ATLANTIS PRESS

Advances in Computer Science Research, volume 75

7th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2017)

Novel Application of DaaS and Hadoop Technology in Big Data Cloud Computing Platform

Hongsheng Xu[1,2 a *], Ganglong Fan[1,2] and Ke Li[1,2]

[1]Luoyang Normal University, Luoyang, 471934, China

[2]Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang, 471934, China

[a]85660190@qq.com

Keywords: Cloud computing; DaaS; Hadoop;Big data; Software framework

Abstract. DaaS is to dig out the potential value of big data and provide services according to the needs of users. Hadoop is a software framework for distributed processing of large amounts of data, and in a reliable, efficient, scalable processing way, relying on the horizontal expansion, to improve the computing and storage capacity by increasing the cheap commercial servers. The paper presents novel application of DaaS and Hadoop technology in big data cloud computing platform. Users can easily in the application on the development and operation of big data processing.

Introduction

The next ten years will be a "big data" to lead the wisdom of the era of science and technology. As the social network gradually mature, mobile bandwidth increased rapidly, cloud computing, networking applications more abundant. Sensing equipment more, mobile terminal access to the network data and the resulting growth rate will be more than any other time in history there are more, faster.

Data acquisition technology is a necessary condition for data processing; we need to have data acquisition means, the information collected, the data processing technology to the upper application of data acquisition. In addition to the various types of sensing equipment such as hardware and software facilities, mainly related to the ETL (data collection, conversion and loading) process, cleaning of. For data filtering, calibration, conversion and other pretreatment, and it is the effective conversion of data into the format and type of suitable [1]. At the same time, in order to collect and store the data access support heterogeneous, but also must design the enterprise data bus, data exchange and sharing between the various enterprises.. Convenient is application and service.

The data service should contain multiple meanings. First of all, the public can provide data access service for the user, the user can access any content data. For example, a user wants to check the weather conditions over the past ten years, the weather data service providers can provide users with the past ten years. For this kind of data service can be provided according to different countries and regions, in the quarter, given the data. So, public access to data is flexibility, multi angle, full range.

Information should be placed in priority with more confidence, more secure network. Including monitoring means, the era of big data: sniffer monitor, on submarine cable transit directly tapped at the receiving end; routing hijacking, currently 9 global router 13 top-level domain name in the server in the United States, they control 70% of the world data; network intrusion, break through the password to tamper with the password, it's with someone's house keys directly to get something.

Hadoop big data framework based on cloud computing, using the power of cluster computing and high-speed storage, realizes a distributed operating system, providing high transmission rate to access the data in the form of flow, adapt to the application of big data [2]. Moreover, data mining, semantic engine development, visual analysis technology, can to extract information from the mass data in depth analysis, control, data value-added "accelerator".

Data mining is generally not what the predefined theme, mainly in the existing data of various algorithms based on the above calculation, so as to forecast (Predict) effect, so as to realize the analysis of some high level data needs. A typical algorithm for clustering Kmeans, and it is for statistical learning

for the classification of SVM and NaiveBayes. The main is use of the tool Hadoop Mahout. The characteristics and challenges of the process is mainly used for mining algorithm is very complex, and involves the calculation of the amount of data and computing are large, commonly used data mining algorithms are based on a single thread.

Big data to cloud data center has a large physical resources and efficient scheduling management function of the cloud computing platform. Cloud computing management support platform can provide flexible and efficient deployment for large data center and enterprise operation and management of the environment, through the underlying hardware and operating system virtualization technology to support heterogeneous application, to provide safety, high performance, high scalability, high reliability and scalability of cloud resource management solutions, reduce application system development, deployment, operation and maintenance costs, improve the efficiency of resource use. The paper presents novel application of DaaS and Hadoop technology in big data cloud computing platform.

The combination of DaaS and intelligent decision service in big data

With the wide application of cloud computing, the system construction is bound to influence, thereby affecting the operation mode and development system of the whole business system and electronic commerce technology [3]. Based on cloud database relational database service will be the main development direction of cloud database, cloud database (CloudDB), provides the ability of parallel processing of massive data and good scalability and other characteristics, and provide support in the online analysis processing (OLAP) and online transaction processing (OLTP) database provides cloud service ability, superior performance, and become an ideal platform for cluster environment and cloud computing environment. It is a highly scalable, secure and fault-tolerant software, customers can reduce the cost of IT through the integration of multiple data management in business, improve the performance of decision service for all applications and real-time make better.

DaaS is the twin brother of SaaS, as one of the "as a service" family members, it will provide data as a commodity to any organization or individual needs of SOA (service oriented architecture, service oriented architecture) is a business driven, coarse-grained, loosely coupled architecture services, support integration on the business, make it become a kind of connection, reusable business tasks or services, is the most effective method to implement DaaS, as is shown by equation (1) [4].

$$c(\alpha) = \frac{\rho'\theta'' - \rho''\theta'}{((\rho')^2 + (\theta')^2)^{\frac{3}{2}}} \tag{1}$$

Cloud computing technology is the most ideal solution. The survey shows: at present, IT professionals of cloud computing in many key technologies are most concerned about the large-scale data parallel processing technology of large data parallel processing and general no ready-made solution for the application of industry, cloud computing platform software, virtualization software does not need to own the development of large-scale data processing applications, but the industry has no ready-made and generic software need to be developed specifically for the specific application requirements, involving many parallel algorithms, query optimization technology research and design of index system, provides the driving force for the development of large data processing technology.

The data source is also called real-time uninterrupted data stream. The stream data refers to the data as a data stream in the form of processing [5]. The data stream is distributed in time and the number of a series of infinite collection of data records; data recording is the smallest unit of data flow. For example, the field data generated by the sensor may be the Ever fount in networking for stream processing system. We will separate the details in the next section. Analysis of statistics and analysis of data dynamic and real-time calculation and real-time data for the monitoring system, dispatching management has important practical significance.

Such a cloud database to be able to meet: A. data processing: analysis of operating system similar to the search engine and telecom operators level such a large-scale application, need to be able to handle PB level data, at the same time to deal with millions of traffic. B. cluster management: large-scale

distributed applications can be easily deployed application and management low latency.C. Read and write speed: fast response speed can greatly improve the user satisfaction of it. D. construction and operation cost: basic requirements for cloud computing applications is that in the cost of hardware, software cost and manpower cost are greatly reduced [6].

DaaS solutions can provide the following advantages: agility. Through the integration of data access, customers can quickly move on it, and no longer need to consider the source of the underlying data. If the customer needs a slightly different data structure or call the specific location of the data, DaaS by minimal change to meet very fast the demand, as is shown by equation(2).

$$x_{i+1}^{2}(t+1)=\left(1-d_{i}^{2}(t)\right)x_{i}^{2}(t)+\left(\frac{N^{2}(t)}{N^{3}(t)+N^{2}(t)}\right)s_{i}\alpha N^{1}(t) \tag{2}$$

Store unstructured data using the file system, and improve the backup and disaster recovery strategy, compared with the cluster + commercial database scheme before minicomputers expensive enterprise this set of economic benefits of big data solutions, not only the loss of performance, but also won the scalability. Before our solution at the early stage of design a data center, we must take into account the scalability after the implementation of the program. The usual method is to estimate the business volume and the amount of data in the next period of time, adding extra calculation unit (CPU) and storage, to always be prepared.

Now a variety of data sources, such as Internet companies: SNS, micro-blog, video website, e-commerce website; networking, mobile terminal equipment, goods, personal location, sensor data; China Unicom, mobile, telecommunications and Internet communications operators; astronomical telescope images, video data, meteorology the satellite image data [7]. These data have, you can through big data related technologies, such as analysis technology, storage technology, computing technology to explore the value of data, and provide services.

With the development of cloud computing, there has been a lot of cloud platform and distributed system. The model of cloud computing is a business model, is the essence of data processing technology. The data has become a valuable asset, as a saying goes: who owns the big data, which will have the cloud provides storage for future. Data assets are access and calculation. The inventory of assets,assets and it is so that the national governance, corporate decision-making, personal services is a kind of data service idea.

Application of DaaS and Hadoop Technology in Big Data Cloud Computing Platform

The lack of resource utilization, data spread to the entire enterprise IT system leads to complexity management continuously, this is a problem every CIO [note] has to face. The predicament in reality also promoted the development of technology, data service (Data-as-a-Service, DaaS) by the resource centralized management, to improve the efficiency of IT the performance of the system and the direction. So DaaS has many CIO favored in the past few years, the main technology which contains data virtualization, data integration, SOA, BPM and PaaS.

Data mining is generally not what the predefined theme, mainly in the existing data of various algorithms based on the above calculation, so as to predict the effect, so as to realize the analysis of some high level data needs [8]. A typical algorithm for clustering for K-Means and it is SVM and Naive for statistical learning Bayes classification, the main use of the tools are Hadoop Mahout. The characteristics and challenges of the process is mainly used for mining algorithm is very complex, and involves the calculation of the amount of data and computing are large, and commonly used data mining algorithms are based on a single thread.

Eucalyptus is trying to clone the AWS open source cloud computing platform, to achieve a similar Amazon EC2 function, used by computing cluster or workstation cluster to achieve flexibility, the use of cloud computing. It provides the interface compatibility with EC2 and S3 storage system. The application of these interfaces can interact directly with Eucalyptus, Xen[10] and KVM support virtual technology, as well as for system management and user settlement cloud management tool.Eucalyptus

consists of five main components, respectively for the cloud controller CLC, cloud storage service Walrus, cluster controller CC, SC storage controller and the node controller NC.Eucalyptus through the Agent to the management of computing resources, the component can collaborate with each other to provide cloud services [9].

Data service is any service and related data can occur in a centralized location, such as aggregation, data quality management, data cleaning, and then provide data to different systems and users, and no longer need to consider what these data from the data source, as is shown by equation(3).

$$\begin{cases} \mathrm{f}(i_1) = a_0 + a_1 i_1 + a_2 {i_1}^2 +, \cdots, a_{t-1} {i_1}^{t-1} \\ \mathrm{f}(i_2) = a_0 + a_1 i_2 + a_2 {i_2}^2 +, \cdots, a_{t-1} {i_2}^{t-1} \\ \cdots\cdots\cdots\cdots \\ \mathrm{f}(i_t) = a_0 + a_1 i_t + a_2 {i_t}^2 +, \cdots, a_{t-1} {i_t}^{t-1} \end{cases} \tag{3}$$

So the cloud database must use relevant technology support for cloud environments, such as data node dynamic stretching and hot swappable, provide multiple copies of the fault detection and fault tolerant mechanism and transfer mechanism of all data, SN (Share Nothing) architecture, management center, and other processing nodes on the connectivity of any node is connected to the whole work cloud system and task tracking, data compression technology to save disk space and reduce the disk IO time. Cloud database is the traditional route database upgrade and database application based on close to Xiang Yun, to better adapt to the cloud computing model, such as automated resource allocation management, virtualization support, high scalability, can play an important role in the future will it.

The MapReduce programming model is the heart of Hadoop, for parallel computing of large-scale data sets. It is this kind of programming mode, to achieve a large-scale expansion across a Hadoop cluster in hundreds or thousands of servers; HDFS distributed file system provides Hadoop processing platform for massive data storage based on NameNode, which provides service for file metadata, DataNode block storage file system.

Experiments and Analysis

Build a DaaS platform for customer needs, including the main elements include: data acquisition (Data acquisition): from any data source, such as data warehouse, email, portal, third party data source. The data management and standardization: manual or automatic data standard. Data aggregation: This is a quality control mechanism strong service and technology driven, not simply write a ETL program. 100 data service: Web service, extraction and report, to allow the end user to more easily consumption data.

The traditional storage for massive data processing, through the establishment of data center construction, including hardware and software system of large data warehouse and its supporting operation, equipment (including servers, storage, network equipment, etc.) more and more high-grade, data warehouse, OLAP and ETL, BI and other platforms more and more huge, but they need more and more investment, in the face of the growth rate of the data, more and more powerless, so based on the traditional data center construction technology, operation and promotion more and more difficult. In addition to the general use of the traditional database, the data warehouse and the BI tool to complete the processing and analysis of data mining [10].

HBase is built on HDFS, used to provide high reliability, high performance, column storage, scalable, real-time database system to read and write data storage can be loosely unstructured and semi-structured, it is a large data warehouse based on Hadoop, can be used for data extraction, transformation and loading (ETL) storage. Mass storage, query and analysis of data in the Hadoop; Pig is a large-scale data analysis platform based on Hadoop, can make SQL analysis of data requests into a series of optimized MapReduce algorithm provides a simple programming interface for operation and calculation of sea quantity data complex parallel.

Summary

The paper presents novel application of DaaS and Hadoop technology in big data cloud computing platform. The application of knowledge and skills, personnel, processes and technology platform is the essence of the DaaS strategy is the key requirements of it. DaaS data management is more centralized, so that more users do not need to pay attention to the underlying data, and fully focused on how to use these data. As a high-performance communication middleware of binary system, Avro provides data serialization function and RPC service between Hadoop platforms.

Acknowledgements

This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (17B520026).

References

[1] Abouzeid A, Bajda-Pawlikowski K, Abadi D, Silberschatz A, Rasin A. HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proc. of the VLDB Endowment, 2014,2(1):922-933.

[2] Hongsheng Xu, Ruiling Zhang. Novel Approach of Semantic Annotation by Fuzzy Ontology based on Variable Precision Rough Set and Concept Lattice, International Journal of Hybrid Information Technology Vol.9, No.4 (2016), pp. 25-40.

[3] Deelman E, Singh G, Su MH, Blythe J, Gil Y, Kesselman C, Mehta G, Vahi K, Berriman GB, Good J, Laity A, Jacob JC, Katz DS. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. Scientific Programming, 2015, 13(5): 219-237.

[4] Wang Shan, Wang Hui Ju, Xiong Qin, Zhou Xuan. Architecture big data: challenges, current situation and Prospect, Chinese Journal of computers, 2013,34 (10): 1741-1752.

[5] H.-s. XU, R.-l. ZHANG, "Semantic Annotation of Ontology by Using Rough Concept Lattice Isomorphic Model", International Journal of Hybrid Information Technology, Vol.8, No.2, 2015, pp.93-108.

[6] Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. Taverna: a tool for building and running workflows of services. Nucleic acids research, 2014, 34(suppl 3): W729-W732.

[7] Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G. Understanding big data. New York et al: McGraw-Hill, 2013.

[8] Zhao Y, Li Y, Tian W, Xue R. Scientific-Workflow-Management-as-a-Service in the Cloud, Cloud and Green Computing (CGC), 2012 Second International Conference on. IEEE, 2012: 97-104.

[9] Tao Xuejiao, Hu Xiaofeng, Liu Yang. Big data research, Journal of system simulation, 2013,25S:142-146.

[10] Labrinidis A, Jagadish HV. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.