2013 International Conference on Advanced Cloud and Big Data

Cloud-based Big Data Mining & Analyzing Services Platform integrating R

Feng Ye, Zhijian Wang
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
{yefeng1022, zhjwang}@hhu.edu.cn

Feng Ye, Zhijian Wang, Fachao Zhou, Yapu Wang, Yuanchao Zhou
College of Computer and Information
Hohai University
Nanjing, China
{yefeng1022, zhjwang}@hhu.edu.cn, {ZFCISBEST, wangyapu0714, zhouyuanchao1230}@gmail.com

***Abstract*—As an important resource and productive element, big data permeates all the domains, such as: E-commerce, traffic management or smart city. When possessing the capability of aggregating the information and then mining and analyzing deeply the latent knowledge, it will bring endless innovative achievements. Therefore, big data mining and deep analytics is becoming one of the research hotspots, and has attracted more and more attention from academia, industry as well as government. However, because of the "3Vs (volume, velocity and variety)" characters of the big data, there is no single tool or a one-size-fits-all solution for big data processing. This paper reports our own experiences in building a cloud-based big data mining & analyzing services platform by integrating R for providing rich data statistical and analytic functions. The architecture of the services platform is discussed in details, which includes four layers: infrastructure layer, virtualization layer, dataset processing layer and services layer. Following the whole architecture, the implementation of K-Means algorithm service is introduced as an example. Finally, we propose the conclusion, and explore the research directions in the future.**

Keywords-big data; data mining and analyzing; cloud computing; RHadoop; K-Mean clustering

I. INTRODUCTION

As data volumes continue to increase exponentially, the data tsunami can easily overwhelm traditional analytics tools or platforms designed to ingest, analyze and report. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [1]. The challenge we are facing is not only how to store and manage diverse data but also to effectively analyze the data to gain insight knowledge to make smarter decisions. Currently, a number of works [2-8] have been presented. These researches introduce big data curation [9], mining and analyzing from different aspects, such as status quo, ideas or implementations. For example: [2] introduces the "Lambda Architecture" which provides a general purpose approach to implement arbitrary functions on massive dataset in real time; a scalable deep analytics platform [8] has been implemented. Because of the complexity, there is no single tool or one-size-fits-all solution for deeply mining and analyzing the big data. Moreover, extracting valuable knowledge from massive datasets requires further studies, experiments as well as scalable and smart services, programming tools and applications achieved.

The paper reports our own experiences in building a cloud-based big data mining & analyzing services platform by integrating R for providing rich data statistical and analytics functions. The idea and features of this services platform include: 1) In order to adapt to different size of the dataset, the services platform proposed should not only apply to TB or PB level, but also for small-scale datasets. For small scale data mining task, a common PC is sufficient to fulfill the goals. For big data mining task, a typical processing framework will rely on computers cluster with a high performance computing platform, where a data mining and analytical task is deployed by running some parallel programming tools, such as Map-Reduce. It will bring optimal treatment strategies and cost-effective infrastructure resource utilization. 2) A variety of open-source software and tools are integrated to make full use of the respective characters for data processing, for example: utilize NoSQL [7] for managing multiple dataset and R for data analytics and visualization. 3) Based on the idea of SaaS (Software-as-a-Services) of cloud computing, complex data mining, analytics and visualization functions will be encapsulated into cloud services, so it will be very convenient for the non-programmer users to use friendly Web-based interface to process massive data on the large clusters and get some useful knowledge visually.

The remainder of this paper is structured as follows. Section 2 introduces the related works. After analyzing the key technologies and tools used, we propose the architecture of cloud-based big data mining & analyzing services platform in section 3, which includes four layers: infrastructure layer, virtualization layer, dataset processing layer and services layer; in Section 4, the implementation of k-Means algorithm services is introduced as an example. Finally, the conclusion and the future work follow.

II. RELATED WORKS

According to the most common definition, big data refers to massive, heterogeneous, and often unstructured digital content that is difficult to process using traditional data management tools and techniques [6]. Currently, many works are very valuable to tackle the challenges. In [2], Lambda Architecture and the principles of building systems

978-1-4799-3261-0/14 $31.00 © 2014 IEEE
DOI 10.1109/CBD.2013.13

for big data have been presented. This general approach can implement an arbitrary function on an arbitrary dataset in real time by decomposing the problem into three layers: batch layer, the serving layer and the speed layer. However, limited to the contents of the book not yet published, we just get a good idea and a partial solution. C.Ji, et al [3] described a systematic flow of survey on the big data processing and discussed the key issues in the context of cloud computing. But there is also not a concrete realization of the service or platform.

In [4], the authors present a taxonomy for analytic workflow system and a conceptual architecture of Cloud-based Analytics-as-a-Service, which is a big data analytics service provisioning platform in cloud. However, the implementation is preliminary, lacking the details of how to process the massive dataset. DataCloud [5] is a massive data mining and analysis framework and its key part is RABBIT, which is a kind of massive data analysis tool. However, RABBIT seems not a widely known and used. D.Talia [6] shows a data analysis workflow application designed using the data mining cloud framework's graphical programming interface, but seems to lack rich statistical functions for user. In [7], NoSQL could be used to manage multiple dataset, and it is only the basis of big data mining and analyzing.

Our work is similar to Ricardo [8], which is software system that integrates R statistical tool and Hadoop to support parallel data analysis. Ricardo includes three main components, namely: the R statistical software package, the Hadoop large-scale DMS and the Jaql query language [8]. The biggest difference between Ricardo and our system is the connection between R and Hadoop: we use Rhadoop [10], which is the most mature (and best integrated) project for R and Hadoop [19]; but Ricardo implements the R-Jaql bridge. And both systems consider the solution for processing different "dataset size".

In summary, it is easy to see that there is not a one-size-fits-all solution for big data processing, and there are still huge required research and development efforts needed for the challenge.

III. CLOUD-BASED BIG DATA MINING & ANALYZING SERVICES PLATFORM

In order to tackle the challenge of big data mining and analyzing, we consider to integrating many different key tools and technologies, from infrastructure resource management to rich statistical computation and graphic functions.

A. Cloud Computing and Apache CloudStack

According to NIST, cloud computing [11-12] is defined as "A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [12]. As a highly available, highly scalable Infrastructure-as-a-Service (IaaS) cloud computing platform, Apache CloudStack [13] is an open source software designed to deploy and manage large networks of virtual machines. Our purpose is to utilize Apache CloudStack to manage infrastructure resource by constructing a private cloud, and then integrate relational and NoSQL database for storing multiple datasets. Therefore, it can serve as an effective paradigm for addressing both the computation and data storage requirements of big data mining and analyzing applications.

B. Map-Reduce programming model and Hadoop

Map-Reduce [14-15] is a programming model for expressing distributed computations on massive amounts of data and an execution framework for large-scale data processing on clusters of machines. It consists of two common built-in functions are Map and Reduce. The Map function receives a key/value pair as input and generates intermediate key/value pairs to be further processed. The Reduce function merges all the intermediate key/value pairs associated with the same (intermediate) key and then generates final output. The great strength of Map/Reduce is that both maps and reducers are easily parallelizable, and this programming model fully embodies the idea of "divide and conquer".

Map-Reduce programming model has since enjoyed widespread adoption via an open-source implementation called Hadoop [16-17]. Though Hadoop isn't the best solution for all problems, it is open source, popular, and runs well in the cloud. Based on this reason, Hadoop can be deployed on virtual machines cluster of the private cloud using Apache ClouadStack and adopted for providing distributed big data processing ability in our services platform.

C. R language and RHadoop

R [18-19] provides both an open-source language and an interactive environment for statistical computation and graphics. More and more users choose R to scatter diverse data. But R itself does not have a big data processing capacity. Combining the advantages of R and Hadoop will be feasible for big data mining and analyzing.

There are three ways to combing Hadoop and R language: using the RHadoop package in R, using Segue, and using Hadoop streaming [19]. Because RHadoop is the most mature (and best integrated) project for R and Hadoop, it is selected in our service platform. And it is a collection of four R packages that allow users to manage and analyze data. According to [10], it consists of four packages:

a) plyrmr: this R package enables the R user to perform common data manipulation operations on very large data sets stored on Hadoop.

b) rmr: this R package allows an R user to perform statistical analysis via MapReduce on a Hadoop cluster.

c) rhdfs: this R package provides basic connectivity to the Hadoop Distributed File System. R users can browse, read, write, and modify files stored in HDFS.

d) rhbase: this R package provides basic connectivity to HBASE. R users can browse, read, write, and modify tables stored in HBASE.

At this point, the cloud, Hadoop and the R could be integrated together for the ultimate target of big data mining and analysis.

D. Services Platform Overview

The architecture proposed is shown in Figure1, and there are four tiers: infrastructure layer, virtualization layer, dataset processing layer and services layer. Infrastructure layer provides the hardware foundation for big data processing, such as PCs, various servers and network equipment. Various resources are abstracted into different resource pools, such as data resource pool, network resource pool.

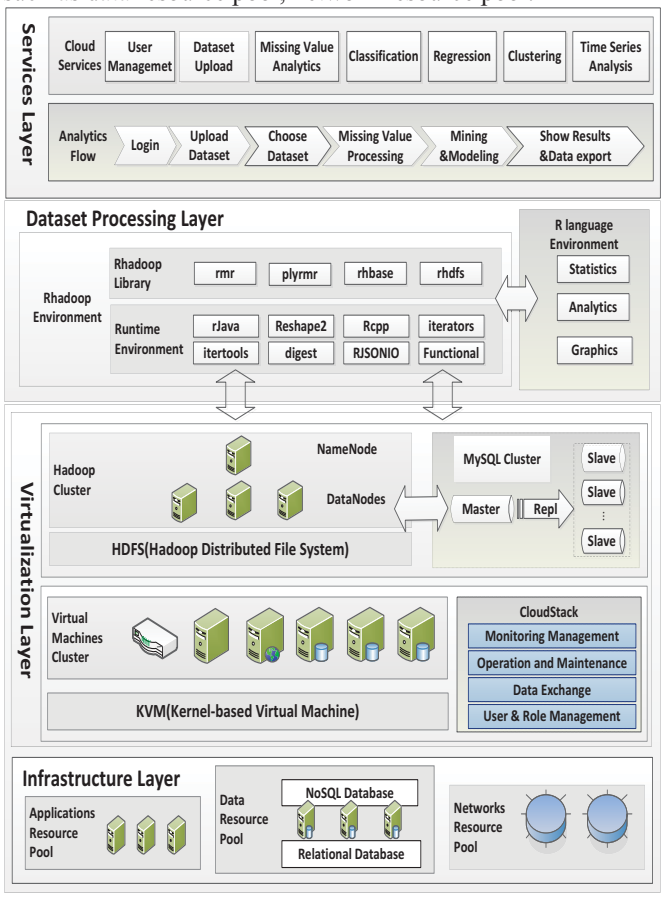

Figure.1 The Architecture of Cloud-based Big Data Mining & Analyzing Services Platform

In virtualization layer, Apache CloudStack is installed, configured and deployed to construct virtual machines cluster and then used to manage the infrastructure resource. Hadoop, NoSQL or relational databases and other tools can be installed in virtual machines cluster. In this layer, according to a variety of business requirements, multiple data management solutions can be coexist, such as MySQL cluster and Hadoop Distributed File System (HDFS). Different data management solutions and diverse storage tools are applicable for different data sizes or types, for example: if local resource of single virtual machine is sufficient for data processing, it is not necessary to use Hadoop.

Above the virtualization layer, it is dataset processing layer. It consists of R language runtime environment and RHadoop environment. R language runtime environment provides many functions, such as: statistics, analytics, graphics and so on. In RHadoop environment, there are RHadoop library and runtime environment. Dataset processing layer support connection and analytic functions and data manipulation operation via MySQL and Hadoop cluster. From the service layer, it receives various requests, and processes these requests. Then the results are returned to the service layer. General speaking, this layer provides the statistical analysis abilities to implement big data analytics.

In services layer, there are four main aspects to be considered: computing mode selection, cloud services, data mining workflow and user interface. Firstly, two computing modes are considered: Map-Reduce mode and single machine mode. According to the volume of dataset, the user can choose one of two modes. Secondly, based on the idea of SaaS, many data mining and analyzing algorithms are implemented into cloud services, including: support vector machines, decision tree and neural network for classification; K-Mean, PAM (partitioning around medoids), CLARA (Clustering LARge Applications), AGNES (AGglomerative NESting) and DIANA (DIvisive ANAlysis) for clustering; and ARIMA (auto-regressive integrated moving average). Thirdly, the platform constructs a data mining and analytics workflow and the users can set the data source, choose the services, setting services parameters and check the results. At last, in order to hide the complexity of the R language, the services platform provides WYSIWYG Web-based user interface for users. The composite images of the user interface and result display interface is shown in Figure 2.

Figure 2. The composite images of the user interface and result display interface

IV. K-MEANS CLUSTERING SERVICES

Many Map-Reduce algorithms for big data analysis have been implemented and discussed in [20]. The use of

parallelization algorithms is the key to achieve better scalability and performance for processing big data. In this subsection, we take the basic idea about K-Means clustering service on this platform as an example.

The details of the K-Means clustering service using pseudo-code is as follows.

The input of the service: HDFS file input path i, HDFS file output path o, field delimiter $pattern$, the number of cluster $num.cluster$, Maximum number of iterations $num.iter$;

The output of the service: cluster centroid C.

1) $C = \text{matrix}(num.clusters)$;

// Initializing the cluster centroid C

2) for i in $1: num.iter\{$

3) C = values(from.dfs(mapreduce(input = i, output = o, vectorized.reduce = FALSE, in.memory.combine = FALSE, map = Map, reduce = Reduce, combine = Combine))));

// Execute Map-Reduce program to get the latest cluster centroid，and return it to R

4) $C = \text{as.matrix}(C)$;

// convert the cluster centroid into matrix

5) $C = \text{reset}(C);\}$

// Reset C to remain the total number of centroid same

6) return C;

In 3), the implementation needs to call Hadoop to execute the Map-Reduce process. The Map function, Combine function and Reduce function can be described as follows.

1. Map function

Input：row number k， value v

Output：the column number of the optimal cluster centroid, the key-value Composed by the v

1) $content = \text{readTable}(v, pattern)$;

//read the data into a data frame

2) $content = \text{dealNa}(content)$;

//tackle the missing value

3) $D = \text{dist\_fun}(C, content)$;

//get the centroid distance to C

4) $nearst = \text{maxCol}(D)$; //get the column number of the optimal cluster centroid

5) $\text{keyval}(\text{c}(nearst), content)$;

2. Combine function

Input：the output of Map function

Output：key-value

1) $\text{keyval}(k, v)$;

3. Reduce function

Input：the output of Combine function

Output：new cluster centroid C

1) $count = \text{size}(v)$; // the number of value

2) $dArray = getMatrix(v)$; // change v into matrix

3) $tm = \text{getAvgOfCol}(dArray, count)$;

//get the average value of each column

4) $\text{keyval}(k, \text{getDataFrame}(tm))$;

//convert tm into a data frame, and return value pairs

After implementing the K-Means clustering service, we test its performance under both single machine computing mode and Map-Reduce mode. The hardware includes 3 PCs, which has 4 core, 2.3G dominant frequency, 4G memory. And one of them acts as a master of Hadoop cluster, and the others act as slaves. The test dataset use and expand an individual household electric power consumption dataset from famous UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption), which initially includes 2075259 instances.

The result is shown in Table I. From the Table, it is easy to find that: in our experiment, good PC can handle only about 1.6G dataset using K-Means clustering; the cloud-based big data mining & analyzing services platform is available for user, and the performance is acceptable.

TABLE I. K-Mean Clustering Executing Time Analytics

Dataset Volume(G)	Computing Time(s)	
	Single Machine Model	*Map-Reduce mode(s)*
0.1	13	17
0.2	35	19
0.8	206	37
1.6	memory overflow	59
6.4	memory overflow	162
25.6	memory overflow	650

V. CONCLUSION AND FUTURE DIRECTION

In this paper, we report our own experiences in building a cloud-based big data mining & analyzing services platform, including: introducing the key technologies used, the architecture and the K-Mean clustering service. Though the services are incomplete and the services platform is preliminary, it is a feasible solution and the experimental result shows that the performance is acceptable for users.

The better management and analysis of big data will become the next frontier of innovation, competition and productivity. And multidisciplinary and trans-disciplinary effort continues to deliver new techniques and tools for the analysis of very large collections of data. The future directions of our work will focus on two aspects: 1) combine business requirement s in the field of water information, and utilize the services platform to analyze the dataset related; 2) develop more data statistical and analytic services which is based on Map-Reduce programming model.

ACKNOWLEDGMENT

The starting point for the research comes from the tournament title 7 of the 2nd University Student Software Design Competition of "China Software Cup" in 2013.

The research is supported by the Water Science and Technology Project of Jiangsu Province No.2013025 and in part by the National Natural Science Foundation of China under Grant No. 61300122.

REFERENCES

[1] IBM. What is big data: Bring big data to the enterprise, 2012, http://www-01.ibm.com/software/data/bigdata/

[2] N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems (MEAP Edition). Manning Publications, 2012, pp. 1-25.

[3] C. Ji, Y. Li, W. Qiu, et al. Big Data Processing in Cloud Computing Environment. San Marcos, 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), 2012, pp. 17-23.

[4] F. Zulkernine, M. Bauer, A. Aboulnaga. Towards Cloud-based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud. Santa Clara, 2013 IEEE International Congress on Big Data, 2013, pp. 62-69.

[5] G. Zhang, C. Li, Y. Zhang, et al. DataCloud: An Efficient Massive Data Mining and Analysis Framework on Large Clusters. Haikou, 2012 Ninth Web Information Systems and Applications Conference (WISA), 2012, pp. 198-203.

[6] D. Talia. Clouds for Scalable Big Data Analytics. Computer, Vol.46 No.5, 2013, pp. 98-101.

[7] L. Zhao, S. Sakr, A. Liu. On the Spectrum of Web Scale Data Management [A]. L. Wang, R. Ranjan, J. Chen, et al. Cloud Computing: Methodology, Systems and Applications [M]. Boca Raton: CRC Press, 2012, pp. 488-506.

[8] S. Das, Y. Sismanis, K.S. Beyer, et al. Ricardo: Integrating R and Hadoop. New York, Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 987-998.

[9] T. Hey, S. Tansley, K. Tolle. The Fourth Paradigm: Data-Intensive Scientific Discovery. Washington: Microsoft Research, 2009, pp. xii-xiii.

[10] RHadoop. https://github.com/RevolutionAnalytics/RHadoop/wiki

[11] B. Furht and A. Escalante. Handbook of Cloud Computing [M]. New York: Springer, 2010, pp. 1-20.

[12] A. Kalapatapu and M. Sarkar. Cloud Computing: An Overview [A]. L. Wang, R. Ranjan, J. Chen, et al. Cloud Computing: Methodology, Systems and Applications [M]. Boca Raton: CRC Press, 2012, pp. 6-24.

[13] Apache Cloudstack. http://cloudstack.apache.org/

[14] J. Lin, C. Dyer. Data-Intensive Text Processing with MapReduce. Morgan & Claypool Publishers, 2010, pp. 1-64.

[15] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Communications of the ACM, vol. 51 no. 1, 2008, pp. 107–113.

[16] Hadoop. http://hadoop.apache.org/

[17] T. White. Hadoop: The Definitive Guide (Third Edition). Sebastopol: O'Reilly, 2012, pp. 1-16.

[18] J. Maindonald and W.J. Braun. Data Analysis and Graphics Using R (Third Edition). New York: Cambridge University Press, 2010, pp. 1-41.

[19] J. Adler. R in a Nutshell (Second Edition). Sebastopol: O'Reilly, 2012, pp. 213-500.

[20] Kyuseok Shim. MapReduce Algorithms for Big Data Analysis. Proceedings of the VLDB Endowment, 2012, pp. 2016-2017.