Design and Implementation of Big Data Training Project Platform Based on Hyper-converged Architecture

Jun Zheng[1], Yu Bai[2], and Wei Huang[3](✉)

[1] Guizhou Police College, Guiyang, China
[2] College of Mathematics and Information Science, Guiyang University, Guiyang, China
[3] Department of Information Engineering, Guizhou Light Industry Polytechnic, Guiyang, China
39189550@qq.com

Abstract. Building high quality big data training courses is the necessary foundation and urgent need for training big data talents, and it is also an urgent task at present. In this study, hyper-converged architecture technology is used to design and implement a big data training project platform for college students, providing different simulation scenarios for big data training. The system software architecture design adopts the principle of loose coupling and hierarchical design, from bottom to top, it is mainly divided into data acquisition layer, physical layer, device virtualization layer, data layer, business processing layer, service layer, User Interface layer. The platform deploys private cloud OpenStack on CentOS7. The big data analysis process and technology are the data acquisition module (Scrapy and NoSQL), data cleaning module (Kettle), data mining and analysis module (Python scripts), and data visualization module (Power BI). The big data training project platform based on Openstack provides reference for the construction of related platforms.

Keywords: Hyper-converged · Big data · Training courses · Teaching research · Education and teaching reform

1 Introduction

In February 2022, the project of "counting in the East and counting in the West" was officially launched. As one of the eight national hub nodes approved to build a national integrated computing network, Guizhou will also deeply participate in this national super project. At present, Guizhou is speeding up the construction of the first national big data comprehensive pilot zone and digital economy development innovation zone, and 37 key data centers have been fully put into operation or are under construction. Up to now, a total of 18 large and ultra-large data centers have been established in Guizhou Province, among which 8 are ultra-large data centers, making it one of the regions with the most large and ultra-large data centers in the world. The server carrying capacity of Guizhou Province reached 2.25 million units, with an average on-set rate of 56.5%. The backbone networks of 16 provinces and 32 cities in China are directly connected to each other.

F. Huang et al. (Eds.): ICAIE 2023, AHCS 15, pp. 598–603, 2024.
https://doi.org/10.2991/978-94-6463-242-2_73

Big data technology is developing rapidly in data analysis, transaction processing and other fields, and the architecture system of big data technology has formed a pattern dominated by open-source technology. With the decrease of hardware equipment price and energy consumption and the rapid development of large-scale artificial intelligence model, the demand for computing performance increases rapidly, and the continuous reduction of delay requirements, which promotes the development pattern of large and ultra-large data centers.

The rapid development of big data technology has brought about a sharp increase in the demand for talents. The construction of high quality training courses of big data is the necessary foundation and urgent need of training big data talents, and is also an urgent task at present. Big data training is closely integrated with the industry, and there are different analysis data and analysis processes for different industrial scenarios and needs. For teachers, it is difficult to realize big data training courses in different scenarios. Therefore, it is necessary and urgent for a standard professional big data training platform to support big data teaching.

2 Relevant Technology

Big data has a large amount of data, a variety of data types, and complex data sources. Therefore, even if the construction of big data training tasks on a single machine can meet the analysis of teaching scenarios, it does not conform to the real enterprise working environment and lacks the understanding of operation and maintenance. The direct use of the supercomputer center is difficult to meet the needs of 60 students to conduct concurrent experiments at the same time. Therefore, it is necessary to build virtual machines (VMS) with computing and storage resources in the supercomputer center through virtualization platform software to reasonably schedule computing and storage resources to meet the needs of 60 students for concurrent training in different application scenarios.

The hyper-converged OpenStack cloud computing platform uses the distributed storage mechanism to reduce storage device overhead (Melo et al. 2018), which can save cost for the construction of simulation platform. The hyper-converged architecture achieves efficient hardware integration, flexible resource scheduling, rapid system deployment, secure and reliable data, and unified service operation and maintenance (Zhang et al. 2021), which allows for rapid and flexible system deployment. The hyper-converged system consists of distributed video collection, hierarchical video storage and read/write load balancing (Qiang et al. 2020), which allows flexibility in building course video content.

The hyper-converged cloud platform adopts the open-source cloud computing solution OpenStack (Lima et al. 2019; Sahasrabudhe and Sonawani 2014), which has a modular architecture (Lima et al. 2019), can manage large pools of resources throughout the data center (Chi et al. 2018; Kai et al. 2020; Parakh et al. 2018; Sharma and Joshi 2016; Wang et al. 2022), including compute, storage, and network resources. Therefore, it is possible to build the big data training project platform for college students based on OpenStack. The platform deploys private cloud OpenStack on CentOS7.

3 System Design

3.1 The Architecture Design of the Platform

The software architecture design of the big data training project platform adopts the principle of loose coupling and hierarchical design. From bottom to top, it is mainly divided into data acquisition layer, physical layer, device virtualization layer, data layer, business processing layer, service layer, and User Interface (UI) layer. The data acquisition layer is the interface layer, which accesses external data collection for the entire simulation platform, including the Internet and the Internet of Things. The physical layer provides basic physical facilities for the big data training project and is the physical basis of the whole big data training project platform. It has servers, gigabit Fibre Channel switches, 10-gigabit Fibre Channel switches, and UPS power supplies. In order to accommodate 60 students to use virtual machines concurrently, each virtual machine is configured with 2 CPUs, 32 GB memory, 500 GB hard disk, and one graphics card. Therefore, 120 CPUs, 1600 GB memory, 30 TB hard disk, and 60 graphics cards are required. For this purpose, 60 2U rack servers are configured, each server has two Intel Xeon Silver 4310 processors, 32 GB DDR4–2933 ECC REG RDIMM memory, a 2.5-inch 480GB SATA SDD disk, three 2.5-inch 10 TB SATA disks and one NVIDIA Geforce RTX3090 24G PCI-e turbo GPU card. At the device virtualization layer, a hyper-converged management platform is deployed to provide hardware virtualization and resource scheduling for the entire big data training project platform, including CPU virtualization, storage virtualization, and graphics card virtualization. In order to meet the different training requirements of practical training projects, a variety of database systems are deployed at the data layer, including distributed database and stand-alone database, which provide interfaces for data access to the database and are directly invoked by the business processing layer or the service layer to provide data management and data sharing for the big data practical training project platform. Because student data and teacher data contain sensitive personal privacy data, user access needs to be verified, and the data is encrypted by the national secret algorithm and stored in the database. The business processing layer refers to the common business processing component, which is mainly used to process the common business processing rules and processes of the service layer, including the foreground display, task release, information browsing, query statistics and event approval. It is implemented by Java objects. The service layer is the function of big data practical training project platform to provide external business services, including task publishing service, information browsing service, query statistics service and event approval service. This layer separates the service interface from the implementation to achieve loose coupling, which is mainly implemented by I Service interface. The UI layer is the user interface of the big data practical training project platform and the entrance for users to access the big data practical training project platform. Students, teachers and administrators can all access big data practical training project resources on this layer (Fig. 1).

3.2 The Functional Design of the Platform

The purpose of the big data practical training project platform is to build a comprehensive big data practical training project platform covering a variety of application scenarios

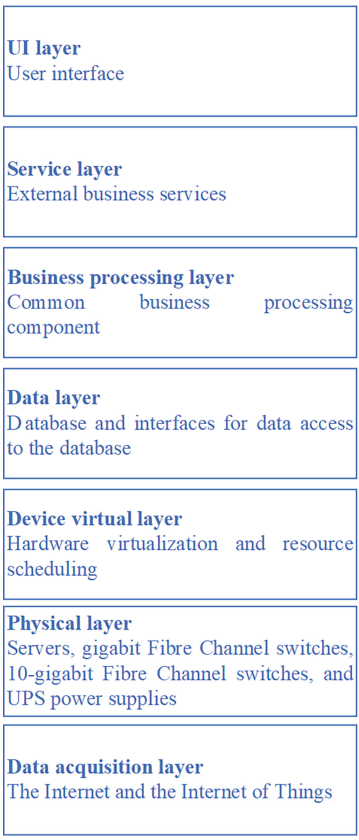

Fig. 1. Architecture of big data practical training project platform

(Fig. 2), including the whole process of data collection, data cleaning, data mining and analysis, and data visualization. The big data analysis process and technology are the data acquisition module (Scrapy and NoSQL), data cleaning module (Kettle), data mining and analysis module (Python scripts), and data visualization module (Power BI). The big data training project platform has three kinds of users with permissions. The first kind is the super manager user, which is mainly responsible for the operation and maintenance of the whole big data training project platform. The second is the teacher user, which can conduct teaching-related management work, including the creation of experimental environment, the allocation of student accounts, the management of teaching data, and the modification of experimental manuals; The third category is student users, who can carry out practical training activities, including viewing experimental manuals, selecting interesting experimental scenarios, and writing and submitting experimental reports. Under the constraints of the platform architecture of the big data practical training project platform, the big data practical training project platform mainly includes data acquisition module, data cleaning module, data mining and analysis module, data visualization module and teaching management module. At present, the big data training project

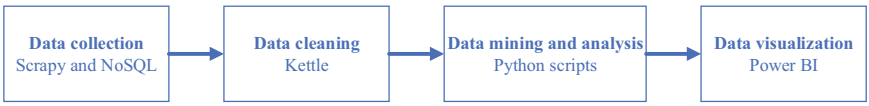

Fig. 2. Big data processing process and technology

platform has built a total of 5 practical training scenario projects, including student portrait decision analysis project, medical treatment decision analysis project, traffic accident cause analysis project, e-commerce red wine decision analysis project, and government and civil situation decision analysis project.

4 Conclusion

Through the construction of big data practical training project platform, it provides practical training close to the actual scene for students majoring in computer related majors, cultivates their system integration ability, data collection ability, data cleaning ability, data mining and analysis ability and data visualization ability, and cultivates their team cooperation ability, so that their big data skills have been comprehensively trained. Also understand the characteristics of different scene data and the corresponding big data technology, which is better for students to adapt to the job.

Acknowledgment. This work was supported by Guiyang University 2021 university-level construction project of new liberal arts, new engineering and new agriculture; "Construction of Integrated Curriculum and Teaching Material System for New Engineering Majors" under grant number 0221003005136.

References

Chi Y, Li G, Chen Y, Fan X Design and Implementation of OpenStack Cloud Platform Identity Management Scheme. In: 2018 International Conference on Computer, Information and Telecommunication Systems (CITS), 11–13 July 2018 2018. pp 1–5. https://doi.org/10.1109/CITS.2018.8440198

Kai Z, Youyu L, Qi L, Hao SC, Liping Z Building a private cloud platform based on open source software OpenStack. In: 2020 International Conference on Big Data and Social Sciences (ICBDSS), 14–16 Aug. 2020 2020. pp 84–87. https://doi.org/10.1109/ICBDSS51270.2020.00027

Lima S, Rocha Á, Roque L (2019) An overview of OpenStack architecture: a message queuing services node Cluster Computing 22:7087-7098 https://doi.org/10.1007/s10586-017-1034-x

Melo C et al. Availability models for hyper-converged cloud computing infrastructures. In: 2018 Annual IEEE International Systems Conference (SysCon), 23–26 April 2018 2018. pp 1–7. https://doi.org/10.1109/SYSCON.2018.8369580

Parakh P, Narayan DG, Mulla MM, Baligar VP SLA-aware Virtual Machine Scheduling in OpenStack-based Private Cloud. In: 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 20–22 Dec. 2018 2018. pp 259–264. https://doi.org/10.1109/CSITSS.2018.8768760

Qiang W, Bo S, ying QZ, dong Lg, Fan D A Hyper-Converged Video Management System Based on Object Storage. In: 2020 12th International Conference on Advanced Infocomm Technology (ICAIT), 23–25 Nov. 2020 2020. pp 74–79. https://doi.org/10.1109/ICAIT51223.2020.9315468

Sahasrabudhe SS, Sonawani SS Comparing openstack and VMware. In: 2014 International Conference on Advances in Electronics Computers and Communications, 10–11 Oct. 2014 2014. pp 1–4. https://doi.org/10.1109/ICAECC.2014.7002392

Sharma MA, Joshi MO Openstack Ceilometer Data Analytics & Predictions. In: 2016 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 19–21 Oct. 2016 2016. pp 182–183. https://doi.org/10.1109/CCEM.2016.045

Wang H, Zhang X, Ma Z, Li L, Gao J An Microservices-Based OpenStack Monitoring System. In: 2022 11th International Conference on Educational and Information Technology (ICEIT), 6–8 Jan. 2022 2022. pp 232–236. https://doi.org/10.1109/ICEIT54416.2022.9690713

Zhang Y, Ren J, Liu F, Wang Z, Song Y, Yin L, Peng Y A Novel Hyper-Converged Architecture for Power Data Centers. In: 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 29–31 July 2021 2021. pp 391–394. https://doi.org/10.1109/ICPICS52425.2021.9524221