PRIMAVERA Data Management and Analysis using JASMIN

Jon Seddon[1], Ag Stephens[2], Malcolm Roberts[1], David Hein[1]

[1] Met Office, UK. [2] Science and Technology Facilities Council (STFC), UK.

PRIMAVERA is a European Union Horizon2020 funded project that aims “to develop a new generation of advanced and well-evaluated high-resolution global climate models, capable of simulating and predicting regional climate with unprecedented fidelity, for the benefit of governments, business and society in general”.

PRIMAVERA includes the running of seven climate models with common initial conditions and common forcings at low and high-resolutions to generate a multi-model ensemble of climate simulations. The volume of data from these simulations is estimated at 2.5 petabytes. The simulations will be run on High Performance Computers (HPCs) throughout Europe. The project required a central location where data from these simulations could be brought together and analysed. JASMIN's fast disk storage, tape archive, extensive compute resources and high-bandwidth connections to Europe means that it is the ideal platform for PRIMAVERA's data management and analysis requirements. JASMIN will allow the 100 PRIMAVERA scientists from across Europe to collaborate together on this multi-model ensemble of high-resolution climate data to improve future climate simulations.

The Data Challenge

The PRIMAVERA stream 1 simulations will produce 2.5 petabytes of high resolution model data. The project has 440 terabytes of group workspace storage available to it.

Data Management Plan

Figure 1 shows the workflow of data that has been devised for PRIMAVERA. Data is transferred across the Internet from the HPCs throughout Europe to JASMIN group workspaces using JASMIN’s high-performance data transfer service.

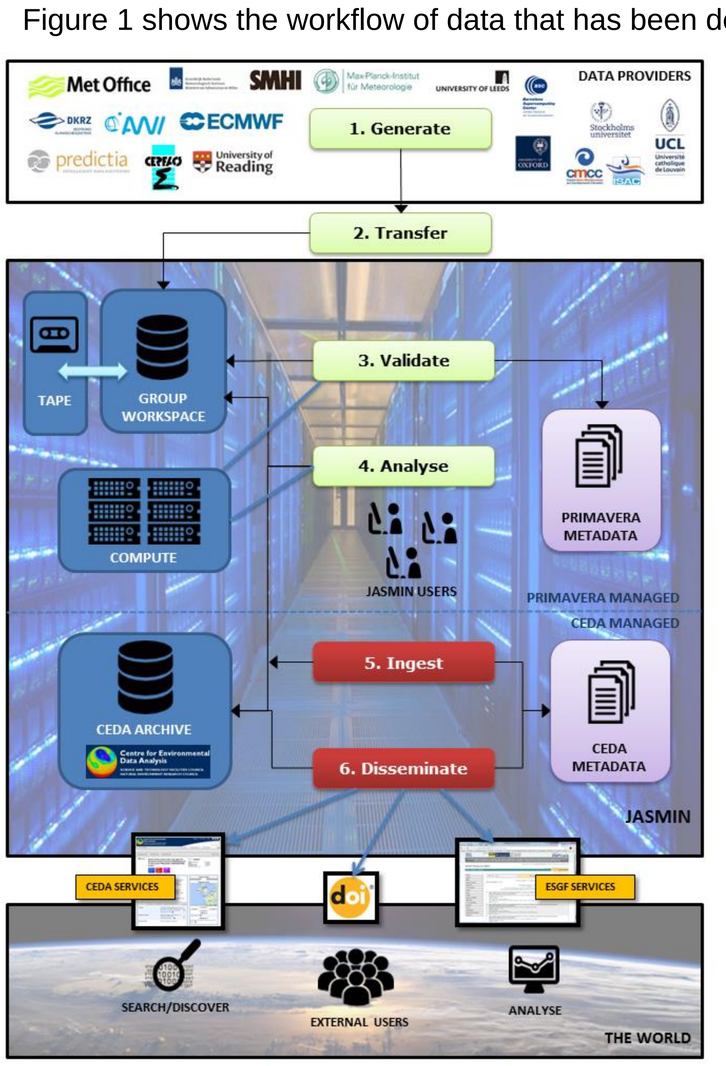

Figure 1. The work flow devised for the PRIMAVERA data.

On arrival at JASMIN, data is validated and the metadata recorded in the database. The files are then written to elastic tape, making disk space available for further chunks of data to be transferred to JASMIN.

Once all time periods for a variable have been received, this variable will be ingested to the CEDA archive and made available to the wider community through the ESGF.

Data Management Tool

A Data Management Tool (DMT) has been developed and implemented on a server in the JASMIN private cloud. The DMT consists of a database to store the metadata from each validated file and a web interface to allow data providers and data users to query the received data. The interface and database have been developed by PRIMAVERA funded staff at CEDA and the Met Office. It is implemented using the Django web framework, Python and a PostgreSQL database.

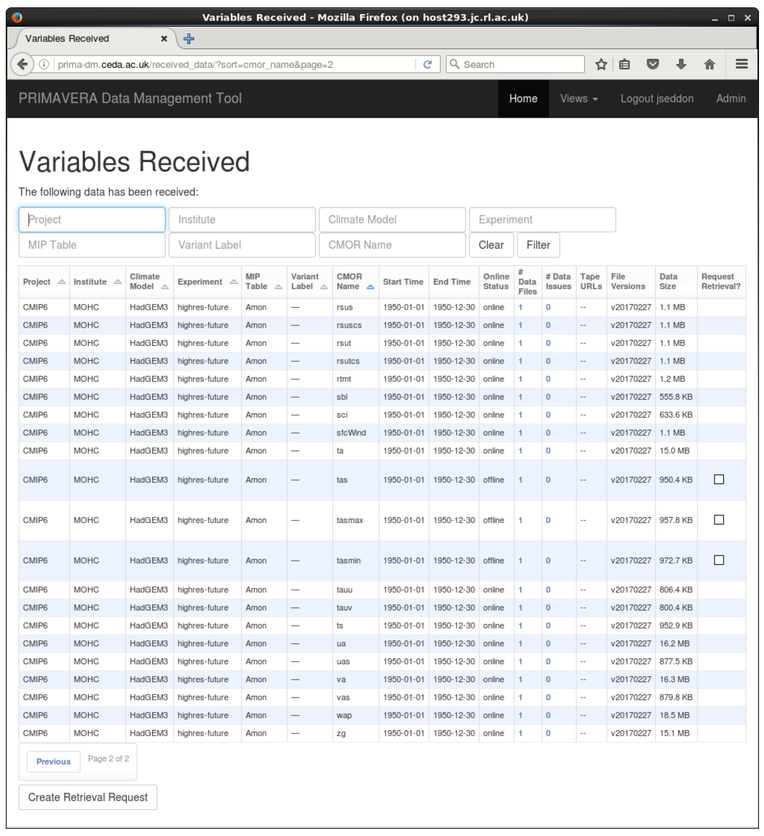

Figure 2. The Data Management Tool’s web interface.

Scientists can use the DMT’s web interface to query what data is available. If the required data is only on tape then they can use the web interface to request that it is restored to disk. Figure 2 shows one of the DMT’s queries available to scientists; the data that has been received so far can be seen, along with its location and the facility to request that it is restored to disk.

Data Analysis at JASMIN

The combination of the fast storage, interactive analysis servers and the LOTUS compute cluster allows PRIMAVERA scientists to bring their analysis to the data. There is no longer a need for scientists to download a copy of the data to their home institutes. However, because all of the data cannot be held on disk at once, the workflow shown in Figure 3 is necessary. This workflow is not as convenient as having all of the data set constantly online, but this is not possible because of the size of PRIMAVERA’s high-resolution data sets.

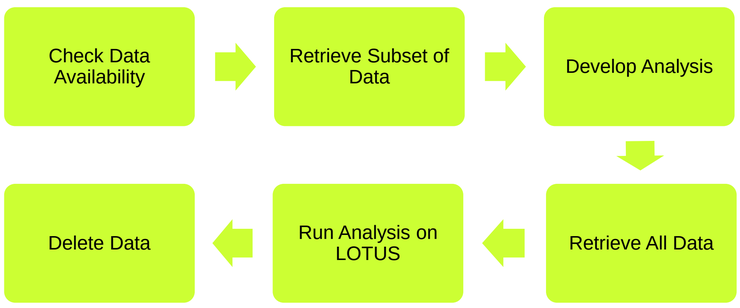

Figure 3. The workflow for analysing PRIMAVERA data.

Conclusions

JASMIN is the ideal environment to host the data storage and analysis facilities for the PRIMAVERA project because:

- its fast connections to the Internet allow the data to be rapidly brought together in one location;
- users can bring their analysis to the data, rather than having to download their own copy of the data, which isn’t feasible in a data set of this size;
- the JASMIN cloud allows custom user interfaces to the data to be developed;
- the JASMIN staff have the expertise to get the most out of the facility and out of the PRIMAVERA data.

Met Office Hadley Centre, FitzRoy Road, Exeter, Devon, EX1 3PB United Kingdom
Tel: +44 1392 886236 Fax: +44 1392 885681
Email: jon.seddon@metoffice.gov.uk