Massive Model Output Visualisation

Shaun Verrall

www.github.com/ver078/output-hpc-vis

AGRICULTURE AND FOOD
www.csiro.au

I am a software engineer working on Agriculture Production SIMulator (APSIM) software. APSIM is a process based model aimed at simulating crop and animal growth at the paddock scale. It outputs customisable reporting data in space delimited text files at daily, month, yearly or crop event time scales. APSIM is used to create simulations that mimic the real world to generate what if scenarios for farming management. Output is compared against real world experiments in order to improve the process based science in the model.

My Synthesis Project

Clusters are currently been used more and more to generate massively large data sets of output files. It is not uncommon to run millions of simulations on clusters with millions of corresponding output files. These large numbers of output files pose a serious problem in analysis of the data. Standard approach when analysing data is to first plot or visualise the data and then once an understanding of your data is obtained then filter out only the data of interest. This approach can no longer be maintained as the sheer quantity of data causes problems for visualisation tools.

My approach is to tip the standard approach on its head, and filter first using customisable sensibility filters applied to each output file and then visualise the data that either passes or fails these tests. In order to speed up the filter process this is done in parallel using the same sort of clustering technology that originally ran the simulations.

My Digital Toolbox

The CSIRO High Performance computing cluster using Slurm array jobs is used to run a customised R script in parallel on every output file in the dataset. Standard R plotting packages are used to plot the data eg. ggplot.

I did not know R before data school or how to run jobs on the CSIRO High Performance Computer cluster.

Favourite tool

The Git Bash command line has been a surprise favourite of mine. Lets you combine the power of command line Git with the standard tools of ssh for logging in and running jobs on the cluster.

My time went …

Setting up the Slurm array jobs. Automating this part is proving most difficult.

Next steps

Create an R package that allows users can easily install. This will allow users from R studio running on their desktop to run the filtering on the cluster and then plot and visualise the results in R studio on their desktop.

Possible idea to test would be to visualise the data spatially on a map of Australia using leaflet or as a shiny app that uses javascript to select points on the map and display more detailed information about the point selected.

Sample Data

Single Simulation Data (this x 600000)

Name	Type	Value
sim	list [2]	List of length 2
head	list [6 x 2] (S3: data.frame)	A data.frame with 6 rows and 2 columns
out	list [62 x 62] (S3: data.frame)	A data.frame with 62 rows and 62 columns

Meta Data (head)

Filter

	name	value
1	phenology	late
2	sowing	01-Apr
3	nitrogen	40
4	site	10007
5	soilwater	1
6	genotype	CanopyDevelopment-7

Single Ouput File (out)

Filter

	SimulationName	SimulationID	CheckpointID	CheckpointName	Clock.Today	Wheat.AboveGroundWtAtFlowering	Wheat.AboveGroundWtAtMaturity	Wheat.Grain.HI	Wheat.Grain.Number	Whea
8	10007	1	1	Current	1957-11-10	504.349	800.169	0.283	14441.354	
9	10007	1	1	Current	1958-11-21	355.445	364.678	0.078	10459.149	
10	10007	1	1	Current	1959-11-01	381.842	802.252	0.386	10425.998	
11	10007	1	1	Current	1960-11-13	552.635	910.590	0.297	15615.336	
12	10007	1	1	Current	1961-10-30	676.510	1249.361	0.328	18180.895	
13	10007	1	1	Current	1962-11-11	15.872	15.724	0.006	364.361	
14	10007	1	1	Current	1963-11-06	347.925	402.447	0.119	9777.356	
15	10007	1	1	Current	1964-11-12	56.888	84.922	0.284	1393.906	
16	10007	1	1	Current	1965-11-10	295.558	319.906	0.110	8073.735	
17	10007	1	1	Current	1966-11-16	202.195	231.004	0.159	5345.289	
18	10007	1	1	Current	1967-11-07	373.061	381.044	0.063	10596.790	
19	10007	1	1	Current	1968-11-17	163.900	217.948	0.259	4152.410	
20	10007	1	1	Current	1969-11-03	352.404	359.980	0.084	9987.783	
21	10007	1	1	Current	1970-11-08	222.196	359.948	0.338	5878.417	
22	10007	1	1	Current	1971-11-20	300.699	471.006	0.232	8210.045	
23	10007	1	1	Current	1972-11-12	16.706	16.346	0.011	370.630	
24	10007	1	1	Current	1973-11-15	20.458	19.835	0.017	466.372	
25	10007	1	1	Current	1974-11-08	335.484	367.036	0.106	9342.043	
26	10007	1	1	Current	1975-11-12	484.162	651.172	0.192	13525.470	
27	10007	1	1	Current	1976-10-30	56.366	56.481	0.084	1069.035	
28	10007	1	1	Current	1977-10-28	212.986	212.443	0.069	5856.263	
29	10007	1	1	Current	1978-11-13	274.591	309.802	0.108	8127.995	
30	10007	1	1	Current	1979-11-06	193.190	369.769	0.370	5283.565	
31	10007	1	1	Current	1980-11-05	198.467	211.841	0.110	5443.201	

Filtered Output File (with Provenance maintained)

Name	Type	Value
sim_filtered	list [3]	List of length 3
00e2ced275084228a4922c614e31dad7_1_meta	list [6 x 2] (S3: data.frame)	A data.frame with 6 rows and 2 columns
00e2ced275084228a4922c614e31dad7_1_data	list [3 x 62] (S3: data.frame)	A data.frame with 3 rows and 62 columns
00e2ced275084228a4922c614e31dad7_1_num_rows	integer [1]	3