International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

A WORKFLOW MANAGEMENT SYSTEM FOR SCALABLE DATA MINING ON CLOUDS

Supreetha S[1] ,Pooja H.R[2], Roshan s olaty[3], Bhumika B.R[4]

[1]*Assistant Professor, Dept. of Computer Science & Engineering, Sapthagiri college of Engineering*
[2,3,4]*Department of Computer Science Engineering, Sapthagiri College Of Engineering, Banglore-57, Karnataka, India*

Abstract-*Data generation is increasing every day and the amount of data generated is enormous so that extraction of useful information is extremely increasing. The world is moving to a on-command on-demand world that means information is needed when and where it is required. The extraction of useful information from data is often complex process that can be conveniently modeled as a data analysis workflow. When large data sets has been provided it becomes a challenge for the present data mining algorithms to analize and retrieve information. Data analysis workflows may take long time for execution therefore efficient systems are required for scalable execution. Nowadays since all the data is being stored in the cloud as they offer storage and processing services that are scalable in addition to a software platform for its manageability. Workflow for analysing data along with estimated datasets and data mining algorithms are used for eliciting valued information from data. Here various management systems for visual workflow which are suitable for cloud are discussed along with DMCF visual workflow language which is scalable in obtaining and executing such workflows on a public and private cloud.*

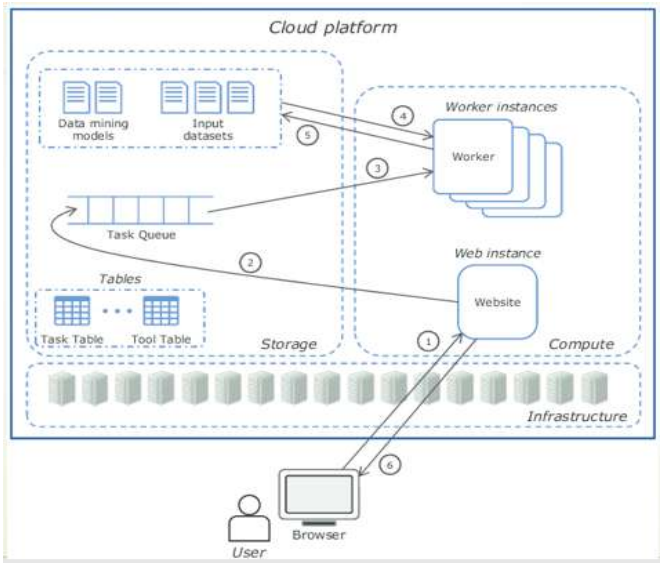
***KeyWords*:CloudComputing,DataMining,Workflow,DMCF**

1.INTRODUCTION

In order to manage such a complicated task of storing the data cloud is required providing computing resources as a SaaS (software as a service).Combination of data mining in cloud is recent trend in knowledge discovery and pattern matching because no large number of results are effectively accomplished and accessed by the cloud clients. The objective here is to design cloud software technologies and integrate it to implement an effective environment for designing and executing scalable data analysis workflow.

Data analysis workflows may take long time for execution therefore efficient systems are required for scalable execution. Nowadays since all the data is being stored in the cloud as they offer storage and processing services that are scalable in addition to a software platform for its manageability. Workflow for analysing data along with estimated datasets and data mining algorithms are used for eliciting valued information from data. In the era of big data, a huge amount of data can be generated quickly from various sources (e.g., smart phones, sensors, machines, social networks, etc.). Toward these big data, conventional computer systems are not competent to store and process these data. Due to the flexible and elastic computing resources, cloud computing is a natural fit for storing for execution of processing pipelines among heterogeneous event processing engines as a workflow.

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 analysed 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.



2. Related works

In ClowdFlows: A Cloud Based Scientific Workflow Platform

The paper presents a simple cloud based platform for com-position, execution, and sharing of interactive data mining workflows. It depends on the principles of service-oriented knowledge discovery, and features interactive scientific workflows. In contrast to comparable data mining platforms, our platform runs in all major Web browsers and platforms, including mobile devices. In terms of cstoring, ClowdFlows provides researchers with an easy way to expose and share their work and results, as only an Internet connection and a Web browser are required to access the workflows from anywhere. Provides cross-platform functionality. Appropriate for beginners, data miners, and non-experts due to intuitive and basic user interface. Mining uninterrupted data streams from the internet not supported.

Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences

Reliability on computational approaches in the sciences has revealed to concerns about how accessible and reproducible computation results truly are. Galaxy http://usegalaxy.org, an open web-based plat- form for genomic research, addresses these problems. Galaxy automatically tracks and manages data provenance and provides support for capturing the context and intent of computational methods. Galaxy Pages are interactive than web-based documents that provide users with a medium to communicate with a complete computational analysis. Management of data analysis and computational operations simplified. Lesser abilities regarding the interfacing with large-scale computational systems and workflow in a parallel and distributed manner.

Scientific Workflow Management and the Kepler System

Many scientific disciplines are now data and information driven, and new scientific knowledge is often gained by scientists putting together data analysis and knowledge discovery "pipelines". A related trend is that more and more scientific communities realize the benefits of sharing their data and computational services, and are thus contributing to a distributed data and computational community infrastructure. However, this infrastructure is only a means to an end and scientists ideally should be bothered little with its existence. The goal is for scientists to focus on development and use of what we call scientific workflows. These are networks of analytical steps that may involve, e.g., database access and querying steps, data analysis and mining steps, and many other steps including computationally intensive jobs on high performance cluster computers. In this paper we describe characteristics of and requirements for scientific workflows as identified in a number of our application projects. We then elaborate on Kepler, a particular scientific work- flow system, currently under development across a number of scientific data management projects. We describe some key features of Kepler and its underlying Ptolemyii system, planned extensions, and areas of future research. Kepler is a community- driven, open source project, and we always welcome related projects and new contributors to join.

Orange4WS Environment for Service-Oriented Data Mining

Novel data-mining tasks in e-science involve mining of distributed, highly heterogeneous data and knowledge sources. However, standard data mining platforms, such as Weka and Orange, involve only their own data mining algorithms in the process of knowledge discovery from local data sources. In contrast, next generation data mining technologies should enable processing of distributed data sources, the use of data mining algorithms implemented as web services, as well as the use of formal descriptions of data sources and knowledge discovery tools in the form of ontologies, enabling automated composition of complex knowledge discovery workflows for a given data mining task. This paper proposes a novel Service-oriented Knowledge Discovery framework and its implementation in a service-oriented data mining environment Orange4WS (Orange for Web Services), based on the existing Orange data mining toolbox and its visual programming environment, which enables manual composition of data mining workflows. The new service-oriented data mining environment Orange4WS includes the following new features: simple use of web services as remote components that can be included into a data mining workflow; simple incorporation of relational data mining algorithms; acknowledge discovery ontology to

describe workflow components (data, knowledge and data mining services) in an abstract and machine-interpretable way, and its use by a planner that enables automated composition of data mining workflows. These new features are show cased in three real-world scenarios.

3. EXISTING SYSTEM

Existing cloud computing system offers not only reliable services with performance guarantees, but also savings on in-house IT infrastructures. However, the datasets that is used for clustering purpose may contain important information, e.g., patient health information, commercial data, and behavioral data, etc, directly outsourcing them to public cloud servers inevitably raises privacy concerns.

Disadvantages:

- . Less Efficiency
- Security is less
- Speed of Transmission is low

4. Overview of the Present Work

The rapid growth of big data involved in today's data mining and analysis also introduces challenges for clustering over them in terms of volume, variety, and velocity. To efficiently manage large-scale datasets and support clustering over them, public cloud infrastructure is acting the major role for both performance and economic consideration. Nevertheless, using public cloud services inevitably introduces privacy concerns. This is because not only many data involved in data mining applications are sensitive by nature, such as personal health information, localization data, financial data, etc, but also the public cloud is an open environment operated by external third-parties. For example, a promising trend for predicting an individual's disease risk is clustering over existing patients health records , which contain sensitive patient information. Therefore, appropriate privacy protection mechanisms must be placed when outsourcing sensitive datasets to the public cloud for clustering.

OBJECTIVES

Data mining:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It provides the tools and techniques to identify valid, potentially useful, and ultimately understandable data. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. It is the practice of automatically searching large storage of data to discover patterns and trends that go beyond simple analysis. The goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. Traditional data mining tasks such as association rule mining, market analysis and cluster analysis commonly attempt to find patterns in a dataset characterised by a collection of independent instances of a single relation.

5. PROPOSED WORK

We are proposing a system K-means clustering over Large-scale Dataset using Map Reduce technique. First we are initializing trained data set for every different cluster which is related to medical Information. After, the clustering algorithm divide file into number of chunks and for every chucks hash code is generated for the security purpose. Before storing into Cloud System, classification algorithm classifies that file belong to which cluster category.

Advantages of the Proposed System

- More Efficiency
- Security is more because of Hash code generation
- Speed of Transmission is high because of deduplication concept is used while uploading file to the Cloud storage

SEQUENCE MODULE

K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function

Hadoop Map Reduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a sub-project of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.

According to The Apache Software Foundation, the primary objective of Map/Reduce is to split the input data set into independent chunks that are processed in a completely parallel manner. The Hadoop Map Reduce framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file system.

A gateway is a hardware device that acts as a "gate" between two networks. It may be a router, firewall, server, or other device that enables traffic to flow in and out of the network.

While a gateway protects the nodes within network, it also a node itself. The gateway node is considered to be on the "edge" of the network as all data must flow through it before coming in or going out of the network. It may also translate data received from outside networks into a format or protocol recognized by devices within the internal network.

A router is a common type of gateway used in home networks. It allows computers within the local network to send and receive data over the Internet. A firewall is a more advanced type of gateway, which filters inbound and outbound traffic, disallowing incoming data from suspicious or unauthorized sources. A proxy server is another type of gateway that uses a combination of hardware and software to filter traffic between two networks. For example, a proxy server may only allow local computers to access a list of authorized websites.

DNA computing is a branch of computing which uses DNA, biochemistry, and molecular, biology hardware, instead of the traditional silicon-based computer technologies. Research and development in this area concerns theory, experiments, and applications of DNA computing. The term "molectronics" has sometimes been used, but this term has already been used for an earlier technology, a then-unsuccessful rival of the first integrated circuits; this term has also been used more generally, for molecular-scale electronic technology

6. SYSTEM IMPLEMENTATION

User Module

- Registration
  In Initialization module , new user register their details in the process and get the username and password

- Login
  User enter the username and password

- View Profile (Edit)
  User can edit his/her details using this module.

- Training Process
  Training data will be uploaded for training process.

- Upload File
  User enter the username and password for uploading, User upload the file into server.

Avoid Redundancy

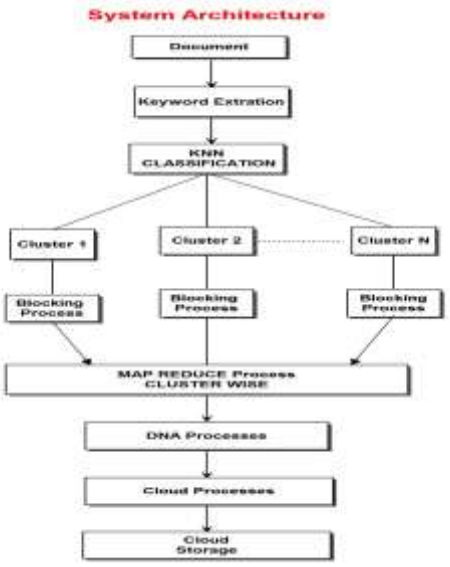
- Here we are using redundancy methods to be implemented for avoids repeated data for encrypted data.

- Then the file ready to load for the cloud resources. The access policy is in plaintext, which may leak some private information about the end-users.

- Based on our observation, the attributes are leaked from the attribute mapping function.

- Download File
  For user Downloading, we verify the user credentials and file will be downloaded .allow the request user to access the file and download to him/her
- Change Password.
  It is used to change user password.

- Logout
  User can logout using this module.



TESTING

Definition

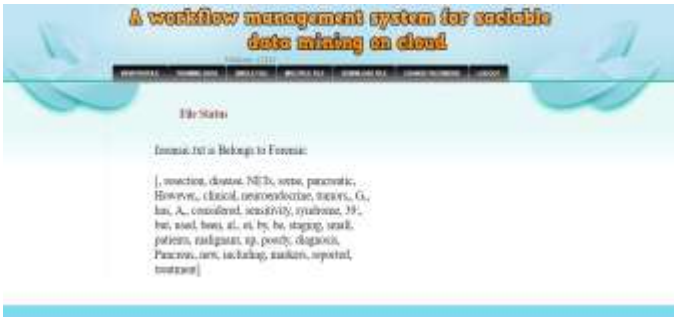
Unit testing is a development procedure where programmers create tests as they develop software. The tests are simple short tests that test functionally of a particular unit or module of their code, such as a class or function.

Using open source libraries like cunit, oppunit and nun it (for C, C++ and C#) these tests can be automatically run and any problems found quickly. As the tests are developed in parallel with the source unit test demonstrates its correctness.

7. SCREENSHOTS

User registration

Uploading file



File status display

8. CONCLUSION

In the existing system large data sets 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 worldwide,making easier to domain experts the use of common patterns specifically designed for parallel execution of data mining applications. Performance and accuracy is more comparing to other applications and time complexity is reduced.

9. FUTURE ENHANCEMENT

In our future enhancement we are outsourcing the data, security the main concern because there may be a malicious hacker can hack our data. Speed of transmission is very low when we are dumping the files into big data. So we need to concentrate about the speed of transmission. Less optimized data is also a main concern and also concentrating on various kinds of data from multiple sources such as image video etc

- **Data Classification based on Security:** A cloud computing data center can store data from various users. To provide the level of security based on the importance of data, classification of data can be done. This classification scheme should consider various aspects like access frequency, update frequency and access by various entities etc. based on the type of data. Once the data is classified and tagged, then level of security associated with this specific tagged data element can be applied. Level of security includes confidentiality, encryption, integrity and storage etc. that are selected based on the type of data.
- **Identity management system:** Cloud computing users are identified and used their identities for accessing the services. A secure trust based identity management scheme is essentially a need by all cloud service provider and users. Various issues of identity management system are identified. Solution to secure id-generation and distribution, storage and life cycle management is a demand for trust based identity management system.

- **Secure trust based Solution for cloud computing Service:** A secure environment for execution of the cloud computing services along with overall security considerations is a challenge. A secure and trusted solution is the requirement that needs to be focused and addressed by the cloud computing infrastructure.

- **Optimization of resource Utilization:** Security considerations and provisions for virtualization along with the optimum use of the cloud infrastructure also needs to be focused and addressed.

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," IEEE Trans. Vis. Comput. Graph., vol. 20, no. 2, pp. 196–210, Feb. 2014.

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

[4] The SAGE Handbook of Social Network Analysis. Newbury Park, CA, USA: Sage, 2014.

[5] T. Hey, S. Tansley, and K. Tolle, Eds., The Fourth Paradigm: Data Intensive Scientific Discovery. Redmond, WA, USA: Microsoft Res., 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, et al., "On the use of cloud computing for scientific workflows," in Proc. IEEE 4th Int. Conf. eScience, 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. Int. Conf. Bioinf. Comput. Biol. Biomed. Informat., 2013, Art. no. 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. 16th Int. Conf. High Perform. Comput. Commun., 2014, pp. 367–374.

[10] D. Talia, "Workflow systems for science: Concepts and tools," ISRN Softw. Eng., vol. 2013, 2013, Art. no. 404525.

BIOGRAPHIES

SUPREETHA S
Assistant Professor
Computer science and engineering department
Sapthagiri college of engineering

POOJA H R
1SG15CS068
Computer science and engineering department
Sapthagiri college of engineering

ROSHAN S OLATY
1SG15CS090
Computer science and engineering department
Sapthagiri college of engineering

BHUMIKA BR
1SG15CS017
Computer science and engineering department
Sapthagiri college of engineering