processes
- bringing multiple datasets together, creating a single interface,
- semi - automatic data access authorisation.

Collect data sets together

Principles

Research often involves bringing together different data sets into a relational database system.

Collect data sets together

Principles

Research often involves bringing together different data sets into a relational database system.

In fisheries research, these data include:

- Fisher reported catch effort and landings,
- Sampling data from landings,
- Government and fishery observer reports,
- Vessel management system records (GPS)

Collect data sets together

Principles

Research often involves bringing together different data sets into a relational database system.

In fisheries research, these data include:

- Fisher reported catch effort and landings,
- Sampling data from landings,
- Government and fishery observer reports,
- Vessel management system records (GPS)

Avoid ***spreadsheet hell*.**

Standardise data preparation

Principles

Researchers always need to update data, and this should be done in a published and standardised way, such as:

- fixing missing values, typos, and other clear mistakes,
- derivation of standard derived values, and,
- imputation of unknown values.

Semi-automatic data authorisation

Principles

Authorisation should be managed at the level of contracting.

Semi-automatic data authorisation

Principles

Authorisation should be managed at the level of contracting.

Streamlined access should be **usual**,

Semi-automatic data authorisation

Principles

Authorisation should be managed at the level of contracting.

Streamlined access should be **usual**, but with the ability to **review** access used.

Solution – trust, but verify

Trust, but verify

Trust researchers, but watch what they do.

TRUST BUT VERIFY

Introducing the Kahawai reporting system

Developed by Dragonfly Data Science for Trident Systems.

Introducing the Kahawai reporting system

Developed by Dragonfly Data Science for Trident Systems.

Based on our experience at Dragonfly.

Introducing the Kahawai reporting system

Developed by Dragonfly Data Science for Trident Systems.

Based on our experience at Dragonfly.

Applying, and automating, ideas we borrowed from software development.

The Kahawai reporting system

The **kahawai** system integrates:

- a standard set of research databases in **PostgreSQL**,
- a safe, isolated, compute environment based on **docker**,
- a continuous deployment platform, integrated with **GitHub**.

The Kahawai reporting system

The **kahawai** system integrates:

- a standard set of research databases in **PostgreSQL**,
- a safe, isolated, compute environment based on **docker**,
- a continuous deployment platform, integrated with **GitHub**.

Bit like TravisCI, or Jenkins, but for fisheries research.

Standard set of research databases

Kahawai provides access to a range of research databases, over an internal network.

Standard set of research databases

Kahawai provides access to a range of research databases, over an internal network.

Data preparation and imputation scripts are incorporated into the research databases.

Standard set of research databases

Kahawai provides access to a range of research databases, over an internal network.

Data preparation and imputation scripts are incorporated into the research databases.

Acts as a platform for collaboration.

Using docker to manage computation

Researchers like to use a range of different software packages, typically at many different versions.

Using docker to manage computation

Researchers like to use a range of different software packages, typically at many different versions.

Docker allows researchers to define their own favourite installation, built on top of Linux.

Using docker to manage computation

Researchers like to use a range of different software packages, typically at many different versions.

Docker allows researchers to define their own favourite installation, built on top of Linux.

Kahawai **builds** docker images with network access, so can download library code from internet.

Using docker to manage computation

Researchers like to use a range of different software packages, typically at many different versions.

Docker allows researchers to define their own favourite installation, built on top of Linux.

Kahawai **builds** docker images with network access, so can download library code from internet.

However, each job **runs** without network access.

Continuous deployment

Continuous deployment is the idea that code should be continuously, and automatically, integrated and deployed.

Continuous deployment

Continuous deployment is the idea that code should be continuously, and automatically, integrated and deployed.

In practice, when code is pushed to the code repository, a new integration run is started.

Access driven by code version control

Access to the system is provided through code version control.

Access driven by code version control

Access to the system is provided through code version control.

Researchers need to write code.

Access driven by code version control

Access to the system is provided through code version control.

Researchers need to write code.

All work is open to review.

What we have learnt

Running since December 2014

The Kahawai reporting system has been running successfully for over a year.

Running since December 2014

The Kahawai reporting system has been running successfully for over a year.

Related processes have been used at Dragonfly for nearly eight years.

Running since December 2014

The Kahawai reporting system has been running successfully for over a year.

Related processes have been used at Dragonfly for nearly eight years.

Is currently been used intensively, regular computer upgrades have been required, and more planned.

Efficiency gains have been made

We have seen an increase in the efficiency of individual researchers.

Efficiency gains have been made

We have seen an increase in the efficiency of individual researchers.

More research outputs are being produced, and at a cost that allows the support of lower value fishing stocks.

Greater transparency

By requiring code to be checked in, we get greater transparency on what researchers are doing.

Greater transparency

By requiring code to be checked in, we get greater transparency on what researchers are doing.

Impossible to make “untraceable edits” to datasets.

Greater transparency

By requiring code to be checked in, we get greater transparency on what researchers are doing.

Impossible to make “untraceable edits” to datasets.

It is possible, and useful, to review a research output, all the way back to the raw data.

Reproducible research

Science is based on the notion of repeatable research.

Reproducible research

Science is based on the notion of repeatable research.

It is also very effective and useful.

For example, when a manager wants to re-run an analysis, but can’t get the attention of a busy scientist.

Enhanced collaboration

Kahawai reports need to *work*, in the sense that they compile and run successfully.

Enhanced collaboration

Kahawai reports need to *work*, in the sense that they compile and run successfully.

This makes it easy for other researchers to collaborate on a project, working together on a single code base is easy.

Enhanced collaboration

Kahawai reports need to *work*, in the sense that they compile and run successfully.

This makes it easy for other researchers to collaborate on a project, working together on a single code base is easy.

The Kahawai reporting system takes care of integrating and building the project.

Flexibility in choice of tools

Docker is a light - weight virtualisation tool. Built on top of Linux.

Flexibility in choice of tools

Docker is a light - weight virtualisation tool. Built on top of Linux.

Researchers frequently use a wide range of software libraries, that need to be installed together in a coherent way.

Flexibility in choice of tools

Docker is a light - weight virtualisation tool. Built on top of Linux.

Researchers frequently use a wide range of software libraries, that need to be installed together in a coherent way.

Each report uses a separate docker image, so no clashes.

Easy access to cloud compute

Jobs are run on remote compute resources. Currently these are in a private cloud.

Easy access to cloud compute

Jobs are run on remote compute resources. Currently these are in a private cloud.

The reporting system makes it trivial to run projects, repeatedly, on the cloud, freeing up local computers.

Next steps

We are building a clone of Kahawai, focused on the general public.

Next steps

We are building a clone of Kahawai, focused on the general public.

It is currently code names “Gorbachev”.

Next steps

We are building a clone of Kahawai, focused on the general public.

It is currently code names “Gorbachev”.

We recently used our prototype to run 9000 simulations, for a total of three years of compute time, over 2 months, using the Amazon spot market.

Thanks

Thanks to the David Middleton and Trident Systems for supporting the Kahawai project.

Thanks to the Ministry for Primary Industries for allowing this development to occur.

Thanks to all the amazing open source tools that make this work possible.