A Cloud-Based Workflow Management Solution for Collaborative Analytics

Henry Kasim, Terence Hung, Xiaorong Li, William-Chandra Tjhi, Sifei Lu, and Long Wang

Institute of High Performance Computing,
Agency for Science, Technology and Research (A*STAR), Singapore
{kasimh,terence,lixr,tjhiwc,lus,wangl}@ihpc.a-star.edu.sg

Abstract. The concept of collaborative analytics is to accommodate reuse and collaboration in data analysis process through sharing of analytics methods, algorithms, and computation resources. However, realizing collaborative analytics is challenging due to the large data sets, high throughput and computational intensive requirements. In this demonstration, we present a cloud-based workflow management solution that allows collaborative analytics to run in the cloud computing environment. Our solution provides sharing of analytics resources, recommendation of analytic workflows, dynamic scheduling and provisioning for scalable data analytics, high availability through fault-tolerance, real-time monitoring and tracking of collaborative analytics status. Examples of a generic data mining analysis and climate change analytics are given to show that our work can be applied for a wide variety of study in the real-life world.

1 Introduction

Unraveling useful insights from raw data in a complex domain usually requires interconnecting data collection, preprocessing, analysis and post-processing steps into a sophisticated analytics workflow. In the real world, this workflow design is often done inefficiently through a dynamic and continuous process. To improve this process, analysts today resort to systematic sharing and reuse of analytics methods and resources, resulting in collaborative analytics.

Through collaborative analytics, user is assisted in workflow design as a wide-range of shared analytics workflows designed on collaborative environment are made available for reuse. With collection of workflows to consider, there is also a challenge to run multiple workflows concurrently without have to concern about the limitation of the computing resources. Due to the above challenges, we present a cloud-based workflow management solution for collaborative analytics, as shown in Fig 1.

In Fig 1, Step 1, a variety of data analytics workflows can be saved and shared to *Collaborative Analytics Workflow Database*. Step 2, a user submits his/her data into the collaborative analytics platform and requests for collaborative analytics solutions.

G. Pallis et al. (Eds.): ICSOC 2011, LNCS 7221, pp. 243–245, 2012.

Step 3, the system recommends relevant data analytics solutions based on the user's data. Step 4, a user activate one or multiple solutions for his/her data analytics. Step 5, the *Cloud Workflow Management* performs cloud-based data analytics in a parallel and elastic manner. Step 6, the final analytics results are returned to the user.

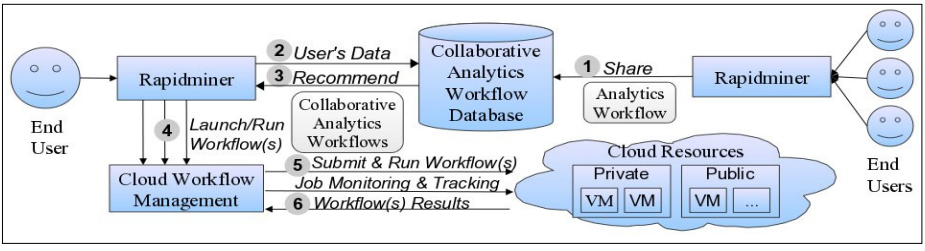

Fig. 1. Cloud-based workflow management solution for collaborative analytics

To handle the collaborative analytics computation resources requirements, we provide a cloud workflow management solution. Cloud workflow management is responsible to schedule and provision the computation resources to execute the recommended analytics workflow efficiently. This solution also support high throughput data processing by running multiple analytics workflows concurrently, high availability through fault-tolerance, and transparency through monitoring and tracking of workflow status. The cloud workflow management encapsulates the complexity of running distributed analytics workflows in a cloud environment.

Our current prototype implementation is built on top of existing analytics workflow suite called RapidMiner [1]. We designed and incorporated new features into RapidMiner to share and recommend analytics workflows based on user's data. Specifically for the recommendation, user's data and the data stored in the collaborative analytics workflow database are characterized based on their statistical properties [2]. The data characteristics of the analytics workflows which are being recommended to the user are similar to the user's data.

2 Demonstration

Our demonstration will showcase two collaborative analytics workflows that we have deployed in our cloud test-bed. The first one is the data mining analysis which extracts statistical pattern from the user data, and the second one is the climate change analysis which analyzes and forecast the weather conditions. These collaborative analytics workflows run in our cloud environment powered by twenty-four cores cluster with hybrid heuristic for scheduling data analytics workflow applications [3].

The data mining analysis demonstrates the sharing of the data mining workflow design, the recommendation of the top three best-suited workflow designs based on the user's data and the workflows execution in the cloud environment. It is a generic use case of collaborative analytics which can be applied to various fields of study.

The climate change analysis demonstrates the uses of collaborative analytics for computational and data-intensive spatio-temporal analysis. It shows the feasibility to deal with spatial and temporal data (i.e. transportation, logistics, geodetics). It supports various data visualisation tools to visualize the analytic results, for example ncview application for thematic data, interactive heat map across time, episodic and trajectory values (i.e., high pressure cells and low pressure centre trajectories).

References

1. Rapidminer – Analytical ETL, Data Mining, and Predictive Reporting, http://rapid-i.com/
2. Castiello, C., Castellano, G., Fanelli, A.M.: Meta-data: Characterization of Input Features for Meta-learning. In: Torra, V., Narukawa, Y., Miyamoto, S. (eds.) MDAI 2005. LNCS (LNAI), vol. 3558, pp. 457–468. Springer, Heidelberg (2005)
3. Rahman, M., Li, X., Palit, H.: Hybrid Heuristic for Scheduling Data Analytics Workflow Applications in Hybrid Cloud Environment. In: Proc. High-Performance Grid and Cloud Computing Workshop 2011, USA, May 16-20 (2011)