IT Architecture of the MARS Crop Yield Forecasting System

Rob Lokers, Daniel van Kraalingen and Hendrik Boogaard

Wageningen UR, Alterra-CGI, P.O. Box 47, 6700 AA Wageningen, the Netherlands. Tel. +31-(0)317-481850. Email: rob.lokers@wur.nl

Introduction

The Crop Growth Monitoring System (CGMS) provides operational services and analysis tools to the Joint Research Centre of the European Commission (JRC) in the area of crop monitoring and crop yield forecast, as part the MARS Crop Yield Forecasting System. For an overview of the global structure of the CGMS processes we refer to the article of Boogaard et al. also published in this issue.

The operational services of CGMS are fully automated. The required technology needed to support this critical system has through the years been on the edge of technological possibilities. Recent developments that demanded for innovative IT solutions are:

- A vast increase of the functional scope of the system
- An extension of the supported regions of interest
- Improvement of spatial and temporal resolutions of datasets (e.g. weather data)

This article focuses on three aspects of the CGMS: issues related to data and data processing, the IT environment that has been implemented, and the developed tools for viewing and analyzing data.

Data and data processing

The amount of data involved in the processing of the CGMS has grown tremendously in the past years. This was caused by several developments:

- The spatial coverage of the system was extended from Europe to a global coverage by the addition of several new regions of interest (e.g. Asia, the American continent).
- Higher resolution input data have become available and higher resolution outputs are demanded for analysis tasks.
- The extension of the system with the capability to use weather forecast data as input for the model calculations has heavily increased the amount of input and model data, because a large number of possible realizations of future weather need to be processed.

Currently, the expected total size of the datasets that are required to operate the system with all its planned extensions is about 15 Terabytes. Managing this amount of data, timely processing of model results and provision of the required data to the end users requires specific measures in the area of data handling and processing.

Structure and timing of the various process steps within CGMS is complex. Various operations are performed in parallel, results of these operations have to be combined in later steps and the implemented periodical cycles require that results from previous runs are available and valid. A project management board provides administrators with the status of the different process steps through time and provides an overall view on the status of the system. This includes not only the internal status of operations, but also the status of preparatory external process steps (e.g. the processing of raw weather data by the external data provider).

A system for distributed computing of independent subtasks was developed to be able to timely process all processing tasks using multiple CPU's or machines. Coordination is carried out through a database-centric architecture which avoids a technically more complex system for inter-process communication. A central process is responsible for the creation of new processing tasks in a task status table based on the delivery of new data. The distributed calculation processes poll this task table to see if tasks are waiting to be processed by the particular process and update the status of tasks after processing.

Data storage and data handling

The hardware configuration consists of a database server, separate disk storage and several servers (CL-x) responsible for all data processing tasks, servers for data caching (MS-x) and a web server hosting the web enabled viewers as shown in Figure 1.

The database server has a high speed database connection to a storage area network (SAN) EVA-6000. On the server Oracle

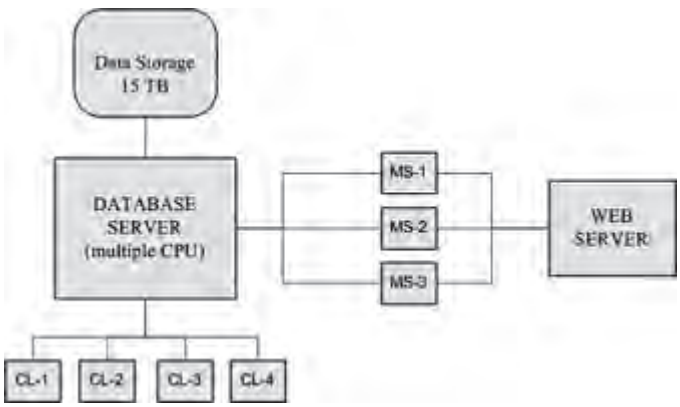

Figure 1 Overview hardware

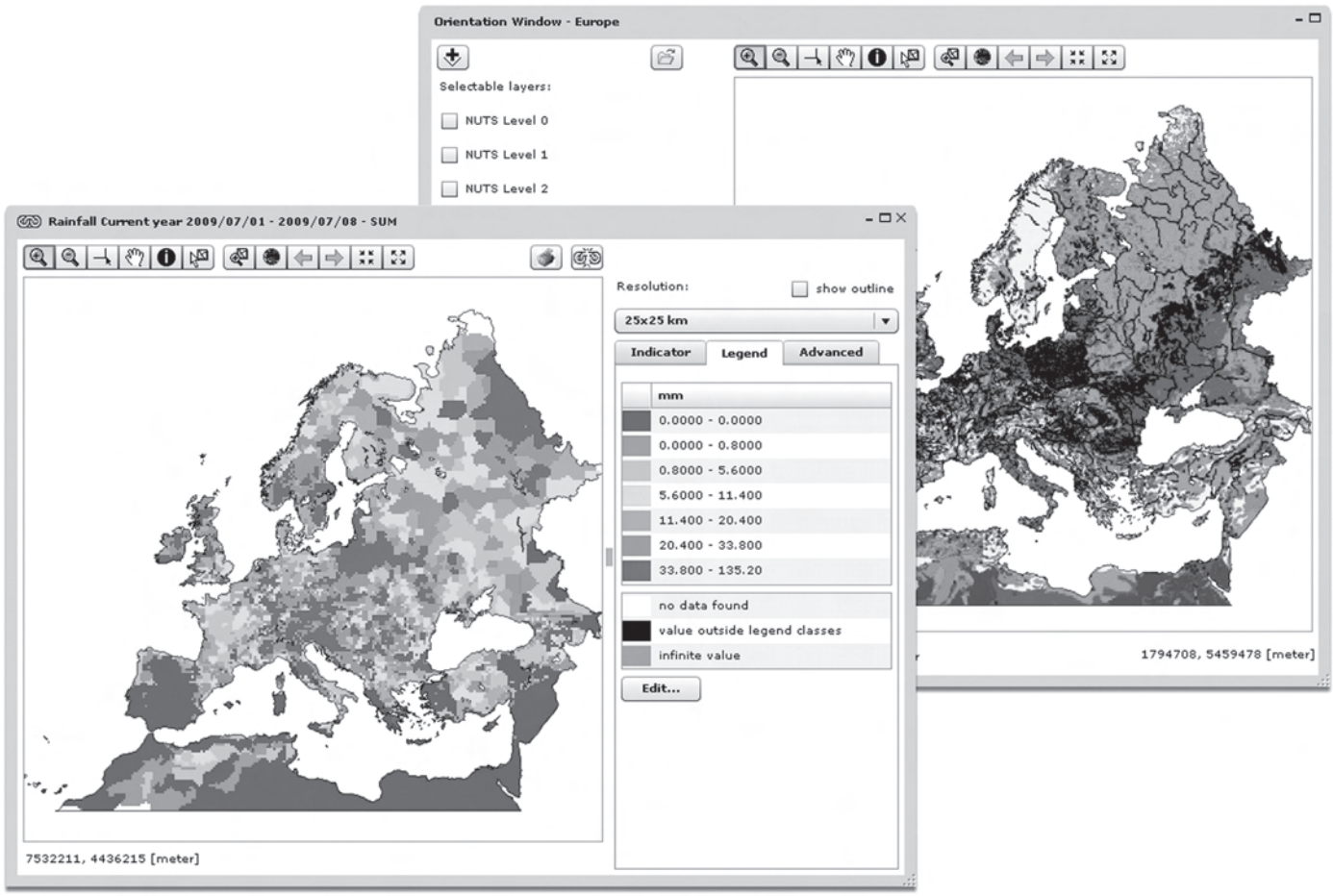

Figure 2 Example of the CGMS map viewer

11g is installed on a Red Hat Enterprise Linux (RHEL) version 5 Operating System. Because of the required processing capacity at peak times four data processing servers are available, hosting all CGMS related software, to ensure products are ready according the required time schedule. In addition five servers are necessary to host the web enabled viewers and to cache frequently asked data sets (see next paragraph). All these servers are located in the same failsafe hosting environment as the database server.

End users of the CGMS require fast access to recent weather and crop data. It concerns all kinds of temporal aggregations of basic or derived indicators or statistical functions on these indicators. To provide the required information to the end users quickly, optimization of the data access procedures is inevitable. An in-memory cache mechanism was developed that uses so-called hash tables. Hash tables require very little searching through datasets and therefore require little time for locating required information. The cache mechanism has lead to performance improvements of roughly 60 to 100 times for fetching data in comparison with optimized database access.

The CGMS production system is located in Wageningen and operated by Alterra-CGI. To guarantee fast access for JRC end users and to have a fallback infrastructure available, the full dataset is replicated to a second database at JRC in Italy. A customized replication mechanism was developed to facilitate the daily replication of large amounts of data that are modified. This replication mechanism is based on FTP exchange of optimized datasets and is able to replicate all data changes that occur through the operational processes.

Visualization and analysis of results

Results of the CGMS are used by 2 groups of users. The analysts at JRC use the system results for all kinds of analyses, varying from yield prediction and elaboration of bulletins to climate change studies. These users require an advanced set of tools to view and analyze the available data. A second group are the web users, who can access a subset of the available information through an extranet.

For visualization and analysis of results by the JRC analysts several tools have been built through the years. A flexible map viewer application was developed that allows the mapping of spatial data from the system. Web users can access data through a website (http://www.marsop.info).

Currently a large redesign and extension of the instruments for visualization and analysis is performed. This new environment offers advanced features for spatial and temporal analysis through mapping and graphing. It consists of:

- Orientation map: This reference map allows users to analyze an area and select their regions of interest. Users can explore relevant thematic maps and make selections of relevant regions to be further explored in linked maps or graphs.

- Linking of maps and graphs: Maps and graphs of all available indicators can be viewed. The map extent and selected areas of interest (as selected in the orientation map) can be shared between all maps and graphs.
- Visualization of aggregated indicator data: Indicators can be aggregated to various spatial and temporal aggregation levels. Simple examples are a map of the mean temperature over a specific period for all EU countries or a graph of the rainfall sum over the year compared with the long term average.
- Favorites allow analysts to save the settings of their generated maps and graphs. This allows for fast recreation of often used sets of maps and graphs (e.g. as input for the bulletins)
- Legend management allows users to create their own legends for specific maps and to share these legends with other users.

Figure 2 presents weather data for a selected part of Europe through the map viewer.

The chosen IT technologies to set up the map and graph viewers comply with state-of-the art development tools and standards:

Rich Internet Application (RIA)
The new viewers are developed in the RIA environment Adobe Flex. Adobe Flex allows for the development of advanced web enabled user-interfaces that extend the current possibilities of HTML-based environments. This allows for the use of one single development environment for the advanced functionality needed by analysts and the more basic functionality for web users.

Open source web mapping
The generation of maps for the viewer applications is supported by the use of the OS map server GeoServer (www.geoserver.org). Geoserver allows for the seamless integration of raster images and feature based data (e.g. polygons). The use of Styled Layer Descriptors (SLD) facilitates generation of polygon based maps.

Model View Control (MVC) based architecture
The development of complex web enabled user interfaces in an asynchronous, event driven environment like Adobe Flex can potentially lead to complex and even unmanageable event structures. Through the use of an application framework based on the MVC pattern, the handling of events and requests is structured in such a way that manageability and maintainability is guaranteed.