International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

SECURE DATA MANAGEMENT AND WORKFLOW MANAGEMENT SYSTEM ON CLOUDS

[1]L.VANAJA, [2]R.SIVAKUMAR

[1]M.TECH DEPT OF CSE, KAKINADA INSTITUTE OF TECHNOLOGICAL SCIENCES, RAMACHANDRAPURAM, ANDHRAPRADESH, INDIA, 533255

[2]ASSISTANT PROFESSOR, KAKINADA INSTITUTE OF TECHNOLOGICAL SCIENCES, RAMACHANDRAPURAM, ANDHRAPRADESH, INDIA, 533255

ABSTRACT

The extraction of useful information from data is often a complex process that can be conveniently modeled as a data analysis workflow. When very large data sets must be analyzed and/or complex data mining algorithms must be executed, data analysis workflows may take very long times to complete their execution. Therefore, efficient systems are required for the scalable execution of data analysis workflows, by exploiting the computing services of the Cloud platforms where data is increasingly being stored. The objective of the paper is to demonstrate how Cloud software technologies can be integrated to implement an effective environment for designing and executing scalable data analysis workflows. We describe the design and implementation of the Data Mining Cloud Framework (DMCF), a data analysis system that integrates a visual workflow language and a parallel runtime with the Software-as-a-Service (SaaS) model. DMCF was designed taking into account the needs of real data mining applications, with the goal of simplifying the development of data mining applications compared to generic workflow management systems that are not specifically designed for this domain. The result is a high-level environment that, through an integrated visual workflow language, minimizes the programming effort, making easier to domain experts the use of common patterns specifically designed for the development and the parallel execution of data mining applications. The DMCF's visual workflow language, system architecture and runtime mechanisms are presented. We also discuss several data mining workflows developed with DMCF and the scalability obtained executing such workflows on a public Cloud.

I. INTRODUCTION

The past two decades have been characterized by an exponential growth of digital data production in many fields of human activities, from science to enterprise. Very large datasets are produced daily from sensors, instruments, mobile devices and computers, and are often stored in distributed repositories. For example, astronomers analyze large image data that every day comes from telescopes and artificial satellites [1]; physicists must study the huge amount of data generated by particle accelerators to understand the laws of Universe [2]; medical doctors and biologists collect huge amount of

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

information about patients to search and try to understand the causes of diseases [3]; sociologists analyze large social networks to find how users are influenced by others for various reasons [4]. Such few examples demonstrate how the exploration and automated analysis of large datasets powered by computing capabilities are fundamental to advance our knowledge in many fields Unfortunately, large datasets are hard to understand, and in particular models and patterns hidden in them cannot be comprehended neither by humans directly, nor by traditional analysis methodologies.

To cope with big data repositories, parallel and distributed data analysis techniques must be used. It is also necessary and helpful to work with data analysis tools and frameworks allowing the effective and efficient access, management and mining of such repositories. In fact, often scientists and professionals use data analysis environments to execute complex simulations, validate models, compare and share results with colleagues located world-wide [6]. Extracting useful information from data is often a complex process that can be conveniently modeled as a data analysis workflow combining distributed datasets, preprocessing tools, data mining algorithms and knowledge models. Workflows provide a declarative way of specifying the high-level logic of an application, hiding the low-level details that are not fundamental for application design. They are also able to integrate existing software modules, datasets, and services in complex compositions implementing discovery processes in scientific and business applications. Cloud systems can be effectively used to handle data analysis workflows since they provide scalable processing and storage services, together with software platforms for developing data analysis environment on top of such services [7].

The objective of the paper is to demonstrate how Cloud software technologies can be integrated to implement an effective programming environment and an efficient runtime system for designing and executing scalable data analysis workflows. Specifically, the paper describes design and implementation of the Data Mining Cloud Framework (DMCF), a system that integrates a visual workflow language and a parallel runtime with the Software-as-a-Service (SaaS) model for enabling the scalable execution of complex data analysis workflows on Clouds. The main contribution of DMCF is the integration of different hardware/ software solutions for high-level programming, management and execution of parallel data mining workflows.

Through its visual programming model, DMCF minimizes the programming effort, making easier to domain experts the use of common patterns specifically designed for the development and the parallel execution of data mining applications. This was done by introducing a visual workflow language, called VL4Cloud, which includes visual patterns useful in real data mining applications, in particular: data preprocessing (data partitioning and

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

filtering); parameter sweeping (the concurrent execution of many instances of the same tool with different parameters to find the best result); input sweeping (the concurrent execution of many instances of the same tool with different input data); tool sweeping (the concurrent execution of different tools on same data); combinations of parameter, input, and tool sweeping patterns for the highest flexibility; data and models aggregation (e.g., models evaluations, voting operations, models aggregation). For supporting these patterns, VL4Cloud provides novel data-mining-specific visual workflow formalisms, data and tool arrays, which significantly ease the design of parallel data analysis workflows. A data array allows representing an ordered collection of input/output data sources in a single workflow node, while a tool array represents multiple instances of the same tool. Thanks to data and tool arrays, workflows are more compact compared to those designed using other visual formalisms that oblige developers to replicate node chains to obtain the same semantic. The DMCF runtime was designed to enable the parallel execution of data analysis workflows on multiple Cloud machines, so as to improve performance and ensure scalability of applications. To this end, the runtime implements data-driven task parallelism that automatically spawns ready-to-run workflow tasks to the Cloud resources, taking into account dependencies among tasks and current availability of data to be processed.

Parallelism is effectively supported by the data and tool array formalisms of VL4Cloud, because the array cardinality automatically determines the parallelism degree at runtime. In addition, data and tool arrays improve generality of workflows and therefore their reusability. In fact, once defined, a workflow can be instantiated many times not only changing the input data or the tools (as in the other workflow formalisms), but also redefining the parallelism level by specifying a different cardinality of the data/tool arrays.

Finally, DMCF is provided according with the SaaS model. This means that no installation is required on the user's machine: the DMCF visual user interface works in any modern Web browser, and so it can be run from most devices, including desktop PCs, laptops, and tablets. This is a key feature for users who need ubiquitous and seamless access to scalable data analysis services, without needing to cope with installation and system management issues. Thanks to the SaaS approach, user have online access to a large repository of ready-to-use data handling and mining algorithms. Many of such algorithms are taken from open source projects (more than 100 algorithms from the Weka and Waffles libraries) and several of them were designed by scratch (e.g., tools for data splitting, data merging, voting). In addition, it is easy for every user to add his/her own algorithm in the system using a visual configuration tool. Through a guided procedure, the configuration tool allows users to upload

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

executable files of the new algorithm, and to specify its input and output parameters.

II. EXISTING SYSTEM

Galaxy [13] is a web-based platform for developing genomic science applications, now used as a general bioinformatics workflow management system. A Galaxy workflow is a reusable template that a user can run repeatedly on different data. The Galaxy [13] software runs on Linux/Unix based servers, and therefore several organizations execute Galaxy on private or public Cloud IaaS. In order to improve Galaxy's capabilities with respect to interfacing with large scale computational systems and running workflows in a parallel manner, Galaxy has recently been integrated with Swift/T [14], a large-scale parallel programming framework discussed below.

Taverna [15] is a workflow management system mostly used in the life sciences community. Taverna can orchestrate Web Services and these may be running in the Cloud, but this is transparent for Taverna, as demonstrated in the BioVel project. Recently, the Tavaxy system has been developed for allowing the integration of Taverna and Galaxy [16]. In particular, Tavaxy allows users to create and execute workflows employing an extensible set of workflow patterns that enables the re-use and integration of existing workflows from Taverna and Galaxy, and allows the creation of hybrid workflows.

Orange4WS [17] is a service-oriented workflow system that extends Orange, a data mining toolbox and a visual programming environment for the visual composition of data mining workflows. The system enables the orchestration of web-based data mining services and the collection of information in various formats, as well as design of repeatable data mining workflows used in bioinformatics and e-science applications.

Kepler [18] is a visual workflow management system that provides a graphical user interface for designing scientific workflows. Data is encapsulated in messages or tokens, and transferred between tasks through input and output ports. Kepler provides an assortment of built-in components with a major focus on statistical analysis and supports task parallel execution of workflows using multiple threads on a single machine.

E-Science Central (e-SC) [19] allows scientists to store, analyze and share data in the Cloud. Its inbrowser workflow editor allows users to design a workflow by connecting services, either uploaded by themselves or shared by other users of the system. One of the most common use cases for e-Sc is to provide a data analysis back end to a standalone desktop or Web application. In the current implementation, all the workflow services within a single invocation of a workflow execute on the same Cloud node.

III. PROPOSED SYSTEM

The proposed system is to demonstrate how Cloud software technologies can be integrated to implement an effective programming environment and an efficient runtime system for designing and executing

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

scalable data analysis workflows. Specifically, the paper describes design and implementation of the Data Mining Cloud Framework (DMCF), a system that integrates a visual workflow language and a parallel runtime with the Software-as-a-Service (SaaS) model for enabling the scalable execution of complex data analysis workflows on Clouds. The main contribution of DMCF is the integration of different hardware/ software solutions for high-level programming, management and execution of parallel data mining workflows.

Through its visual programming model, DMCF minimizes the programming effort, making easier to domain experts the use of common patterns specifically designed for the development and the parallel execution of data mining applications. This was done by introducing a visual workflow language, called VL4Cloud, which includes visual patterns useful in real data mining applications, in particular: data preprocessing (data partitioning and filtering); parameter sweeping (the concurrent execution of many instances of the same tool with different parameters to find the best result); input sweeping (the concurrent execution of many instances of the same tool with different input data); tool sweeping (the concurrent execution of different tools on same data); combinations of parameter, input, and tool sweeping patterns for the highest flexibility; data and models aggregation (e.g., models evaluations, voting operations, models aggregation). For supporting these patterns, VL4Cloud provides novel data-mining-specific visual workflow formalisms, data and tool arrays, which significantly ease the design of parallel data analysis workflows.

IV. MODULE DESCRIPTION:

4.1 User

4.1.1 User Registration:

A registered **user** is a **user** of a website, program, or other system who has previously registered.
Registered **users** normally provide some sort of credentials (such as a username or e-mail address, and a password) to the system in order to prove their identity: this is known as logging in

4.1.2 Send Request

User can send the request for work schedule to the cloud service provider. Download Work schedule Cloud service providers send the request to the User for downloading the work schedule.

4.2 Cloud Service Provider:

4.2.1 Workload

Cloud service provider can load the amount of work.

4.2.2 Work schedule

Work is to be assigned for the user.

4.2.3 Authentication

User can authenticate for the available request.

4.2.4 Send work

After authenticated the user, CSP can send the work to the User.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

V. CONCLUSION

Cloud systems can be used as scalable infrastructures to support high-performance platforms for data analysis applications. Based on this vision, we designed DMCF for large-scale data analysis on the Cloud. The main contribution of DMCF is the integration of different hardware/software solutions for high-level programming, management and execution of parallel data mining workflows.

We evaluated the performance of DMCF through the execution of workflow-based data analysis applications on a pool of virtual servers hosted by a Microsoft Cloud data center. The experimental results demonstrated the effectiveness of the framework, as well as the scalability that can be achieved through the execution of data analysis applications on the Cloud. Besides performance considerations, we point out that the main goal of DMCF is providing an easy-to use SaaS interface to reliable data mining algorithms, thus enabling end-users to focus on their data analysis applications without worrying about low level computing and storage details, which are transparently managed by the system

VI. REFERENCES

[1] A. Burd et al., "Pi of the sky-all-sky, real-time search for fast optical transients," New Astronomy, vol. 10, no. 5, pp. 409 – 416, 2005.

[2] O. Rubel, C. Geddes, M. Chen, E. Cormier-Michel, and E. Bethel, "Feature-based analysis of plasma-based particle acceleration data," Visualization and Computer Graphics, IEEE Transactions on, vol. 20, no. 2, pp. 196–210, 2014.

[3] T. Tucker, M. Marra, and J. Friedman, "Massively parallel sequencing: The next big thing in genetic medicine," The American Journal of Human Genetics, vol. 85, no. 2, pp. 142–154,
2009.

[4] The SAGE Handbook of Social Network Analysis, 0th ed. SAGE Publications Ltd, 2014.

[5] T. Hey, S. Tansley, and K. Tolle, Eds., The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

[6] D. Talia and P. Trunfio, "How distributed data mining tasks can thrive as knowledge services," Commun. ACM, vol. 53, no. 7, pp. 132–137, 2010.

[7] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good, "On the use of cloud computing for scientific workflows," in eScience, 2008, 2008, pp. 640–645.

[8] G. Agapito, M. Cannataro, P. H. Guzzi, F. Marozzo, D. Talia, and P. Trunfio, "Cloud4snp: Distributed analysis of snp microarray data on the cloud," in Proc. of the ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics 2013 (ACM BCB 2013), Washington, DC, USA, 2013, p. 468.

[9] A. Altomare, E. Cesario, C. Comito, F. Marozzo, and D. Talia, "Trajectory pattern mining over a cloud-based framework for urban computing," in Proc. of the 16th Int. Conference on High Performance

Computing and Communications (HPCC 2014), Paris, France, 2014, pp. 367–374.
[10] D. Talia, "Workflow systems for science: Concepts and tools," ISRN Software Engineering, 2013.
[11] M. Bux and U. Leser, "Parallelization in scientific workflow management systems," CoRR, 2013.
[12] J. Yu and R. Buyya, "A taxonomy of scientific workflow systems for grid computing," SIGMOD Rec., vol. 34, no. 3, pp. 44–49, 2005.
[13] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," Genome Biology, vol. 11, no. 8, p. R86, 2010.
[14] K. Maheshwari, A. Rodriguez, D. Kelly, R. Madduri, J. Wozniak, M. Wilde, and I. Foster, "Enabling multi-task computation on galaxy-based gateways using swift," in IEEE Int. Conference on Cluster Computing, 2013, pp. 1–3.
[15] K. Wolstencroft et al., "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud," Nucleic Acids Research, vol. 41, pp. 557–561, 2013.
[16] M. Abouelhoda, S. Issa, and M. Ghanem, "Tavaxy: Integrating taverna and galaxy workflows with cloud computing support," BMC Bioinformatics, vol. 13, no. 1, 2012.
[17] V. Podpeˇcan, M. Zemenova, and N. Lavraˇc, "Orange4ws environment for service-oriented data mining," Comput. J., vol. 55, no. 1, pp. 82–98, 2012.
[18] B. Ludscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the kepler system," Concurr. Comput.: Pract.
Exper., vol. 18, no. 10, pp. 1039–1065, 2006.
[19] H. Hiden, S. Woodman, P. Watson, and J. Cala, "Developing cloud applications using the e-Science Central platform," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 371, no. 1983, 2013. [20] J. Kranjc, V. Podpeˇcan, and N. Lavraˇc, "ClowdFlows: A Cloud Based Scientific Workflow Platform," in Machine Learning and Knowledge Discovery in Databases, ser. Lecture Notes in Computer Science. Springer, 2012, vol. 7524, pp. 816–819.
[21] E. Deelman et al., "Pegasus: A framework for mapping complex scientific workflows onto distributed systems," Scientific Programming, vol. 13, no. 3, pp. 219–237, 2005.