A Framework for Cloud-Based Large-Scale Data Analytics and Visualization: Case Study on Multiscale Climate Data

Sifei Lu*, Reuben Mingguang Li*†, William Chandra Tjhi*,
Kee Khoon Lee*, Long Wang*, Xiarong Li* and Di Ma‡
*Institute of High Performance Computing
Agency for Science, Technology and Research, Singapore
Email: {lus,lirm,tjhiwc,leekk,wangl,lixr}@ihpc.a-star.edu.sg
†Department of Geography
National University of Singapore, Singapore
‡Institute for Infocomm Research
Agency for Science, Technology and Research, Singapore

***Abstract*—In this paper, we present a cloud framework to provide cloud clustering, workflow scheduling and management, fault tolerance and distributed data storage, data analytics and visualisation services. Using a practical case study, we show that in the process of analyzing multiscale climate data, typical problems plaguing data analysts are faced. These include large datasets and limited computational resources; data complexity and limited knowledge; and varying data structures/formats and the need to integrate different tools. The implementation of our framework to climate studies was a success. This can be seen in its ability to perform spatio-temporal data analysis and visualization of a large multi-dimensional climate dataset with reduced processing time. The framework demonstrates great flexibility and simplicity for end users intending to perform data analysis by aiding the integration of data and tools and enabling interactive visualization on-the-fly. This is coupled with effective utilization of computational resources and data storage systems.**

Keywords—component; cloud computing; workflow scheduling; climate data analysis; spatio-temporal visualization; multiscale data

I. Background

Cloud computing is characterized by rapid yet elastic on-demand resource pools, which are suitable for extraction, transformation, processing, integration and analysis of big transaction and interaction data. Data analytics exist in most field of study from computational fault detection to finance to climate studies. Among the biggest challenges facing analysts are:

- large datasets and limited computational resources;
- data complexity and limited knowledge;
- varying data structures and format and the need to integrate different toolsets.

In this paper, we present a cloud-based framework that (i) effectively utilizes cloud computing and data storage resources; (ii) harnesses a workflow management and scheduling engine; and (iii) remotely executes RapidAnalytics [1] services to provide big data processing ability.

Our framework (Fig. 1) is designed to be a countermeasure to the above-mentioned data analytics challenges, aiding algorithm developers, analysts and end users engaged in data analytics and visualization. In particular, on-the-fly visualization allows data to be visualized at intermediate points in the workflow, which is useful for ensuring that the workflow is run correctly before committing to a lengthy process.

To demonstrate the capabilities of this framework, a practical case study was chosen. The case study in question involves large multiscale and multidimensional climate model datasets used in global climate change studies. Both the integration of analytics/visualization tools and the effective implementation of cloud platform will be presented in this paper.

II. Objectives

We not only integrated tools for processing spatio-temporal climate datasets but also contributed to a working cloud platform prototype showcasing the original framework that is being developed. More specifically, our objectives are:

- To provide a cloud platform with rapid elastic computing resource and storage service;
- To contribute to a working prototype with integrated data analytics tools for data extraction, transformation and analytics in a workflow designing platform;
- To develop a workflow engine to orchestrate the cloud clustering service, elastic computing nodes and storage, process execution and monitoring;
- To show the feasibility of the above framework (Fig. 1) to various fields of study through application on a case study that requires spatio-temporal analysis and visualization of a large climate dataset.

978-0-7695-4622-3/11 $26.00 © 2011 IEEE
DOI 10.1109/CloudCom.2011.95

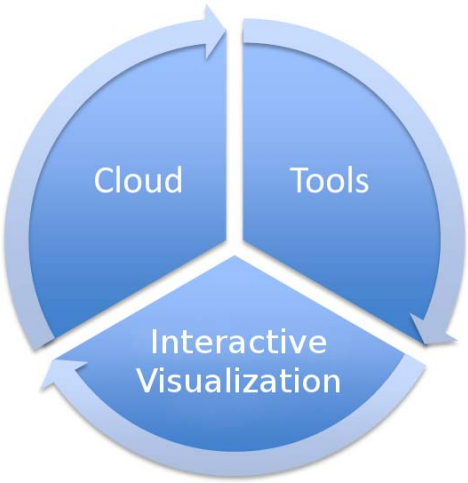

Fig. 1. The framework built on developed and reusable tools on a common cloud computing to handle large resource requirements

III. Case Study on Multiscale Climate Data

The case study is built around the present-day concern of global climate change. Since 1988, the Intergovernmental Panel for Climate Change (IPCC) has been conducting research, with the help of various experts and institutions around the world, on the possible trajectories of climate behaviour in the future. Changes to climate behaviour are of interest due to the possible impacts to humans and the environment. The number of models outputs contributed in Fourth Assessment Report (AR4) alone adds up to around 24 [2].

Each model is run over several possible scenarios (drawn up in the Special Report on Emission Scenarios), consisting of many oceanic and atmospheric output variables. These variables cover a global field at horizontal spatial resolutions as high as 1 decimal degree (360 X 180 grid of the world), spanning up to tens of vertical levels, and temporal resolutions as high as 6-hourly intervals over a period as long as 300 years. For example, the MPI ECHAM5 model spans a grid of 192 X 96 points with 16 levels ($\sim$300,000 data points) for a single time-slice [3]. At daily resolution, a single output variable from the ECHAM5 model, a mathematical model of the global climate developed by Max Planck Institute of Meteorology, will comprise $\sim$300,000 spatial points multiplied by $\sim$365,250 temporal intervals per century per scenario. Extracting subsets from the dataset is a challenge in itself, not to mention its eventual analysis and visualization.

A. Dataset

Multi-dimensional climate projection data (at the global scale) are stored in binary NetCDF files and divided by variables, models and scenarios. In some cases, each variable is split into decadal files. These are obtained the Earth System Grid (ESG) data portal [4].

B. Desired outcome

We aim to obtain useful information at different scales (global, regional, local) from a single dataset using a single workflow on the integrated platform. To cover a range of spatio-temporal taxonomies, we have identified three types of spatio-temporal data to be analysed. They are (i) thematic data, (ii) episodic data, and (iii) trajectory data (Fig. 2).

Thematic spatio-temporal data is similar to a 2-D thematic map except that the thematic data varies not only in space but also in time.

Episodic spatio-temporal data relates to specific events or episodes that occur in a specific spatio-temporal coordinate.

Trajectory spatio-temporal data is associated with certain events or objects manifest in spatial and temporal domains and the spatio-temporal paths taken by these events or objects. We will pay specific attention to sea level pressure as it has considerable influence on atmospheric circulations, e.g. low pressure centres being associated with cyclones and tropical storms (e.g. Fig. 3).

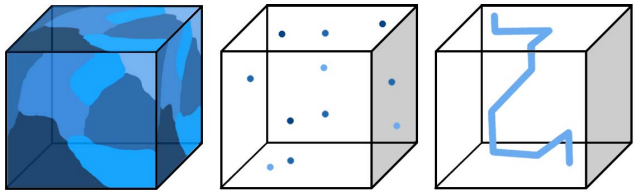

Fig. 2. Visualizing spatio-temporal data in different ways. Left to right: Thematic visualization, episodic visualization and trajectory visualization.

IV. Methodology

The framework is intended to be software-independent, allowing for as many known data structures, classes and toolsets to be integrated as possible without reliance on any specific software.

For the purpose of practical illustration, the framework integrates data analytics tool RapidMiner and RapdAnalytics, Amazon EC2, S3 and Eucalyptus private cloud with cloud clustering service and workflow engine.

A. RapidMiner/ RapidAnalytics

RapidMiner is a popular open source tools with a user friendly GUI for data mining and predictive analysis, the ability of the Java-based standalone software is limited by the memory of the machine and the performance of the cores.

One solution to overcome the limited support for parallelization is to integrate distributed computing framework into RapidMiner for recognition and machine learning application

Fig. 3. The trajectory of Typhoon Vamei, an unprecedented tropical cyclone event due partly to a tropical depression in 2001. Source: Wikipedia

[5]. Radoop is the latest extension for RapidMiner to execute distributed processes using MapReduce algorithms on Hadoop [6]. RapidAnalytics is the server version of RapidMiner; enhanced with secured remote analysis web service, scheduled remote execution and shared repositories.

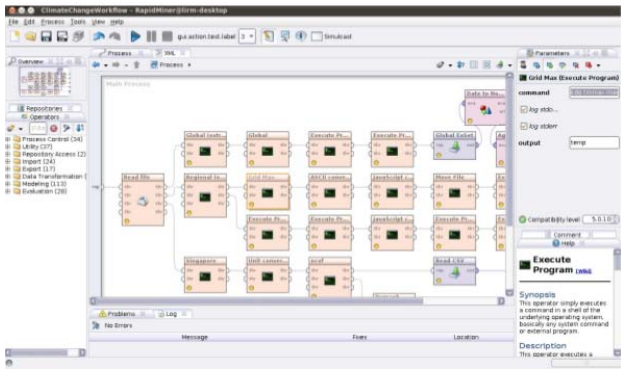

Fig. 4. The RapidMiner GUI and custom functions (linked to the cloud resource) added to the source.

B. Analytics Tools

Custom tools were added to RapidMiner to process spatio-temporal climate data. Parts of the source code were also edited to enable the new formats to be converted into native RapidMiner/RapidAnalytics data types so that Rapidminer functions can be performed on these datasets (see Fig. 4). Some examples of added toolsets include:

- data transformation tools to convert binary files to ASCII or even JavaScript objects (for Jscript-based web visualization tools)
- extraction tools to extract a spatio-temporal subset of the original dataset;
- filtering tools to narrow down data points using conditions
- visualization tools for spatio-temporal data. Some open-source tools integrated include climate data operators, ncview and Protovis.

C. Virtualization

Amazon EC2 enables users to create virtual machines based on the predefined Linux or Windows image with software and configuration data preloaded, and to manage virtual servers and terminate them after execution completed.

The system overview of the data analytics workflow showcase in hybrid cloud is depicted in Fig. 5.

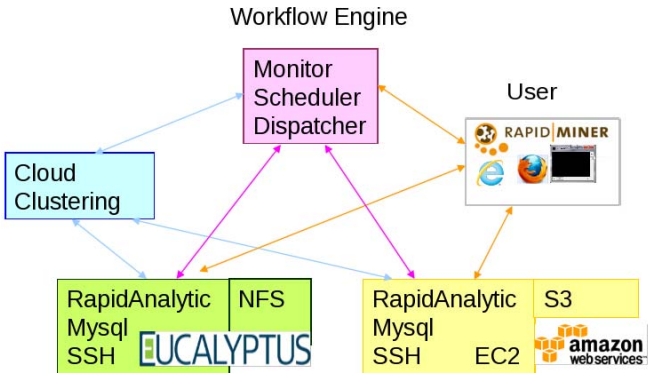

Fig. 5. Schematic showing the architecture of the cloud framework used in this paper.

There are 5 steps to complete a data analytics workflow application.

1) Users use RapidMiner to design the standalone data analytics workflow, and verify the data flow, process logic, visualization result with a small set of data.
2) RapidMiner processes and workflow are reorganised (see Fig. 6) and mapped to RapidAnalytics web service. It can be verified with scheduled batch jobs.
3) The Amazon machine image is built with third-party tools and customized configurations as well as RapidAnalytics web services
4) A similar Eucalyptus virtual machine can be built based on the configuration in last step.
5) The service is published in the workflow engine after configuration of the application for workflow engine is customised.

In our case study, a workflow was created in RapidMiner (including the custom added functions) to obtain episodic, trajectory and thematic spatio-temporal visualizations in the global, regional and local scales of interest from the original dataset.

V. Results

A. Climate Data Analysis and Visualization

1) Visualization (ncview): Fig. 7 (top) shows the graphical plots of thematic, episodic and trajectory data extracted and processed from the original dataset. Interactive heat map

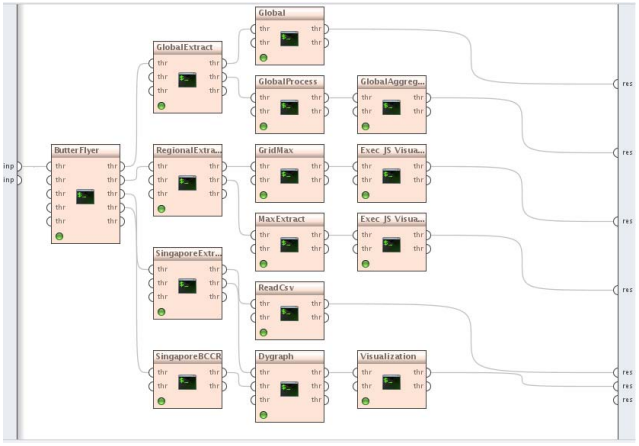

Fig. 6. Reorganised workflow for the purpose of monitoring process chain.

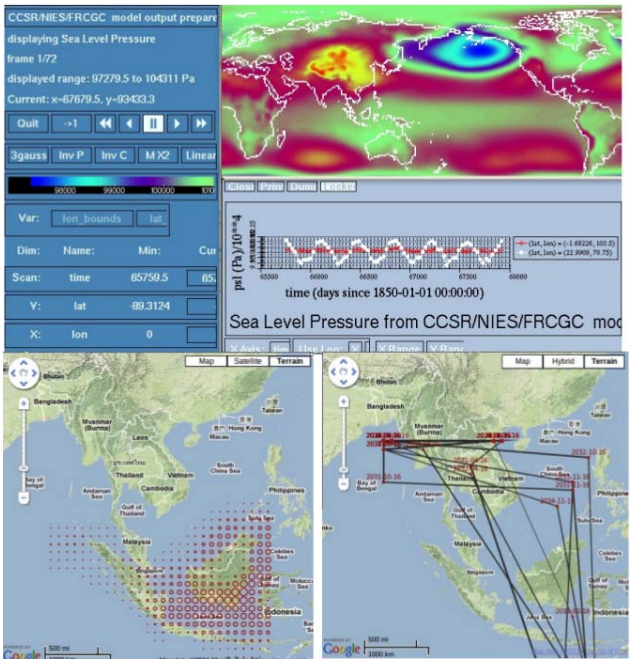

Fig. 7. Top - thematic visualization of SLP data with animation representing time flow. Time series for specific sites (point coordinate) can also be visualized as line graphs. Bottom left - episodic plot of maximum monthly SLP above a certain value. Bottom right - trajectory of minimum monthly SLP

animates the changes to field data across time. Clicking on specific locations on the map will generate time series charts for the set of coordinates. This function can be used to compare the time series of two different locations. Note the enclosed isobars representing pressure cells in the map.

2) Episodic and trajectory visualizations (Protovis) : Subset of binary data is extracted using integrated Climate Data Operators (CDO) tools and then transformed into ASCII and then eventually a JavaScript object in a .js file. Tools allow for a smooth reading of outputs from CDO into the input format required for the web-based Protovis. Information such as high pressure cells and the seasonal shift of low pressure centre trajectories over a selected region can be gathered (see Fig. 7 bottom).

B. Cloud Platform

When an application is triggered from a local RapidMiner instance, the framework relies on cloud clustering service to get available Amazon Elastic Compute Cloud (EC2) instances or Eucalyptus instances based on request Image ID. Then the workflow engine will schedule, monitor and manage the execution of RapidAnalytics web services (refer to Fig. 8 bottom). It supports fault tolerance and the basic workflow scheduling algorithm simplified from the paper by Rahman *et al.* (2011)[7]. The client side RapidMiner will get the processed result from server and invoke few tools to visualise the result.

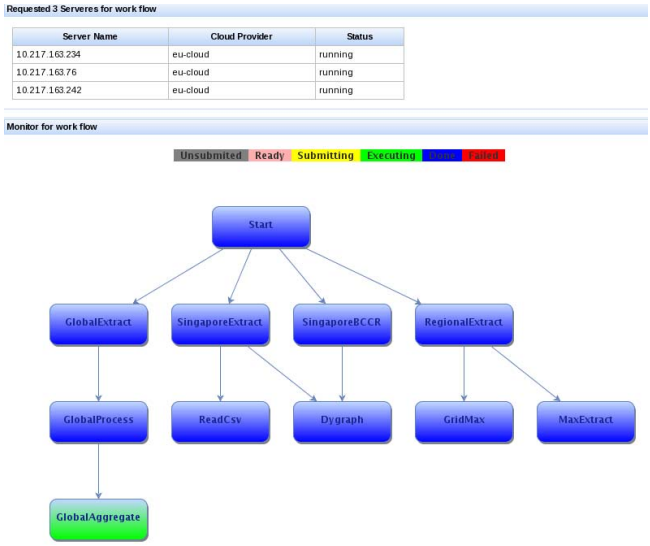

Fig. 8. Workflow execution display showing virtual instances and IP addresses. Boxes represent services; blue=complete, green=in progress. Connectors represent dependency.

Table 1 shows that the speedup ratio is 1.97 and 2.33 for monthly data over 100 years and daily data over 20 years respectively. The reasons for speedup ratio less than 3 (as 3 instances are used) are:

- There exists a special longest path in the workflow; the last task in the path always runs alone;
- The test environment is pure software virtual machines which share Disk I/O resource to transfer big data.

Nevertheless, the increase in speedup from 1.97 to 2.33 when data resolution was increased in our benchmark hints that a large speedup (relative to the number of instances) can be achieved when analyzing data with very fine resolution.

VI. Conclusion

We have developed a cloud-based framework and demonstrated through a practical case study its benefits in

TABLE I

Cloud Processing Time

Application	1 Instance	3 Instances	Speedup
100 years (monthly resolution)	217s	110s	1.97
20 years (daily resolution)	850s	365s	2.33

integrating toolsets with varying operational requirements on a common platform and allowing users to visualize multidimensional datasets of real world phenomena. From the case study, it can be seen that the cloud-based framework is able to provide elastic computing resources and data storage in hybrid cloud. At the core of the framework lies the generic workflow engine which is effective in scheduling, managing analytics workflow and executing tasks in virtual instances.

Such an approach allows for integrative design of analytics and visualization workflows in a distributed computing environments. One possible improvement to the framework would be to auto- generate the workflow configuration XML file for workflow engines. Further works will be focused on the increasing support for data from various fields of study and hence, multi- disciplinary applications. From the cloud perspective, big data storage, retrieving, transferring in the cloud environment is another research direction.

ACKNOWLEDGMENT

The authors would like to thank Dr. Henry Palit for the private cloud environment setup and Dr. Ta Duong for providing the cloud clustering service.

REFERENCES

[1] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*, 2006.

[2] IPCC, Climate Change 2007: The Physical Science Basis. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B.M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

[3] IPCC, Model Summary Data for MPI ECHAM5. Available from: http://www.ipcc-data.org/ar4/model-MPIM-ECHAM5-change.html. Retrieved 5 May 2011.

[4] WCRP CMIP3 Multi-Model Database. Available from: http://esg.llnl.gov:8080/index.jsp.

[5] A. Arimond, C. Koer, and F. Shafait, Distributed Pattern Recognition in RapidMiner, *Proceedings of the RapidMiner Community Meeting and Conference (RCOMM 2010)*, 2010.

[6] Z. Prekopcsk, G. Makrai, T. Henk, C. Gspr-Papanek, Radoop: Analyzing Big Data with RapidMiner and Hadoop , *Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011)*, 2011

[7] M. Rahman, X. Li, H. Palit. Hybrid Heuristic for Scheduling Data Analytics Workflow Applications in Hybrid Cloud Environment, *Proc. High-Performance Grid and Cloud Computing Workshop 2011, in conjunction with International Parallel and Distributed Processing Symposium (IPDPS 2011), Anchorage, Alaska, USA, pp. 961*, 2011.

[8] J. Varia. Architecting Applications for the Amazon Cloud, *Cloud Computing: Principles and Paradigms*, R. Buyya et al. (eds.). Wiley Press, New York, USA, 2010.