CLOUD STORAGE AND SEARCH FOR MASS SPATIO-TEMPORAL DATA THROUGH PROXMOX VE AND ELASTICSEARCH CLUSTER

Yicheng Zheng[1], Feng Deng[1,2], Qingmeng Zhu[1], Yong Deng[1]

[1]Science and Technology on Integrated Information System Laboratory
Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
dengfeng19901003@126.com

Abstract: Cloud computing is currently becoming a popular topic in recent years. The innovative application of cloud computing emerges endlessly. In this paper, the cloud computing platform is architected by virtualization tool Proxmox VE and the open source search engine Elasticsearch under the concept of virtualization. Based on this platform, we explore the feasibility and advancement for storing and searching spatio-temporal data. The time period index is imported as spatio-temporal data records both time and location. In this paper, the experiment is conducted using AIS data, which contains vessel motion characteristics. The result shows that the combination between virtualization and Elasticsearch can effectively store and index the spatio-temporal data with high reliability and efficiency. This paper is an innovation in the application of cloud virtualization. The method can widely and strongly support further research based on mass spatio-temporal data.

Keywords: cloud computing platform; virtualization; Proxmox VE; Elasticsearch; spatio-temporal data

1 Introduction

Cloud computing makes great advancement and innovation in IT industry [1]. The cloud computing has defined a new style of computing strategy in which resources are easily virtualized and provides scalable services that can be accessed over the network[2]. And resources are delivered as users' demand.

Virtualization is a crucial component in cloud computing system. Cloud computing system can be considered as a virtual resource pool. By developing multiple virtual machines in the servers, the total computing and storage capability can be portioned to different tasks [3]. Thus the efficiency of the system increases. And the stability is guaranteed by migrating virtual machines from one server to another. There existing multiple kinds of virtual machines, such as kernel based virtual machine (KVM), Xen, VMware workstation etc. Here, the research and experiment are based on one kind of KVMs, Proxmox VE. The specific technology review will be presented in the following session.

The application of virtualization is widely used in cloud computing. Lucas Nussbaum [4] constructs High Performance Computing (HPC) clusters through Xen and KVM. Then compare their capabilities in different conditions. Virtualization technology is used in many data centers to provide services through internet. In the Internet Data Center (IDC) of Tencent, virtualization improves the efficiency of servers. In the data center of BMW Group, virtualization helps reduce the electricity consumption by 70%.

In this paper, we mainly focus on the storage and search of spatio-temporal data based on the platform constructed by Proxmox VE, an environment for virtualization and Elasticsearch, an open source engine for text data storage and searching based on Lucene. Spatio-temporal data is a dataset that records geometries changing over time. The traditional method to process spatio-temporal data is trough rational, objected-oriented database such as oracle [6] and distributed Hbase [7]. With mass spatio-temporal data, traditional oracle database cannot handle the large sale of the data and the searching efficiency decreasing exponentially. Though the distributed Hbase can store mass spatio-temporal data, it does not provide rapid and characteristic index. Based on the virtualization technology, the combination between Proxmox VE and Elasticsearch is an innovative solution to the problem, with high performance efficiency, high reliability, enough capability and effective searching index.

Experiment is conducted with AIS data. AIS is short for the Automatic Information System, which is compulsorily equipped in vessels by International Maritime Organization (IMO). AIS data is the information broadcasted by the system. AIS data contains much information about vessel motion, such as longitude, latitude, timestamp, speed, course over ground etc. It is obviously one kind of spatio-temporal data.

The structure of the paper is as follows: In the second session of the paper, the technology adopted in the article is introduced. In the third session of the paper, the architecture of the cloud computing platform is proposed. In the fourth session of the paper, the solution for storing and searching AIS data is presented. In the fifth session of the paper, the evaluation about the system is provided. The reliability and effectiveness of the method is

978-1-4799-4719-5 /14/$31.00 ©2014 IEEE

confirmed. In the sixth session of the paper, the conclusion and further work is raised.

2 Technology review

The technology and concept involved in the architect of the paper is as follows:

2.1 The infrastructure as a service(IAAS)

The infrastructure as a service is one of the most important three concepts in could computing. The other two is platform as a service (PAAS) and software as a service (SAAS). In the application of IAAS, consumers can get service of infrastructure through internet. In this paper, the architecture of the platform is exact an example of IAAS. Users acquire service of infrastructure to store and search spatio-temporal data.

2.2 Proxmox VE

Proxmox VE is short for Proxmox Virtual Environment. It is a GPL-licensed, open-source virtualization platform for virtual machines management [8]. Proxmox supports creating and managing both container-based virtualization with Open VZ and full virtualization with KVM following with a friendly web interface (as shown in Figure 1). The system possesses features like backup, restore and live migration.

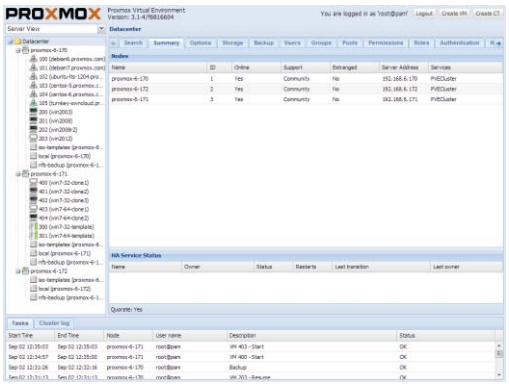


Figure 1 Web interface of Proxmox VE

Though Proxmox VE is an environment integrated with Open VZ, it supports direct installing in the computer, which means the virtual machine can share more resources provided by the computer. And this is one of the main benefits from virtualization.

2.2 Elasticsearch

Elasticsearch is an open source search engine which can provide distributed and real time searching capabilities. It is known as a document database for implementing Lucene as backend for document parsing and structuring [9]. Elasticsearch has following features:

Distributed: Elasticsearch server can start with single node and can be scaled horizontally depending upon concrete requirements. If more capacity is needed, just add more nodes.

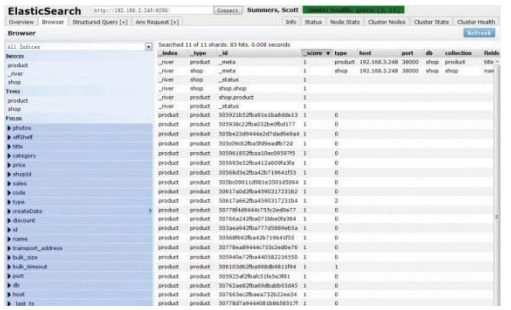
High Availability: Some Elasticsearch can form Elasticsearch cluster. The cluster is error resilient. If any error is detected, it will automatically remove the failed nodes and re-organize itself to make sure data is safe and accessible.

Full Text Search: It provides full query based search capabilities using Lucene.

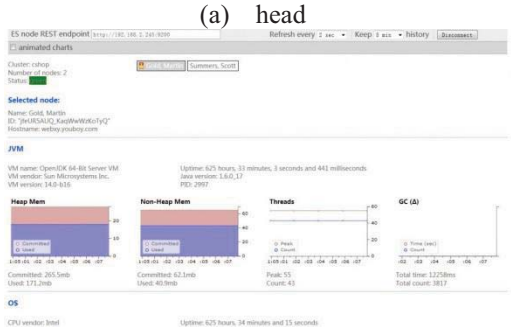
Document Oriented: Elasticsearch stores data or documents in JSON format. All documents are indexed by default with providing result at very fast speed.

Schema-free: Documents can be easily stored in JSON format. Elasticsearch will automatically detect the data structure, data types. And index the data accordingly. User can also define its own mapping and can change if required. Documents are versioned for any changes and provide the conflict management automatically.

Several computers using Elasticsearch engine can form Elasticsearch cluster. Elasticsearch cluster can be easily managed by plugs head and bigdeskbigdesk through web interface (as shown in Figure 2).



(a) head



(b) bigdeskbigdesk

Figure 2 Plugs for managing Elasticsearch through web interface

3 The architecture of the storage and searching system

This session of the paper describes the architecture for the combination of Proxmox VE and Elasticsearch. For saving cost, the powerful servers are simulated by usual PCs. The cloud platform is constructed by Proxmox VE in three same PCs. The hardware configuration is as in table I. Then five virtual machines (shown in table I) are created in the platform to provide Elasticsearch service. These machines form Elasticsearch cluster collaboratively. The structure and application of the platform is the typical reflection of the IAAS concept. Besides, two PCs serving as clients are connected to the

platform through Gigabit Router. Some developing and testing work will be carried out in the client. Operators can also surveil the statement of the Elasticsearch cluster in the clients through plugs head and bigdeskbigdesk. The hardware configuration for clients are shown in Table I as well.

Table I The hardware environment for cloud computing

Device type	Hardware Configuration	Usage
Servers	CPU：Intel(R) Core(R) i7-3770,4 CoresMemory:32GB Hard disk：1TB	Serving as physical notes, running Proxmox virtual environment
Elasticsearch virtual machines	CPU：Intel(R) Core(R) i7-3770,2 CoresMemory:8GB Hard disk：200GB	Building Elasticsearch cluster, providing storage and searching service
Clients	CPU：Intel(R) Core(R) i3-2120,2 Cores Memory:8GB Hard disk：200GB	Developing and testing for the cluster, monitoring running conditions of the platform
Router	Gigabit	Connecting every nodes

The running environment for the cloud computing platform constructed by PC, KVM and search engine can be illustrated in Figure 3.

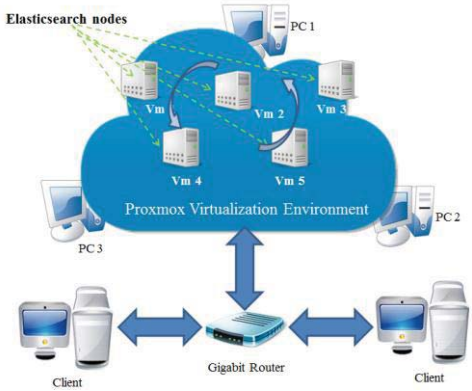


Figure 3 Illustration of the platform operator

Different version of software often has distinct effectiveness to the cloud computing platform. The configuration of software in the platform is shown in Table II.

Table II software environment for cloud computing platform

Software	Description	Version
Proxmox	Managing VE and physical resources	V3.1-3
Cent OS	Upon the KVM, providing service and running environment	V6.5
Elasticsearch	Managing provide service for data storage and searching	V0.90.10
JDK	Providing java environment	V1.7.0_45 （64bit）VM

Under the support of the hardware and the assistance of open source framework, the platform of cloud computing for spatio-temporal storage and searching is established. The structure of the platform is hierarchical as follows: the bottommost layer of the platform is three physical PCs. They are the foundation of the cloud computing platform. Upon them are the Proxmox VE, the Openvz based KVM. Proxmox VE organizes the capability of the PC. Virtual machines can be created and managed on the Proxmox VE through web interface. Thus Center OS is installed as virtual machine. It forms the OS environment that Elasticsearch needed. The topmost layer of the platform is Elasticsearch, which directly interacts with data. The platform architecture motioned in the paper is illustrated in Figure 4.

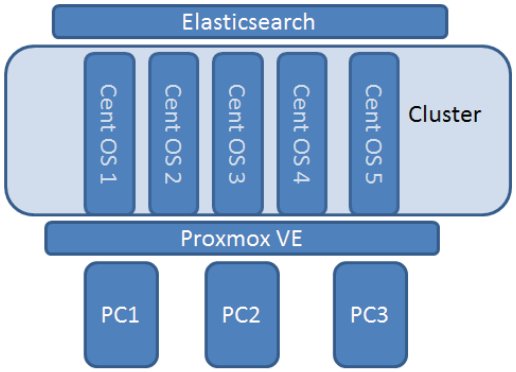


Figure 4 Illustration of the architecture

Spatio-temporal data can be imported to the platform by invoking API of Elastcsearch. As one kinds of Spatio-temporal data, the storage and searching for AIS data will be introduced in the next session.

4 Storage and searching AIS data

This session take AIS data as example to give solution to the management of Spatio-temporal data.

4.1 AIS data

AIS data describes vessel motion in multiple extensions, including Maritime Mobile Service Identify (MMSI), longitude, latitude, position time, flag, ship type, etc. The main attributes and explanations of the AIS data are

shown in Table III.

Table III Attributes and Explanations of AIS data

Attributes	Explanations
MMSI	Maritime Mobile Service Identify of ships
SHIPTYPE	Type of ships
SHIPNAME	Name of ships
FLAG	Flag of ships
COURSE	Course over ground of ships
LONGTITUDE	Longitude of ships
LATITUDE	Latitude of ships
SPEED	Speed of ships
POSITION TIME	Position time of the record
SIZE_LENGTH:	The length of the ship
SIZE_WIDTH	The width of the ship

4.1 Data storage

JSON format data can be applied for storing the data in the server using REST API which is encapsulated in the Elasticsearch engine. The following Figure 5 shows the storage example of AIS data.


```
{
_index: ais
_type: ship_dynamic_info
_id: K7UMzWo3SvKA0tZ8iqLDDQ
_score: 0.5946992
_source: {
COURSE: 307.6
POSITION: 31.6828,121.189
SPEED: 8.99979
POSITIONTIME: 2013-11-02 11:01
MMSI: 413763766
SHIPNAME: YU JIE XIANG 865
FLAG: CHINA
SHIPTYPE: CARGO SHIP
SIZE_LENGTH: 48
SIZE_WIDTH: 8
}
}
```


Figure 5 The storage of AIS data in Json format

As shown in figure 5, the storage of AIS information is accompanied with the creation of index, type, id, and score. In the example of Figure 5, index "ais" and type "ship_dynamic_info" is affirmed by the operators. Id "K7UMzWo3SvKA0tZ8iqLDDQ" and score "0.5946992" is generated by Elasticsearch randomly, using for distinguishing different records.

4.2 Data searching

The Elasticsearch creates index itself when data imports. For improving its searching responding speed, new storage structure which splits the raw data into different month is introduced when importing the data into the cluster. Under this storage structure, timestamp of searching query is to be detected before submitting to the cluster. Thus the searching operation is constrained in a relative limited extent when the result is returned. The operation structure and progress of searching are shown in Figure 6.

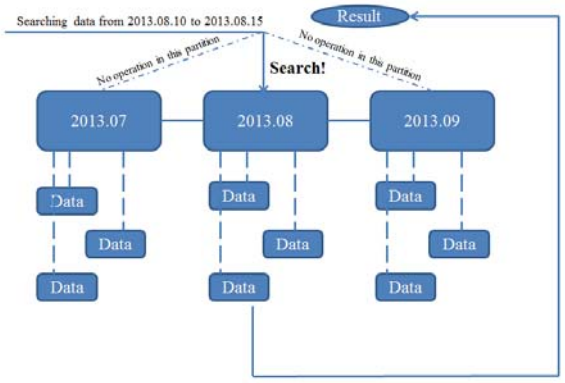


Figure 6 Process of searching data

Figure 6 gives an example that searching data from "2013.08.10 to 2013.08.15". At the beginning of the progress, the timestamp is detected that the searching operation occurs only in august. Then the query is submitted into the partition "2013.08", where the searching and returning executes. If the data wanted step over multiple months, the searching and returning operation will be carried out in multiple months accordingly.

The searching operation under storage according to months reduces the searching scale for every query for data. The efficiency for execution is improved and responding time is reduced obviously.

5 Evaluation

The evaluation of the platform focuses on theses aspects: reliability, responding time, and concurrence.

5.1 Reliability

Every cluster has a host node, in physical or logical. Evaluation for reliability aims for testing the effect when the host nodes shut down. In this paper, the IP addresses of the virtual machines range from 192.168.10.240 to 192.168.10.244. The initial host node is attached by IP 192.168.10.243. The evaluation process begins with the shutting down of initial node.

During the process of Reliability evaluation, the host node is shut down one by one. The migration of host virtual machine is witnessed. And the data stored in the stopped machine transformed into others.

5.2 Responding time

Responding time is the time used for returning results. It depends on the scope of searching query. The illustration of the Responding time is shown in the Figure 7.

Responding time in Figure 7 is the average value after many tests. The returning amounts of the records in third query and fourth query are nearly the same. But the third query which responding time is 0.39s occurs in one month. But the fourth query which responding time is 0.453s occurs in both July and August, which are multiple months. It presents that searching data in more

than one month or more than one time period that is indexed consumed more time. The last searching operation also across the July and August, but with a relative more returning data, the time consuming increases too. It presents the fact that the responding time is positive correlation with the scale of the results.

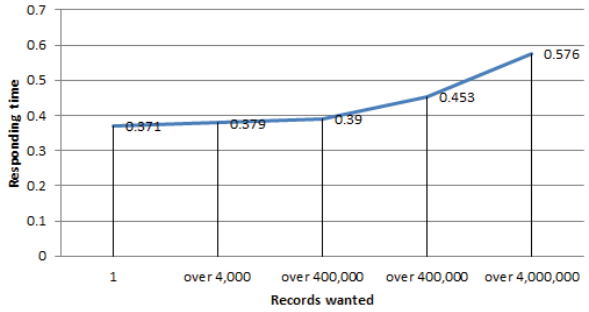


Figure 7 Illustration of responding time

5.3 Concurrence

Concurrence is the time consuming under multiple accesses to the cluster. In this indicator, the searching operation is limited in one time period, which means in the situation of this paper. It occurs in one month. In order to control the result scale, returning less than 100,000 is considered effectiveness. The concurrence results are illustrated in the Figure 8.

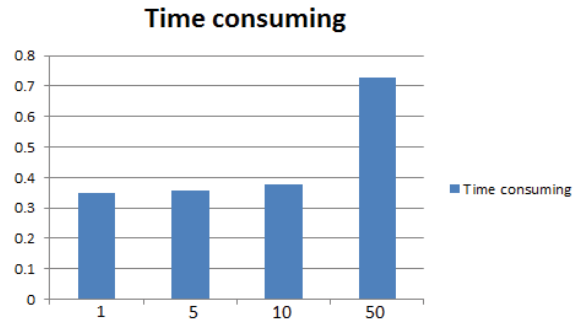


Figure 8 Time consuming for different concurrence

Figure 8 shows the time consuming for the different concurrence. Results show that more concurrences consumes more time. But in concurrence 50 that is not a low indicator, the consuming time still limits in one second, which indicates the high efficiency of the platform in the architecture presented in this paper.

6 Conclusions

This paper proposes a new architecture in cloud computing platform for storing and searching spatio-temporal data. AIS data are used to conduct the experiment to validate the feasibility and efficiency for the architecture. And the result is positive.

However there still existing much research points in the future work. The security of the data and platform is deserved consideration. And the experiment for comparison is needed to make the method more Persuasive.

Acknowledgements

This work was supported by Beijing Municipal Science Foundation No. 4133092 and 863 program (No.2012AA011206).

References

[1] Borko Furht, Armando Escalante. Handbook of Cloud Computing. Springer, Cambridge.

[2] Singh A. Cloud search. Dell Cloud service application, 2013.

[3] Armbrust M, Fox A, Griffith R et al. A view of cloud computing. Communications of the ACM, 2010, 53(4): 50-58.

[4] Nussbaum L, Anhalt F, Mornard O et al. Linux-based virtualization for HPC clusters. Montreal Linux Symposium, 2009.

[5] Qiu O Z, Yue Z. Research on application of virtualization in network technology course. Computer Science & Education (ICCSE), 2012 7th International Conference on. IEEE, 2012: 357-359.

[6] Parent C, Spaccapietra S, Zimányi E. Spatio-temporal conceptual models: data structures+ space+ time. Proceedings of the 7th ACM international symposium on Advances in geographic information systems, ACM, 1999: 26-33.

[7] Tan H, Luo W, Ni L M. CloST: a Hadoop-based storage system for big spatio-temporal data analytics. Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012: 2139-2143.

[8] Li P. Centralized and decentralized lab approaches based on different virtualization models. Journal of Computing Sciences in Colleges, 2010, 26(2): 263-269.

[9] Marek Rogozinski, RafaL Kuc. Elasticsearch Server. packtpub, Birmingham B3 2PB, UK, 2 2013.