The Design of Software Work flow Architecture Based on Cloud Environment

Han Bo[1, a], Xu Lingyu[2, b] and Wang Lei[2, c]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China;

[a] hanbo_321@163.com

Keywords: Cloud Computing; Work flow; Hadoop; Hadoop Distributed File System (HDFS).

Abstract. In order to deal with the large-scale data access and massive marine information to provide reliable real-time cloud computing services, put forward the concept of software work flow and construct a system based on cloud platform with software work flow. Software work flow engine is at the bottom of the whole system, and interacts with the Hadoop platform. The operation designed service flow parse and reconstruction of work flow algorithm for processing user request. Transparent interface provides services for the upper layer flow monitoring and resource management. The engine reduces the complexity of development; improve the scalability of the system. Users can access through Web terminal, customized software services, and real-time monitoring of the cloud platform. On this platform, massive data access, high concurrency and high density visit is also a normal state. Through the construction of the initial prototype system, the availability and efficiency of the platform system are proved.

Introduction

With the implementation of the "Digital Ocean" project, the continuous development of China's marine information work, and built a massive specialized marine environmental information database and the sets used in marine business information system. They provide a powerful information support and technical support for the development of the marine industry in China.

However these independent and similar to the internal structure of the system also exist the following problems: resource consumption, high operation cost; the traditional mode is difficult to adapt to the rapid demand for service deployment, lack of unified computing resources planning; business system stability and low reliability, system maintenance difficult.

Cloud computing technology is used to build the system of ocean environment information application service, to support the construction of marine environment information service of low cost and operating environment, improve the marine resource information can be reuse and sharing and application system scalability. At present, there are some cloud computing platform based on Hadoop technology, including the Blue Cloud of IBM, the cloud plan of Yahoo and Intel, the BigCloud of China Mobile and the cloud computing of Baidu and Alibaba. [1] Hadoop is the most widely used, is used to efficiently deal with massive amounts of data. Yarn is an upgraded version of Hadoop2.0, with the advantages of scalability, high reliability, low cost and high efficiency. [2] In this paper, we put forward the concept of work flow and give a detailed description of Work flow. Study on flow structure of cloud platform software based on the realization of personalized services, marine services flow customization, computing services for large-scale data processing and access to massive ocean of information to provide real-time and reliable cloud.

Software Work Flow

Software work flow [3] is a partly or fully automatic execution of business process. According to a series of rules, documents, information or tasks Software work flow can between different executors to transfer and implement. The two most basic elements in the work flow are the connection between the activity and the activity. Activities corresponding to the business process of the task, mainly reflect the business process of the implementation of the action or operation; the connection between the activities of the business process of the rules and business processes.

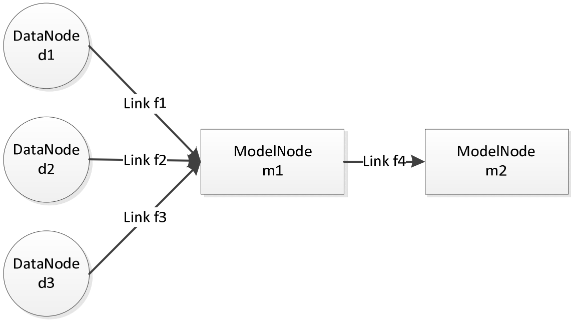


Fig. 1 A Software Work flow

In figure 1, we use two object entities for process modeling: nodes and the direction of the connecting link arcs. The nodes are divided into two types: data nodes and model nodes. [4]

Data node: the user through the resource management center to upload or share data, attributes including data name, owner, nature, data format, data address and data size, etc.

Model node: Model nodes are some features of the execution units. Through the resource management center upload or shared attributes include the model name, model matching format, owner, and address and so on.

Link arc: connect the data node and model node, representation model of nodes through their own specific methods performed on the data. The execution result is a data node.

Definition. A customized software work flow is three tuple (WF, a, R), in which:

WF = (D, M , F) is software work flow instance;

D = { d1 , d2 ,…, dn } is a collection of data nodes, where data node d represents the resources required for activities or events in the business process;

M = { m1 , m2 ,…, mn } is a collection of model nodes, where the model node m represents an activity or event in a business process;

F is a directed link arc between the data nodes and the model nodes, which indicates the relationship between the data nodes and the model nodes;

α is a collection of positive real numbers. Each element α(m) represents the execution time of the model m;

R represents the state of the entire service flow, each of which is submitted to the service flow through the software service flow engine to generate different state transitions through different execution steps.

Different from the traditional work flow, the characteristics of the software work flow is that the process unit is the cloud service. Because the execution of service in the cloud platform more complex and unpredictable, software work flow of the service abstraction is more complex, its automated execution for efficient, real-time, stability, error correction capability requirements are higher.

In general sense, the function of work flow system can be divided into three aspects:

1) Establishment stage function: mainly consider the process and function of related actions and the definition of work flow modeling

2) Running stage function: in a certain operating environment, the work flow process is executed, and the coordination of the activities of each process and the scheduling function is completed;

3) Man-machine interactive function in the running stage: to realize the interaction between the user and the IT application tool in the process of executing various activities;

The figure 2 shows the relationship between three main functions of work flow management system.

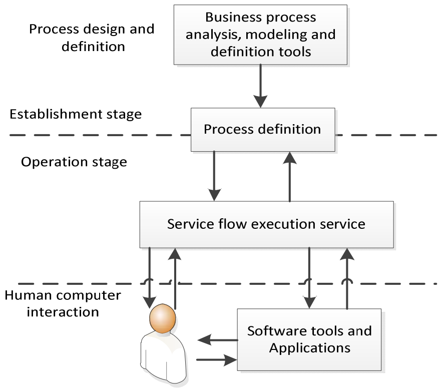

Process design and definition
Business process analysis, modeling and definition tools
Establishment stage
Process definition
Operation stage
Service flow execution service
Human computer interaction
Software tools and Applications

Fig.2 work flow management system

The cloud computing technology and work flow technology combined, originally distributed placed on the cloud platform implementation of complex executable program, the user to customize and customized user is parsed into an interconnected basic service unit consisting of a flow chart, matching the Spark [5] or Hadoop execution system.

Software Work flow Engine

Software work flow engine is responsible for controlling and coordinating the execution of various computational models. Users can upload their own models and data through the visualization platform if they do not find data and models they need in public data area. The inspection system on the platform of visual customization will complete model, to ensure the submission of service flow engine service flow is required and is legal. Software work flow engine analysis and reorganization received work flow, automatic recognition of spark or Hadoop Mapreduce [6] model, using different implementation strategies, monitoring system, delivered to the cloud platform to complete the model operation, and the operation result feedback platform.

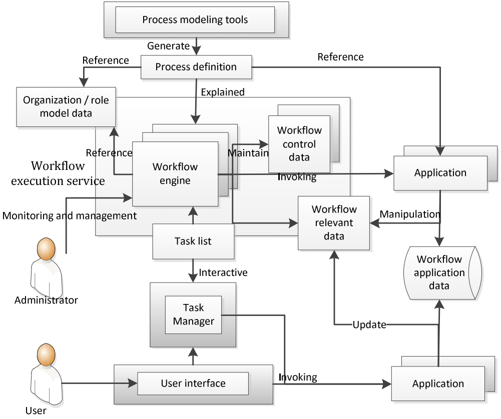


Fig.3 Software Work Flow Engine Model

Figure3 shows the Software work flow engine model. Software service flow is composed of service model of the set of nodes and node relation of a weighted directed acyclic graph DAG [7] (directed acyclic graph), nodes can expressed as a combination. Service model, pointed out the storage location service model distributed file system (HDFS), is the location for storing data in a distributed file system (HDFS). The model input path is the output path of the model, as well as in the distributed file system (HDFS).

Scheduling algorithm

In order to get clear hierarchical topological sort and ensure the system stability, based on the traditional work flow modeling approach and the hierarchical algorithm, we propose PRWF (parse and reconstruction of work flow) service flow analysis and reconstruction algorithm. The execution time of the algorithm is close to the theory system of the shortest execution time, namely process each execution unit under the condition of complete immediate execution in the implementation, so with the highest parallel efficiency and minimal processing overhead service execution flow. The basic steps of the algorithm are as follows:

1) The analysis of the relationship between the nodes and the nodes in the DAG graph;

2) The starting node is labeled as an executable node;

3) Executable node delivery platform, at the same time open up monitoring thread;

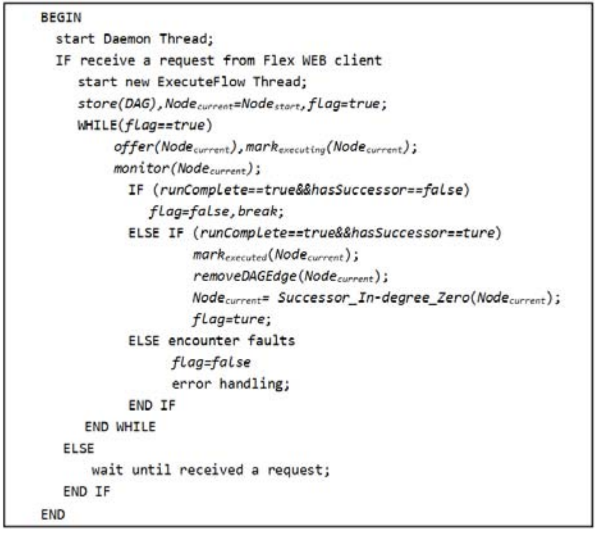
4) Monitor the execution status of the current node, wrong turn 5, and right around 6;

5) Processing error information, report to the upper;

6) to handle the subsequent nodes of the monitoring node, the specific method is: If the monitoring node without following the node, turn 7; If there is a successor node, then: Each successor node control order of the nodes, delete the associated monitoring node and successor node side, check the successor node degree information, If has in degree zero mark executable nodes, turn 3; if the penetration is not zero, not to deal with, looking for the nodes with zero;

7) perform, save the results and submit the reception;

The implementation of the algorithm is shown in the following pseudo code:



```
BEGIN
  start Daemon Thread;
  IF receive a request from Flex WEB client
     start new ExecuteFlow Thread;
     store(DAG),$Node_{current}$=$Node_{start}$,flag=true;
     WHILE(flag==true)
          offer($Node_{current}$),$mark_{executing}$($Node_{current}$);
          monitor($Node_{current}$);
            IF (runComplete==true&&hasSuccessor==false)
               flag=false,break;
            ELSE IF (runComplete==true&&hasSuccessor==ture)
                     $mark_{executed}$($Node_{current}$);
                     removeDAGEdge($Node_{current}$);
                     $Node_{current}$= Successor_In-degree_Zero($Node_{current}$);
                     flag=ture;
            ELSE encounter faults
                  flag=false
                  error handling;
            END IF
      END WHILE
   ELSE
       wait until received a request;
   END IF
END
```

Simulation Experiment

The application system is running on dedicated servers in National Marine Information Center in Tianjin. Now user can upload data, customize the software work flow and interactive with the software work flow on the visual interface.

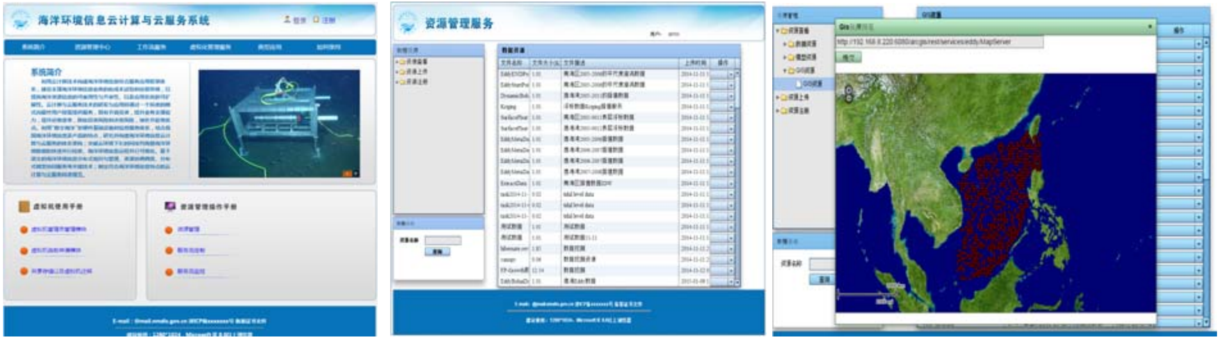
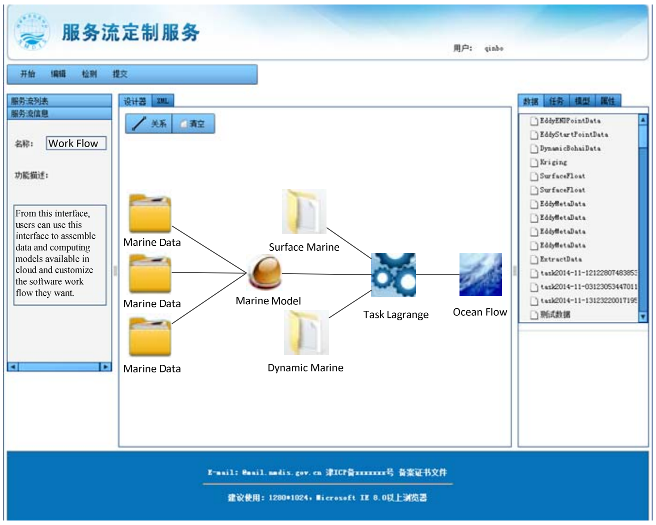


Fig.4 Software Work Flow Visual Custom Interface

Figure4 shows the visual custom interface. Users can use this interface to assemble data and computing models available in cloud and customize the software work flow they want. After customized, user also can check the software work flow in this page before the software work flow submitted.



名称: Work Flow

功能描述:

From this interface, users can use this interface to assemble data and computing models available in cloud and customize the software work flow they want.

EddyENDPointData
EddyStartPointData
DynamicBohaiData
Kriging
SurfaceFloat
SurfaceFloat
EddyMetaData
EddyMetaData
EddyMetaData
EddyMetaData
ExtractData
task2014-11-12122807483853
task2014-11-03123053447011
task2014-11-13123220017195
测试数据

E-mail: @mail.nmdis.gov.cn 津ICP备xxxxxxx号 备案证书文件

建议使用: 1280*1024, Microsoft IE 8.0以上浏览器

Fig.5 Software Work Flow Visual Monitoring and Interaction Interface

Figure5 shows the software work flow monitoring and interaction interface. From this interface, user can know the software work flow submitted in cloud. The information about the software work flow showed in this interface. The surrounded color on the model icon shows the procedure of the computing model. If the software work flow terminated abnormally, this interface will give the error message and show the abnormally termination model.

Summary

This work presents the concept of service flow and designed the whole functions and modules of the platform division combined with the characteristics of marine data. Analysis of the feasibility of software service flow on the Hadoop platform and design the PRWF (parse and reconstruction of work flow) scheduling algorithm of high efficiency. The software service platform has realized the support of large-scale parallel access and processing. By building a prototype system based on the framework, the availability and efficiency of the platform are verified. But how to recommended data and model according to user preferences and improve the prototype system in order to achieve the target system, and how to support the implementation of different language mixed service flow remains to be further study.

Acknowledgment

This work is supported by The Ocean Public Welfare Project of the Ministry of Science and Technology (No.201105033) and The National Natural Science Foundation of China. (No.40976108).

References

[1] Gang Liu, Bin Hou, Zhouwei Zhai.Hadoop. Open-source cloud computing platform [M]Beijing : Beijing University of Posts And Telecommunications Press，2011

[2] The Apache Software Foundation [EB / OL]. [2012 -06 -12]. http : // hadoop .apache.org /common/docs/current/.

[3] C. Plesums, “Introduction to Workflow,” Workflow Management Coalition Handbook 2002, 2002.

[4] W. Aalst, and K. Hee, “Workflow Management Models, Methods and Systems,” MIT Press, 2002

[5] Vincent A. Vons ,, et al. "Silicon nanoparticles produced by spark discharge." Journal of Nanoparticle Research 13.10(2011):4867-4879.

[6] Fadika, Z., et al. "Evaluating Hadoop for Data-Intensive Scientific Operations." Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on IEEE, 2012:67-74.

[7] Chen, Jianxin, and H. Tang. "Research on layering algorithm of DAG." Journal of Huazhong Normal University 42.3(2008):359-363.