THE UNIVERSITY of ADELAIDE

adelaide.edu.au

DATA FARMING USING CSIRO WORKSPACE

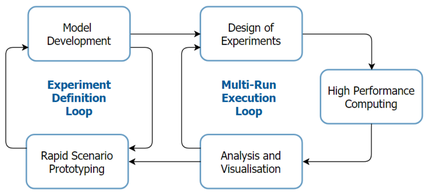
A vehicle for deploying Data Farming capability to end-users

Daniel Cotton, Jaskirat Grover, Nehal Jain, Dillon Thyer, (daniel.cotton, jaskirat.grover, nehal.jain, dillon.thyer)@student.adelaide.edu.au
Supervised by Ali Babar, Anthony Cramp, Meredith Lane, Nguyen Tran, Christoph Treude and Andrew Warhust
School of Computer Science

Background

Data Farming

- Data farming is a concept for "growing" data by exploiting High Performance Computing to run simulations of models of the real-world across a representative subset of potentially billions of combinations of "seed" factors
- Analysis and visualization provides rapid interpretation of results



CSIRO Workspace

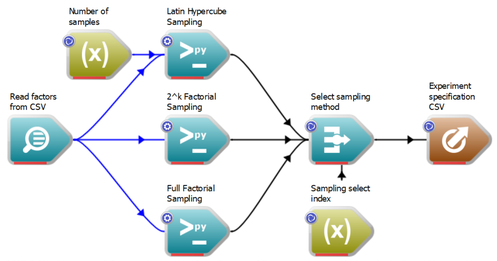
- CSIRO Workspace is a powerful, cross platform scientific workflow framework that enables collaboration and software reuse

Objectives

- Deploy data farming functionality to end-users using CSIRO Workspace
- Build CSIRO Workspace workflows for the multi-run execution loop components: Design of Experiments, High Performance Computing and Analysis and Visualisation
- Demonstrate the multi-run execution workflow using a Wolf Sheep Predation NetLogo model that explores the stability of predator-prey ecosystems

Future Work

- Currently, the HPC workflow supports NetLogo models only so we propose that future work investigates how to generalize this workflow further to enable wider use of models
- Integrating current data farming functionality as Workspace modules/plugins



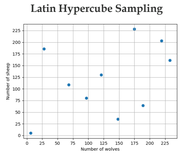
CSIRO Workspace workflow implementing three Design of Experiments sampling methods: Latin Hypercube, 2^k Factorial, and Full Factorial

- The Design of Experiments workflow takes a set of model input factors and offers flexibility in choosing how these factors are sampled
- Users are not limited to using the sampling methods provided and can add their own via simple Python implementation

DOE Sampling Methods

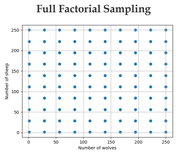
- Sampling methods are used to generate a subset of all possible combinations of inputs accepted by a model

Latin Hypercube Sampling



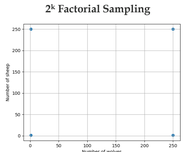
Generates a near-random sample of parameter values from a multidimensional distribution

Full Factorial Sampling



Consists of 2 or more factors where each factor has k levels for which all possible combinations are produced

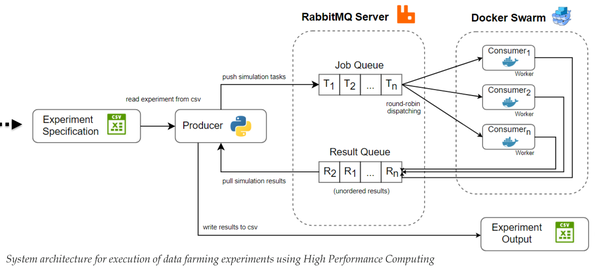
2^k Factorial Sampling



A special case of the general factorial design where there are k factors, each with 2 levels

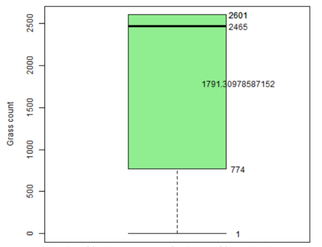
Experiment Execution

- The simulation engine environment is encapsulated in a Docker Image
- A Docker Swarm manager is used to automatically deploy the simulation image to a number of Docker Hosts
- The simulation model is read and distributed to each node using RabbitMQ's fanout exchange
- A Workspace workflow reads in the experiment specification generated by the DOE workflow and pushes each set of values to the job queue
- Worker nodes pull from the job queue, execute the given simulation and push the results to the result queue
- The workflow continuously receives results and writes them to an output file

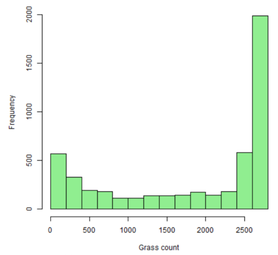


System architecture for execution of data farming experiments using High Performance Computing

- The experiment workflow supports NetLogo models in general with configurable options such as: maximum iterations per simulation and output variables of interest
- The number of simulation tasks per worker node is configurable and advised to be as large as possible to ensure full utilization of multiprocessing



Boxplot of final grass counts for the Wolf Sheep Predation model (5000 simulation configurations)



Histogram of final grass counts for the Wolf Sheep Predation model (5000 simulation configurations)

- R provides functionality, that can be integrated into CSIRO Workspace, for statistical analysis of the results generated by the HPC stage

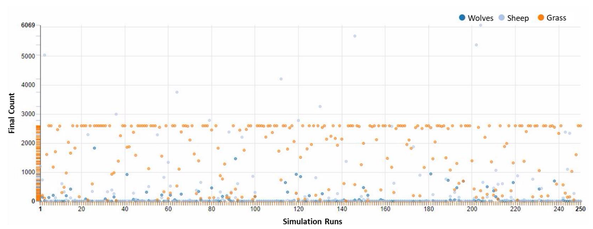
Design of Experiments → **High Performance Computing** → **Analysis & Visualisation**

Design of Experiments (DOE) refers to statistical experiment planning intended to cut down on the number of samples that need to be explored to provide reliable results

High Performance Computing (HPC) provides the means to explore the potentially millions of samples provided from the DOE stage in a reasonable timeframe

The **Analysis and Visualisation (AVIS)** component makes use of various techniques and tools to aid in the translation of large output datasets into information that helps answering questions at hand

- CSIRO Workspace supports key visualisation concepts, focus (view and project manipulation) and linking (simultaneous visualisation exploration), through inbuilt 2D plotting



Scatterplot in Workspace showing a subset of the count of grass, sheep and wolves after a fixed number of iterations for different configurations of the Wolf Sheep Predation model