2015 International Conference on Cloud Computing and Big Data

A Study of Chinese Character Culture Big Data Platform

Guigang Zhang, Jian Wang, Weixing Huang,Yi Yang, Haixia Su,Ye Yue, Yichen Zhai, Manxian Liu, Lijuan Chen
Institute of Automation
Chinese Academy of Sciences
Beijing 100190, China
{guigang.zhang, jian.wang, weixing.huang, yangyi, haixia.su}@ia.ac.cn
{ yueye, 1047327976, 351408369, 1046845469 }@qq.com

***Abstract*—Chinese Characters are important elements of Chinese culture. Digitization techniques of Chinese Characters introduce attractive experiences of Chinese Character culture. The Chinese Character digitization generates large-scale data that is hard handled by traditional methods. To address this issue, in this paper, we propose Chinese Character Culture Big Data Platform that is designed based on the three-level hierarchy of cloud computing. The platform is built by means of open source technologies, for example, Hadoop ecosystem. The platform can effectively store, manage, and analyze large-scale data of digitized Chinese Character for supporting the Chinese Character culture experience application systems. This paper presents design concepts and architecture of the platform, as well as an experiment.**

Keywords—Chinese Characters; Chinese Culture; Big Data; Cloud Computing; Architecture

I. INTRODUCTION

Chinese Character digitization is a major part of National Cultural Resource Sharing Project of China. Although Chinese Characters have more than five thousand years of history，it still keeps alive up to now. Chinese Characters have particular advantages by integrating shape, pronunciation, and meaning of Chinese. Chinese Characters come from everyday life so that they are strongly related to Chinese culture. In order to effectively handle large-scale digitized Chinese Character culture information, in this paper, we propose Chinese Character Culture Big Data Platform (C3BDP). The architecture of the platform is based on Hadoop ecosystem. It provides big data store and management and big data processing. The platform can provide on-line experience services, off-line experience services, Access service of digitized Chinese Character resources, data analysis services. The services are implemented by kinds of system modules. Logically, the platform uses a central management system to manage and coordinate data management module, application system module, system function module. C3BDP provides interactive application systems by application system module for digitized Chinese Character culture. Our experiment shows that C3BDP can provide better capability regarding scalability than the traditional non-big data technology based system.

Section 2 discusses state-of-the-art of digitization techniques of Chinese Character culture and current research work of cloud and big data technologies. Section 3 introduces concepts of Chinese Character Culture Big Data Platform and Section 4 shows architecture of the platform in detail. Section 5 presents an evaluation of C3BDP. Finally, the work is concluded in Section 6.

II. RELATED WORK

A. Chinese Character Culture Digitization Technologies

Chinese character culture digitization technology is leading Chinese Characters to the cultural and creative industries. At present, Chinese Character culture digitization technologies have been developed and applied in many fields, for example, Chinese Character Dance [1] is a new form for dynamically presenting Chinese Characters by means of dancing and stage effects such as stage lights, fireworks. Interactive Chinese Character games are developed according to touching stories related to Chinese Characters. Many Chinese Character culture theme parks have been built in recent years [2]. These parks integrate Chinese culture, technology, and art in order to provide more attractive experiences of Chinese Character culture. Moreover, according to characteristics of Chinese Characters, researchers created cartoon images named Chinese Character Dolls [3].

B. Big Data and Cloud Platform

Currently, Google cloud platform [4] and open source Apache Hadoop ecosystem [5][6][7][8] are mainstream techniques. The techniques are very good at batch processing of large-scale data, while lacking of effective real-time analysis and interactive analysis. Researchers proposed some techniques for the above requirements, such as Google Dremel [9], Apache Spark Streaming [10][11], as well as Apache Storm [15]. In addition, Cloudera provides Impala technique [16] by which users can access big data on Hadoop at a high speed using SQL-like statements.

There are a couple of techniques based on MapReduce programming model [5][6][7][8][17]. Google MapReduce is the template of other kinds of MapReduce models. Hadoop MapReduce is the most famous open source version of Google MapReduce [17]. Haloop [18] and Twister [19][20] are iterative models based on MapReduce. Apache Spark [12][13][14] is an in-memory computing model based on the basic concepts of MapReduce. HadoopDB [21][22] takes MapReduce and PDBMS into account so that it has the batch

978-1-4673-8350-9/15 $31.00 © 2015 IEEE
DOI 10.1109/CCBD.2015.18

processing of MapReduce as well as the semantic query of relational databases. Moreover, several methods were proposed in order to optimize the MapReduce Model, for example, Hadoop++ [23][24], and Microsoft Azure [25].

In culture field, big data technologies have been highly accepted. OCLC(Online Computer Library Center) [26] provides World Cat Local instead of the traditional OPAC(Online Public Access Catalog) in order that each library can upload its catalog data to World Cat Cloud Platform and share the data from other libraries. In October 2010, CALIS (China Academic Library & Information System) [27] organized the "Cloud Computing Technology & Application Workshop China Academic Libraries". Recently, Cloud and Big Data technologies are introducing new chances and revolutionary changes to applications and services of digitized cultural resources.

There are several Cloud and Big Data platforms developed by open source communities and business companies. The most typical ones are Google Cloud Platform, Apache Hadoop, Apache Spark Framework, and Microsoft Azure.

1) Google Cloud Platform

Google cloud is the most famous big data platform in the world. Google proposed three core techniques for the platform: GFS [28], MapReduce [29], and BigTable [30]. GFS (Google File System) is a scalable, high available, distributed file system. GFS is good at large-scale data store. MapReduce is a parallel programming model that focuses on the batch data processing of huge data set. BigTable is a distributed storage system that is built on GFS. It supports PB level data and works very well for storing sparse data.

Based on the three core technologies, Google provides a PaaS technique, GAE(Google App Engine) [31][32], as running environment of applications systems. For safety and security reasons, GAE isolates applications by SandBox technique so that data of an application can be kept invisible against other applications. In the same way, errors of an application will not influence others applications.

Google cloud is a very classic platform that is a de facto standard in big data field.

2) Apache Hadoop Platform

Apache Hadoop is a very famous open source big data platform. The concept of Hadoop came from Google cloud Platform. Hadoop provides two core components: HDFS and MapReduce. HDFS (Hadoop Distributed File System) is a scalable and high available distributed system. Hadoop MapReduce is a parallel programming model that is an open source version of Google MapReduce.

Based on Hadoop platform, Hadoop ecosystem can be built by more advanced components. Apache HBase [33][34] is a columnar database which has concepts similar to Google BigTable. Apache Hive [35][36] is a hadoop-based data warehouse and provides data query with SQL-like statements. Apache Pig [37][38] is a large-scale data processing system. It is able to analyze large data set using MapReduce model with script language. Mahout is a MapReduce-based machine learning library which provides parallel computing models of commonly used machine learning algorithms. Moreover, there are also important components for supporting the Hadoop ecosystem: Zookeeper [39][40] being a coordinator of the Hadoop distributed system, Flume [41] being a log collector, and Sqoop [42] working on data transformation between HDFS and relational databases.

Hadoop is the most widely used big data platform. Many open source projects and commercial products are built upon Hadoop ecosystem.

3) Apache Spark Framework

Apache spark is an in-memory computing framework based on Hadoop. It loads all related data into memory for data processing so that the speed is much faster than traditional disk-based MapReduce, even 100 times faster.

Based on Spark platform, a couple of important components have been proposed. Spark SQL [43][44][45] provides high-speed data access on Hadoop platform with SQL-like statements. Spark Streaming is a streaming computing technique that has pretty good real-time computing performance. Spark GraphX [46][47] is a graph computing framework that can be used to implement the methods based upon Graph theory, for example, algorithms for single source shortest path (SSSP). Spark MLlib [48] is a machine learning library implemented. SparkR [49] is a new component of Spark framework that provides statistical analysis of big data by using characteristics of R programming language.

Apache Spark framework is a new hotspot in big data field. It is growing up very fast and has been applied in many applications instead of disk-based MapReduce model.

4) Microsoft Azure Platform

Microsoft Azure is a cloud platform that depends on a basic component, Windows Azure, that consists of three parts: computing service, storage service, and system management.

Based on the Windows Azure, a couple of services are provided. App Fabric is a middleware collection on Azure platform, which basically works on port mapping and access control. SQL Azure provides relational database services that are mostly related to Microsoft SQL Server. Dot Net Service provides .Net framework as a service for developers.

Microsoft Azure provides an integrated big data service platform. It has good usability and compatibility, particularly for the software projects based on Microsoft techniques.

III. Concepts of Chinese Character Culture Big Data Platform

A. Conceptual Architecture of Chinese Character Culture Big Data Platform

In order to store, manage, and analyze large-scale data of digitized Chinese Character culture, we design a big data system named Chinese Character Culture Big Data Platform (C3BDP)(Fig. 1). The core of C3BDP is Resource Database that stores and manages data resources of C3BDP. The data resources consist of the following types of data sets:

1) *Chinese Character related data:* for example, fonts, colorful fonts, 3D fonts, Chinese calligraphy artist database

2) *System operation data of platform:* for example, logs

3) *Data to be analyzed:* for example, user behavior database

C3BDP provides access services and analysis services of Chinese Character data. Users can use the services by Service Directory. C3BDP defines four basic types of services:

1) On-line experience service: Users may experience different kinds of topics of Chinese Characters and the related culture. The experience topics can be implemented using Human-Computer interaction techniques and Virtual Reality techniques in order to increase attractions of the topics. The topics are classified into three levels:

a) Experience of Chinese Character: For each Chinese Character, users may experience the following topics with regard to basic information of the character:

i. Glyph: shape of the character

ii. Pronunciation: read the character aloud

iii. Semantic: meaning of the character

iv. History: evolution history of the character

v. Art: work of art regarding the character

b) Experience of Chinese Character Culture: Users may experience characteristics of Chinese Characters with the following topics:

i. Chinese characters as an imagery characterization

ii. The creation of Chinese Characters associated certainly with Chinese ancestors' cultural conscious

iii. Constitutes art of Chinese Characters and the way of Chinese thinking

iv. Constitutes art of Chinese Characters and Chinese Aesthetic Spirit

v. Constitutes art of Chinese Characters and Chinese Native Writing

c) Experience of Chinese Character Culture: Users may experience the culture patterns of Chinese Character including the following topics:

i. Novel of Chinese Culture keywords

ii. Documentary of Chinese Civilization keywords

Based on the experience topics, derivative products can be developed. The on-line experience services and derivative products communicate with Resource Database: getting resource data from Resource Database, and sending experience data as well as product data to Resource Database.

2) Off-line experience service: Beside on-line experience service, C3BDP provides off-line experience service as well. The off-line experience can be implemented in the form of devices, for example, the Chinese calligraphy experience device. Corresponding to the three levels of on-line experience, various experience devices can be developed. These devices upload experienced data to Resource Database.

3) Access service of digitized Chinese Character resources: Digitized Chinese Character resources can be used for research, investigation, education, and product development by personal customers and business customers. The customers download required resources from Resource Database and upload new resources to upgrade this database.

4) Analysis service: Analyzing data of Resource Database with suitable algorithms, hidden patterns can be discovered. The patterns are useful for the following purposes:

a) System resource optimization service: weaknesses of the platform can be identified by analyzing system logs for the purpose of system resources optimization.

b) Recommendation service: By use of recommendation algorithms, C3BDP can recommend personalized information to personal customers and business customers.

c) Statistical analysis service: Many patterns can be easily identified by statistical analysis of data stored in Resource Database. This kind of service is also provided to different types of customers.

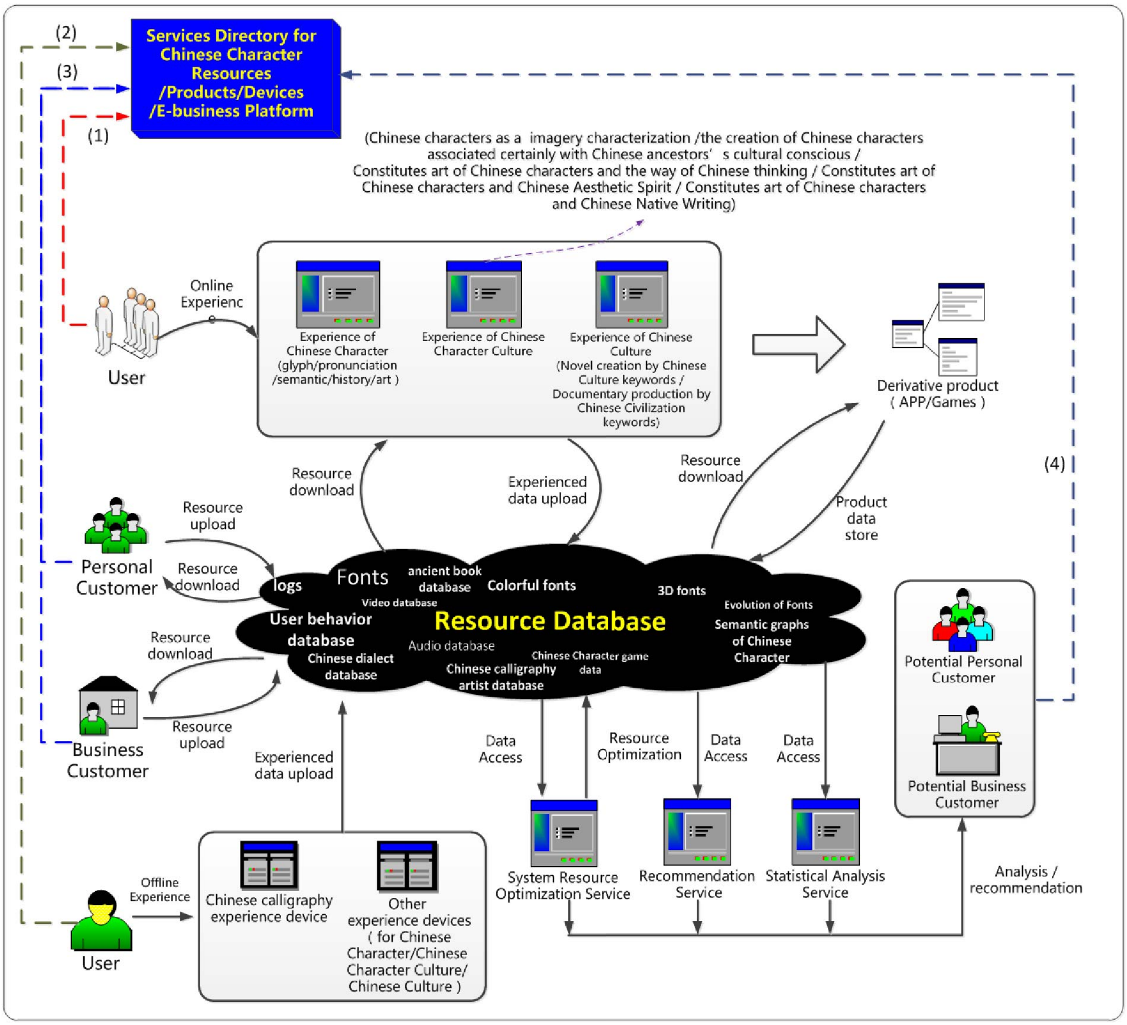

(2)
(3)
(1)
(4)
Services Directory for Chinese Character Resources /Products/Devices /E-business Platform
(Chinese characters as a imagery characterization /the creation of Chinese characters associated certainly with Chinese ancestors' s cultural conscious / Constitutes art of Chinese characters and the way of Chinese thinking / Constitutes art of Chinese characters and Chinese Aesthetic Spirit / Constitutes art of Chinese characters and Chinese Native Writing)
User
Online Experience
Experience of Chinese Character (glyph/pronunciation /semantic/history/art)
Experience of Chinese Character Culture
Experience of Chinese Culture (Novel creation by Chinese Culture keywords / Documentary production by Chinese Civilization keywords)
Derivative product (APP/Games)
Resource download
Experienced data upload
Resource download
Product data store
Personal Customer
Resource upload
Resource download
Business Customer
Resource download
Resource upload
logs
Fonts
ancient book database
Colorful fonts
3D fonts
Video database
Evolution of Fonts
User behavior database
Resource Database
Semantic graphs of Chinese Character
Audio database
Chinese dialect database
Chinese Character game data
Chinese calligraphy artist database
Potential Personal Customer
Potential Business Customer
Experienced data upload
Data Access
Resource Optimization
Data Access
Data Access
System Resource Optimization Service
Recommendation Service
Statistical Analysis Service
Analysis / recommendation
Offline Experience
User
Chinese calligraphy experience device
Other experience devices (for Chinese Character/Chinese Character Culture/ Chinese Culture)

Fig. 1 Conceptual Architecture of Chinese Character Culture Big Data Platform

B. *Central Management System Of Chinese Characters Big Data Platform*

C3BDP has a central management system that coordinates application system module, system function module as well as the data management module (Fig. 2). The central system management system depends on the core of C3BDP, i.e., infrastructure, fundamental storage services and fundamental computing services. The application system module consists of the following online experience applications of C3BDP:

1) Chinese Character stoke analysis system: analysis of Chinese Character strokes and stroke rules

2) Chinese Character coloring system: coloring strategies and configurations for Chinese Characters

3) Chinese Character evolution analysis system: analysis of evolutional process of Chinese Characters

4) Chinese Character structure analysis system: structure analysis of single Chinese Characters

5) Chinese Character 3D font system: 3D fonts for Chinese Characters

6) Chinese Character speech processing system: speech recognition and synthesis of Chinese and Chinese Characters

Beside the application system module, we also design the following supporting system functions:

1) User login (including voice login): user management including authorization and authentication.

2) Resource Administration: management of resources of the platform and application systems.

3) *Resource statistical analysis system:* statistical analysis of Chinese Character resource data and platform resource data.

4) Integration of application systems: coordination and configuration of application systems.

5) Security policy management: definition and configuration of C3BDP security policy including firewall, transmission security, authorization and authentication, and security audit.

6) Other functions: extra functions, for example, data access interface and service interface supporting the resource access service.

The central management system also works on data management module that includes fonts, Chinese Calligraphy library, platform management information database, and other databases. The central management system provides the following purposes:

1) Resource administration purpose: focusing on adding, removing, and updating Chinese Character data and system management data with the following steps:

(1) Administrator login

(2) Administrator obtains the authority from the central management server

(3) Administrator manages resources.

(4) Changes of the resources are submitted to the basic storage system, for example, HDFS.

2) Experience purpose: aiming at the online experience by using the sub-systems with the following steps:

(1) User login

(2) User uses the application systems

3) Analytic purpose: corresponding to the analysis services. The process consists of the following steps:

(1) Analyst login

(2) Analyst requests the statistical analysis function

(3) Analyst obtains the analysis results

(4) Analyst writes reports

(5) Reports are provided to users by personalized recommendation

(6) Reports are provided to leader/decision maker

By the central management system, C3BDP integrates the different modules with the core components of the platform in order to provide the data service, the experience service, and the analysis service.

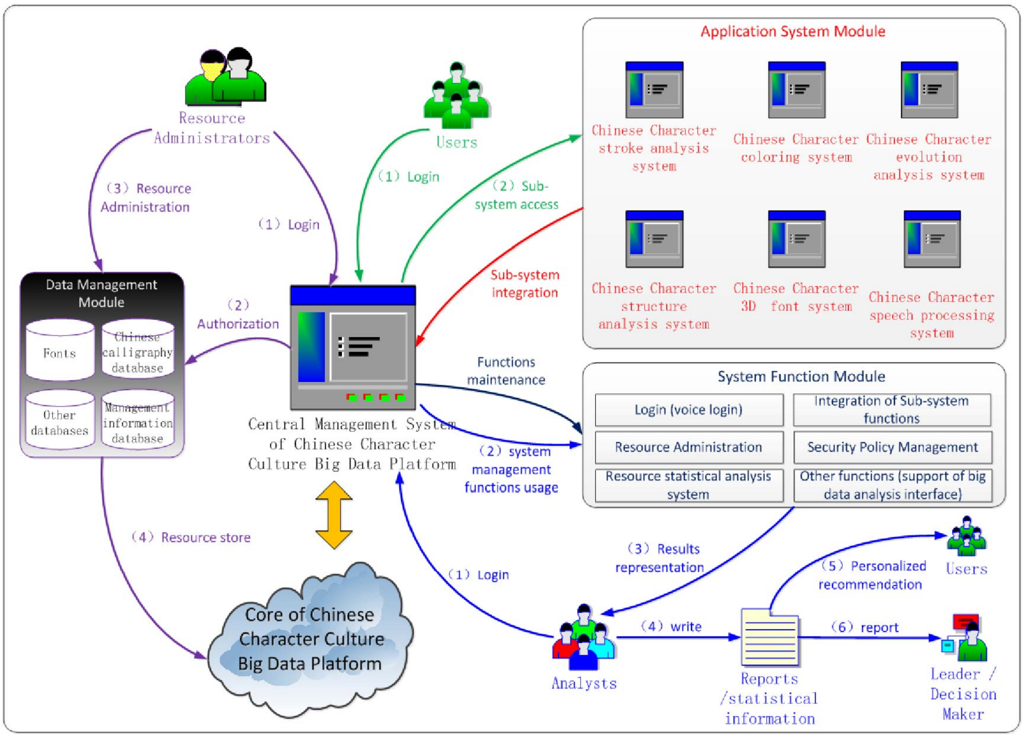

Application System Module
Chinese Character stroke analysis system
Chinese Character coloring system
Chinese Character evolution analysis system
Chinese Character structure analysis system
Chinese Character 3D font system
Chinese Character speech processing system
Resource Administrators
Users
(3) Resource Administration
(1) Login
(1) Login
(2) Sub-system access
Sub-system integration
Data Management Module
Fonts
Chinese calligraphy database
Other databases
Management information database
(2) Authorization
Functions maintenance
Central Management System of Chinese Character Culture Big Data Platform
(2) system management functions usage
System Function Module
Login (voice login)
Integration of Sub-system functions
Resource Administration
Security Policy Management
Resource statistical analysis system
Other functions (support of big data analysis interface)
(4) Resource store
(1) Login
(3) Results representation
(5) Personalized recommendation
Users
Core of Chinese Character Culture Big Data Platform
Analysts
(4) write
(6) report
Reports /statistical information
Leader / Decision Maker

Fig. 2 Central Management System of C3BDP

IV. Architecture of Chinese Characters Big Data Platform

We design the architecture of C3BDP based on the three-level hierarchy of cloud computing. In order to test the availability of our design, we implemented a prototype with a basic architecture. The architecture contains all core elements of C3BDP and it can be easily expanded according to the specific requirements.

A. Topology of Infrastructure of Chinese Character Culture Big Data Platform

IaaS of C3BDP is designed as a computer cluster (Fig. 3). Virtual machines are created with virtualization techniques and connected each other to form a virtual computer cluster. Linux are installed as operating system. Fig. 3 shows the basic topology of the cluster with the following main parts that will be described in the next section:

1) Application environment: Nginx server [50][51], and Tomcat server [52][53]

2) Resource Administration Server: server that monitors resource usage of MySQL Cluster nodes and Hadoop ecosystem nodes.

3) MySQL Cluster [54]

4) Hadoop ecosystem

By using virtualization technology we are able to flexibly administrate system hardware resources and reduce the cost of cluster construction and operation.

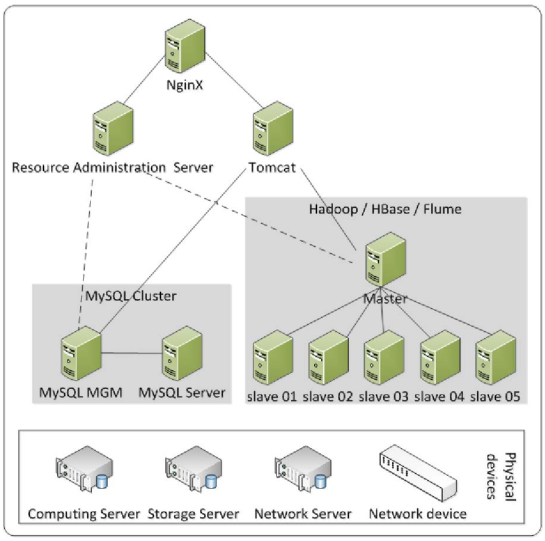


Fig. 3 Topology of virtual computer cluster in C3BDP

B. PaaS of Chinese Character Culture Big Data Platform

PaaS of C3BDP provides kinds of data storage service and computing service, and environments for application systems. The basic PaaS structure consists of the following components (Fig. 4):

1) Nginx: important component in C3BDP as reverse proxy (Firewall), load balancer, and web server.

2) Tomcat: application server that communicate with MySQL Cluster and Hadoop ecosystem. Tomcat can be scaled out with the help of the Nginx server.

3) MySQL Cluster: persistent storage cluster that stores structured data in relational tables. MySQL Cluster manages a high available, scalable, distributed relational database.

4) Hadoop ecosystem: Hadoop platform and important scalable components including HBase and Spark. HBase is a column-oriented database that is suitable to store sparse data. In-memory computing framework Spark provides the streaming computing technique, graph computing technique, and machine learning library. Hadoop provides log collection system Flume as well.

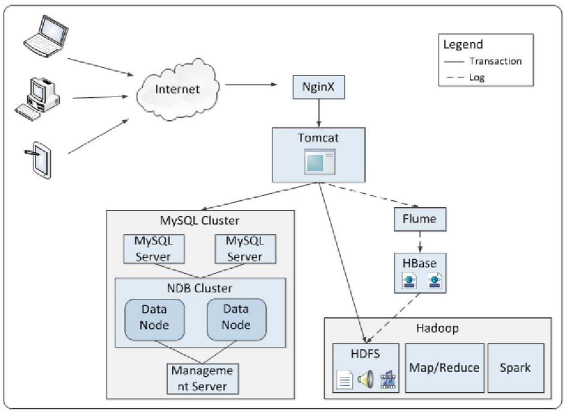


Fig. 4 Design Concepts of PaaS of C3BDP

C. GUI of Chinese Character Culture Big Data Platform

In SaaS, we provide an integrated GUI for monitoring platform as well as the application systems. The typical components of the GUI are illustrated as follows:

1) Hadoop Platform Resource Monitoring by integrating Cloudera Manager GUI [55] (Fig. 5(a))

2) MySQL Cluster Monitoring by integrating phpMyAdmin GUI [56] (Fig. 5 (b))

3) Hadoop HDFS File Management (Fig. 5 (c))

4) Applications of Chinese Character Culture Big Data Platform (Fig. 5 (d)). By Internet browser, users can enter the application systems.

V. EVALUATION

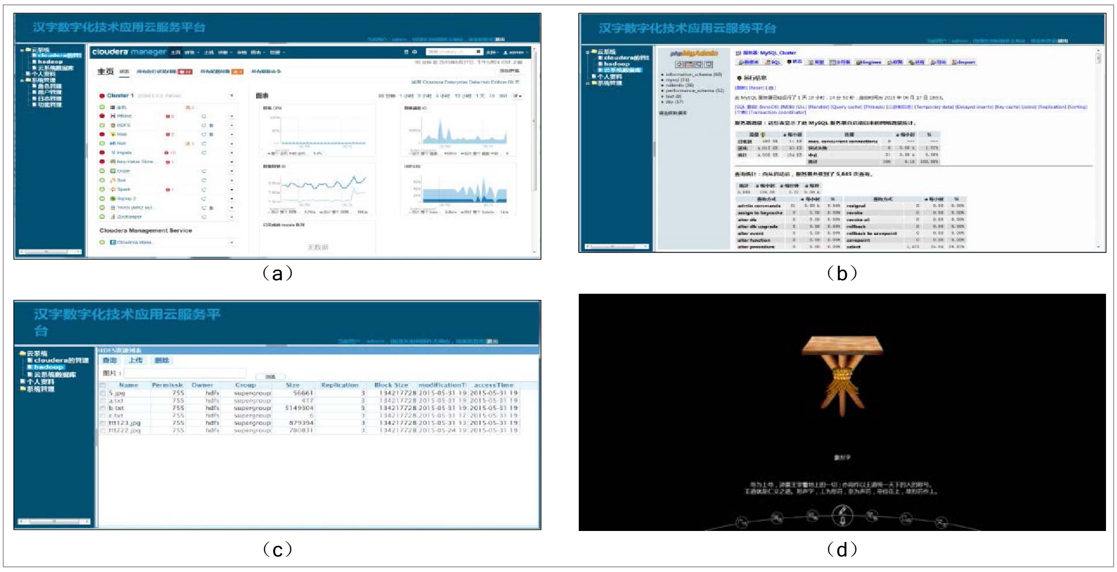
We made an experiment for C3BDP in order to evaluate the scalability of the platform. In the experiment, we compared our platform and the traditional systems that were built with standard B/S architecture rather than big data architecture. Table 1 shows configuration the experiment.

TABLE 1 Experiment Configuration

System	Hardware	Software
C3BDP	CPU: 4core 2.0G Memory: 8G HDD: 100G Network: 1000M	CentOS 6.6 Nginx 1.6.2 Tomcat 7 MySQL Cluster 7.4.6 Hadoop 2.5
Traditional system	CPU: 4core 2.0G Memory: 8G HDD: 100G Network: 1000M	CentOS 6.6 Tomcat 7 MySQL 5.6.23

The hypothesis of the experiment was that C3BDP can be better scalable than the traditional systems.

We sent http requests and monitored the average response time and average throughput rate. The requests involved web site request, database request, and file request.



	Name	Permissic	Owner	Group	Size	Replication	Block Size	modificationT	accessTime
□	5.jpg	755	hdfs	supergroup	56661	3	134217728	2015-05-31 19	2015-05-31 19
□	a.txt	755	hdfs	supergroup	417	3	134217728	2015-05-31 19	2015-05-31 19
□	b.txt	755	hdfs	supergroup	5149304	3	134217728	2015-05-31 19	2015-05-31 19
□	c.txt	755	hdfs	supergroup	6	3	134217728	2015-05-31 17	2015-05-31 19
□	ttt123.jpg	755	hdfs	supergroup	879394	3	134217728	2015-05-31 13	2015-05-31 19
□	ttt222.jpg	755	hdfs	supergroup	780831	3	134217728	2015-05-24 19	2015-05-31 19

（a） （b）

（c） （d）

Fig. 5 GUI of C3BDP

Fig. 6 illustrates the comparison between C3BDP and the traditional System. When number of concurrent requests is less than 200, there are no significant difference between C3DBP and the traditional system regarding response time. Similarly, the average throughput rates of the both systems are also very similar. We conclude that the C3BDP has similar performance with the traditional system when total of concurrent requests under 200.

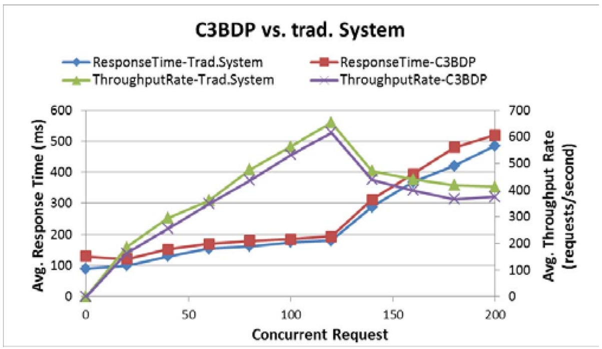


Fig. 6 Comparison between C3BDP and trad. System

As shown in Fig. 6, when the number of concurrent requests is more than 120, the average throughput rate strongly drops from 650 to 400. The traditional system cannot handle this case because its scalability is not good enough. We ran four tests on C3BDP. In turn, we scaled out MySQL cluster, Tomcat, Tomcat again, and Hadoop node. The result is shown in Fig. 7. It shows that C3BDP can provide larger average throughput rate by scaling out the system components. The average throughput rate can rise to over 2000 requests/second.

Based on the results, we can say that the hypothesis is right. That means, C3BDP can provide much better scalability for supporting large count of concurrent requests while providing satisfactory performance in terms of average response time.

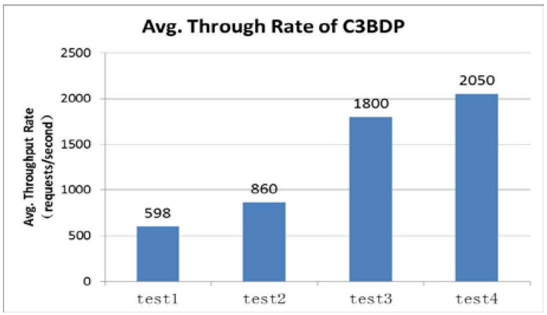


Fig. 7 Average throughput rate of C3BDP

VI. CONCLUSION AND FUTURE WORK

Chinese Characters are core elements of Chinese culture. Recently，data volume of digitized Chinese Characters is rapidly increasing. In this case, data management and analysis is getting hard using traditional methods. To address this issue, in this paper, we propose Chinese Character Culture Big Data Platform (C3BDP). C3BDP provides store and management of large-scale structured data and unstructured data of digitized Chinese Characters. C3BDP provides different kinds of programming models for big data processing as well. C3DBP provides application systems to users. In addition, C3BDP provides a resource administration system for managing resource data. We made an experiment for C3BDP. Results shows that C3BDP has better scalability than the traditional non-big data technology based system. In the future, we will focus on the optimization of C3BDP by using novel approaches, for example, adapted load balancing algorithms. In addition, we will enhance the security management with advanced methods.

ACKNOWLEDGMENT

This research is supported by: (1) the Support Program of the National ‘12th Five-Year-Plan of China’ under Grant No.

2015BAK25B04; (2) the Support Program of the National '12th Five-Year-Plan of China' under Grant No. 2015BAK25B03; (3) National Cultural Resource Sharing Project (key project of '13th Five-Year Plan of China'); (4) Special Project for Civil-Aircraft, MIIT; (5) 2014 annual central cultural industry development fund.

REFERENCES

[1] Chinese Character Dance. http://www.hanziwenhua.com/html/hanzichuangyiwenhuaxiliechanye/20070515/23.html

[2] Chinese Character Culture Theme Park. http://www.hanziwenhua.com/html/hanzichuangyiwenhuaxiliechanye/20070515/24.html

[3] Chinese Character Dolls. http://www.hanziwenhua.com/html/hanzichuangyiwenhuaxiliechanye/20070515/27.html

[4] Google Cloud Computing. https://cloud.google.com

[5] Apache Hadoop. http://hadoop.apache.org/

[6] K.Shvachko, K.H. Rong,S.Radia,et al. The Hadoop Distributed File System: Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium, 2010. Incline Village, NV:IEEE, 2010:1-10.

[7] J. Venner. Pro Hadoop. Apress, 2009.

[8] T. White. Hadoop: The Definitive Guide. O'Reilly Media, Yahoo! Press, 2009.

[9] S. Melnik, A. Gubarev, J.J. Long, et al. Dremel: Interactive Analysis of Web-Scale Datasets. Proc. of the 36th VLDB (2010), pp. 330-339, 2010.

[10] Spark Streaming. https://spark.apache.org/streaming/

[11] M. Zaharia, T. Das, H.Y. Li, et al. Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters. Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing, Pages 10-10, USENIX Association Berkeley, CA, USA, 2012.

[12] Apache Spark. http://spark.apache.org

[13] M. Zaharia, M. Chowdhury, M.J. Franklin, et al. Spark: Cluster Computing with Working Sets:Proceeding HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, 2010. Berkeley, CA, USA: USENIX Association, 2010:10.

[14] M. Zaharia, M. Chowdhury, T. Das, et al. Spark: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing: Resilientdistributed datasets: a fault-tolerant abstraction for in-memory cluster computing, 2012.Berkeley, CA, USA: USENIX Association, 2012:2.

[15] Apache Storm. https://storm.apache.org/

[16] Cloudera Impala. http://impala.io/

[17] J. Dean, S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters: OSDI'04: Sixth Symposium on Operating System Design and Implementation 2004.New York, NY, USA:ACM, 2008:107-113.

[18] Y.Y. Bu, B. Howe, M. Balazinska, M. D. Ernst. HaLoop: efficient iterative data processing on large clusters. Journal Proceedings of the VLDB Endowment Volume 3 Issue 1-2, September 2010, Pages 285-296.

[19] Twister. http://www.iterativemapreduce.org/

[20] Jaliya Ekanayake, Shrideep Pallickara, and Geoffrey Fox MapReduce for Data Intensive Scientific Analysis, Fourth IEEE International Conference on eScience, 2008, pp.277-284.

[21] HadoopDB. http://db.cs.yale.edu/hadoopdb/hadoopdb.html

[22] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel J. Abadi, Avi Silberschatz, Alex Rasin. HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proceedings of the VLDB Endowment, Volume 2 Issue 1, August 2009,Pages 922-933.

[23] Hadoop++. https://infosys.uni-saarland.de/projects/hadoop.php

[24] Jens Dittrich, Jorge-Arnulfo Quiane-Ruiz, Alekh Jindal, Yagiz Kargin, Vinay Setty, and Jörg Schad. Hadoop++: making a yellow elephant run like a cheetah (without it even noticing). Proceedings of the VLDB Endowment Volume 3 Issue 1-2, September 2010,Pages 515-529.

[25] Microsoft Azure. http://azure.microsoft.com/

[26] OCLC. https://www.oclc.org/

[27] CALIS. http://www.calis.edu.cn/

[28] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google File System .19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003.

[29] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, Dec. 2004.

[30] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A Distributed Storage System for Structured Data. OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006.

[31] Sindhu Srivastava, Vani Trehan, Priyanka Yadav, Neha Manga, Sakshi Gupta. Google App Engine. International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[32] GAE. https://cloud.google.com/appengine/docs

[33] Apache HBase. http://hbase.apache.org/

[34] D. Carstoiu, A. Cernian, A. Olteanu. Hadoop Hbase-0.20.2 performance evaluation: New Trends in Information Science and Service Science (NISS), 2010 4th International Conference on, 2010.IEEE,2010,84-87.

[35] Apache Hive. https://hive.apache.org/

[36] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, et al. Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment,Volume 2 Issue 2, August 2009,Pages 1626-1629.

[37] Apache Pig. https://pig.apache.org/

[38] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, et al. Pig latin: a not-so-foreign language for data processing. SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Pages 1099-1110.

[39] Apache Zookeeper. https://zookeeper.apache.org/

[40] Patrick Hunt, Mahadev Konar, Flavio P. Junqueira, et al. ZooKeeper: wait-free coordination for internet-scale systems. USENIXATC'10 Proceedings of the 2010 USENIX conference on USENIX annual technical conference, Pages 11-11 USENIX Association Berkeley, CA, USA, 2010.

[41] Apache Flume. https://flume.apache.org/

[42] Apache Sqoop. http://sqoop.apache.org/

[43] Spark SQL. https://spark.apache.org/sql/

[44] Reynold S. Xin, Josh Rosen, Matei Zaharia, et al. Shark: SQL and rich analytics at scale. SIGMOD '13 Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, Pages 13-24, ACM New York, NY, USA, 2013.

[45] Michael Armbrust, Reynold S. Xin, Cheng Lian. Spark SQL: Relational Data Processing in Spark. SIGMOD '15 Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Pages 1383-1394, ACM New York, NY, USA 2015.

[46] Spark GraphX. https://spark.apache.org/graphx/

[47] Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, et al. GraphX: a resilient distributed graph system on Spark. GRADES '13 First International Workshop on Graph Data Management Experiences and Systems, Article No. 2. ACM New York, NY, USA 2013.

[48] Spark MLlib. https://spark.apache.org/docs/1.2.1/mllib-guide.html

[49] Spark R. https://amplab-extras.github.io/SparkR-pkg/

[50] Nginx. http://nginx.org/

[51] Will Reese. Nginx: the high-performance web server and reverse proxy. Linux Journal, 2008(173), Article No.2. Belltown Media, Houston, TX.

[52] Apache Tomcat. http://tomcat.apache.org/

[53] Jason Brittain, Ian F. Darwin. Tomcat: The Definitive Guide, 2nd Edition.O'Reilly Media, 2007.

[54] MySQL. MySQL AB. http://www.mysql.com/

[55] Cloudera. http://www.cloudera.com

[56] phpMyAdmin. http://www.phpmyadmin.net/