A Review on Workflow Management System for Scalable Data Mining on Clouds

Ranjita L. Ram[1] Prof. Pravin G.kulurkar[2] Prof. Pranita Laddhad[3]
[1]M. Tech. Student [2]Head of Dept. [3]Assistant Professor
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Vidarbha Institute of Technology, Nagpur, India

Abstract— Cloud computing provides flexible services, large performance and scalable data storage to a large and everyday increasing number of users. Clouds enlarged the offer of distributed computing systems by providing advanced Internet services that complement and complete functionalities of distributed computing provided by the Web, Grid computing and peer-to-peer networks. Cloud computing systems provide large-scale infrastructures for complex high-performance applications. Although a few cloud-based analytics platforms are available to-day. As more such platforms emerge, researchers will port increasingly powerful data mining programming tools and strategies to the cloud to exploit complex and flexible software models such as the distributed workflow paradigm. In this paper, we use micro strategic technique and implement a Data Mining Cloud Framework (DMCF) with different application.

Key words: Workflows, Data Analysis, Cloud Computing, Software-As-A-Service, Scalability

I. INTRODUCTION

Cloud computing provides elastic services, high performance and scalable data storage to a large and everyday increasing number of users. Clouds enlarged the offer of distributed computing systems by providing advanced Internet services that complement and complete functionalities of distributed computing provided by the Web, Grid computing and peer-to-peer networks. In fact, Cloud computing systems provide large-scale infrastructures for complex high-performance applications. Most of those applications use big data repositories and needs to access and analyse them to extract useful information. Big data is a new and over-used term that refers to massive, heterogeneous, and often unstructured digital content that is difficult to process using traditional data management tools and techniques. The term includes the complexity and variety of data and data types, real-time data collection and processing needs, and the value that can be obtained by smart analytics. Advanced data mining techniques and associated tools can help extract information from large, complex datasets that are useful in making informed decisions in many business and scientific applications including advertising, market sales, social studies, bioinformatics, and high-energy physics. Combining big data analytics and knowledge discovery techniques with scalable computing systems will produce new insights in a shorter time. Although a few cloud-based analytics platforms are available to-day, current research work foresees that they will become common within a few years. Some current solutions are open source systems such as Apache Hadoop and SciDB, while others are proprietary solutions provided by companies such as Google, IBM, EMC, BigML, Splunk Storm, Kognitio, and Insights One. As more such platforms emerge, researchers will port increasingly powerful data mining pro-gramming tools and strategies to the cloud to exploit complex and flexible software models such as the distributed workflow paradigm. The growing use of service-oriented computing could accelerate this trend. Developers and researchers can adopt the software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) models to implement big data analytics solutions in the cloud. In such a way, data mining tasks and knowledge discovery applications can be offered as high-level services on Clouds. This approach creates a new way to delivery data analysis software that is called data analytics as a service (DAaaS). Here we describe a Data Mining Cloud Framework (DMCF) that we developed according to this approach. In DMCF, data analysis workflows can be designed through visual programming, which is a very effective design approach for high-level users, e.g. domain-expert analysts having a limited understanding of programming. Recently, we extended the DMCF system to support also script-based data analysis workflows, as an additional and more flexible program-ming interface for skilled users. To this end, in [4] we introduced a workflow-oriented language, called JS4Cloud, to support the design and execution of script-based data analysis workflows on DMCF.

II. RELATED WORK

A. Saif Ur Rehman Malik, Samee U. Khan, Sam J. Ewen,"Performance analysis of data intensive cloud systems based on data management and replication: a survey"

Author in this paper, have studied Data Replication and Management, two instrumental technologies that are widely used to manage massive quantities of data on cloud services. The techniques are compared and analyzed based upon the abovementioned features. We also analyze the working of numerous data replication techniques and how data-intensive applications are deployed in the cloud. The knowledge provided in the paper can be further exploited to design and model new mechanisms or approaches in the cloud.

B. Astha Pareek, Manish Gupta"Review of Data Mining Techniques in Cloud Computing Database"Author in this paper use data mining technique

Author in this paper use Data mining is the extraction of hidden information from the huge volume of data. The current business world is utilizing the data mining for gaining the insight into business strategies.in this paper, we focused the implementation of K-Means algorithm in the

Cloud environment and the experimental results shows that it works well in the Cloud.

C. F. Marozzo, D. Talia, and P. Trunfio," A cloud framework for big data analytics workflows on azure"

Author in this paper,use data sets, analysis tools, data mining algorithms and knowledge models that are implemented as single services that can be combined through a visual programming interface in distributed workflows to be executed on Clouds. The first implementation of the Data Mining Cloud Framework on Azure is presented and the main features of the graphical programming interface are described workflow programming features of the framework, we discuss the Data Mining Cloud Framework designed for developing and executing distributed data analytics applications as workflows of services.

D. Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio,"Scalable script-based data analysis workflows on clouds"

Author in this paper, use the Data Mining Cloud Framework (DMCF) which is a software system for designing and executing data analysis and knowledge discovery workflows on the Cloud. In this paper we described our solution for programming and executing parallel script-based data analysis workflows in DMCF. We introduced a workflow language, named JS4Cloud that extends JavaScript to support the development of Cloud-based data analysis tasks and the access to data on the Cloud.

III. Problem Statement

Due to the crucial role that workflow applications play in the scientific community, most current WMSs were developed to enable the execution of these applications in grid computing platforms. When Clouds became main stream, WMSs were enhanced to support it. In this section, we present a brief description of the most prominent WMS found in the state-of-the-art which are related with our work. Pegasus is a mature Workflow Management System that combines features such as portability across a wide range of infrastructures, scalability, data management or transformations. It can be used with popular programming languages among the scientific community through its APIs (application programming interfaces) and also supports submission via web portals. Although it supports multiple cloud providers, it does not dynamic provision resources with different hardware and software requirements.

The advent of Cloud Computing and its core characteristics (rapid elasticity, resource pooling, and pay-per-use, among others) are well-suited to the nature of scientific applications that experience a variable demand during execution. As a consequence, many WMSs derived from projects in the area of grid computing were updated to support the execution on Cloud resources. However, many of their features are optimized for grids and thus are unable to obtain the most key aspects of clouds, such as dynamic provisioning of resources.

For that reason, this work presents a novel multi-platform (clusters and clouds) WMS with support for on-demand provisioning of multi-cloud (public, private and hybrid) customized cloud computing resources. The tool developed in this work has been tested using a comparative genomics pipeline, called search. Promising results were obtained, with significant speedup ratio com-pared to the batch-oriented pipeline. The current working lines include adding support for Grid computing, using a more efficient transference protocol than SSH and implementing data privacy.

IV. Conclusions

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. we point out that the main goal of DMCF is providing an easy-touse 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.

References

[1] Saif Ur Rehman Malik ,Samee U. Khan ,Sam J. Ewen," Performance analysis of data intensive cloud systems based on data management and replication" Distributed Parallel Databases DOI 10.1007/s10619-015-7173-2Springer Science+Business Media New York 2015.

[2] Astha Pareek, Manish Gupta, "Review of Data Mining Techniques in Cloud Computing Database" International Journal of Advanced Computer Research, Volume-2 Number-2 Issue-4 June-2012

[3] F.Marozzo, D.Talia, P.Trunfio," A cloud framework for big data analytics workflows on azure"In Proc. of the 2012 High Performance Computing Workshop, HPC 2012. 2012

[4] Fabrizio Marozzo,, Domenico Talia, and Paolo Trunfio" Scalable script-based data analysis workflows on clouds" In Proc. of the 8th Workshop on Workflows in Support of Large-Scale Science (WORKS2013), pages 124–133, Denver, CO, USA, November 2013. ACM Press

