Development and application of lightweight cloud platform workflow system

Qing Hou*, Qingsheng Xie, Shaobo Li

Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang Guizhou 550003, China

Received 1 October 2014, www.cmnt.lv

Abstract

Based on the principle of colored Petri net, this paper introduces a lightweight cloud platform workflow system, which will deliver BI services in Hadoop. The system aims to achieve the quick development and deployment of business processes via the workflow engine and to provide BI personalized application construction and service by way of rental. The paper expounds the workflow engine's creation principles, activity analysis, scheduling algorithm and specific application. Furthermore, it illustrates the BI parallel cloud platform deployment and system implementation. That platform is applied in the project management of operator construction and achieves some results.

Keywords: workflow engine, cloud computing, business process management, (business intelligence) BI

1 Introduction

The technical principles of workflow can be divided into Petri net, DGA or rule-based description, etc. [1-3] It separates the work into coloring and task and executes according to immobilized standard so that the workflow makes it possible for regular activities of the immobilized program in work to act in the IT system, and achieves the whole process monitoring and data analysis [4]. It is widely applied in fields like project management and office automation.

The traditional centralized workflow under the C/S mode can efficiently resolve general data analysis like Clementine and SPSS. However, with the coming of the big data age, the occurrence of BI makes users pay more attention to the relationship between the explicit data and implicit data. Besides, with explosive growth of the data scale, the increase of structured and semi-structured data, and the increasing demands of independent analysis with burstiness, the traditional workflow fails to meet the requirements of the big data age, such as mass data set and cleaning, OALP (On-Line Analytical Processing) and data mining [5].

Cloud computing takes advantage of distributed technologies to work out the calculation required by workflow and store resources on the relatively cheap facilities such as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service) [6].

The open-source Hadoop has became the foundation on the cloud computing. At present, Hadoop grows to be a huge system including mass data analysis and storage, non-structured data set and processing, task scheduling and monitoring, etc. For instance, the BI platform BC-PDM, based on Hadoop, provides users with the services of analysis and decision-making in the form of Web after putting ETL, OLAP, data mining and report analysis [6].

Based on the principle of colored Petri net, this paper proposes a lightweight cloud workflow engine and provides the successful application that the engine works in a system of an operator's construction project management when the engine is put into the platform of Hadoop. That system accomplishes configurable management of a project through workflow engine and achieves distributed analysis of the process and mass data real-time analysis of the flow processing through processing unit distribution and the distributed processing technique. As the online status works well, the system efficiently supports the operator's internal control management and decision analysis.

2 Workflow engine creations

2.1 CREATION PRINCIPLES OF WORKFLOW ENGINE

When the user's request arrives, the workflow engine will create process instances immediately and create relative files to save the operative information about the process instances. For one process instance, the case structure activities will be made into instantiation for at most once, the loop structure activities for at least once (or N times), and the others for once.

In the meantime, the engine instantiates activities that satisfy the conditions including information about activities ID, initialization time, participants, application, execution state, and so on. The engine will then save the information into the process instance and produce the work list for production users. The work list consists of the to-do task list and the done task lest. The former provides

* ***Corresponding author's*** e-mail: 282102166@qq.com

execution work items while the latter provides information inquiry of the completed tasks.

According to the scheduling process of workflow engine, this paper will categorize the workflow net into process instantiation module, activity instantiation module and task assignment and execution module.

2.2 WORKFLOW ACTIVITY DECOMPOSITION

According to the principles of the engine, the internal data of workflow is comprised of five types including organization structure, activity instance, process instance, activity definition and process definition. The data structure of workflow engine based on such definition is as follows.

1) Activity definition.

Activity information consists of the activity involved process, specific activities and the user information execution. The activities mentioned here can be divided into nine parts, which are general activities, and-join, and-join predecessor activities, and-split, or-join, or-split, or-split ending activities, begin and end.

2) Process definition.

It means to define specific activities list content including predecessor activities and successor activities in the process.

3) Process instance.

It involves the instance that has its creator of the target and the instance that has process definition of the target.

4) Activity instance.

It is saved in the task list and such instance is mainly about the task of specific activities in the process instance.

5) Organization structure.

It stores users coloring and specialization, and allocates relative privilege to correspond to activities in the work list.

2.3 WORKFLOW ENGINE SCHEDULING ALGORITHM

According to the process definition, workflow engine controls the flow of the workflow, allocates corresponding tasks to participants, and then invokes application automatically to execute. The algorithm includes process instantiation module, activity instantiation module and task assignment and execution module. The activity steps are as follows.

1) When the user builds a request, the process instantiation module will add the process instantiation required by the execution into queuing list;

2) To take out the first activity and instantiate it to create an activity instance.

3) To carry on work item allocation, task assignment and module execution, and to takes activity allocation task from the process instantiation queuing list and then store work items into users' work list.

4) Users execute the tasks in the work list and put the completed activities into done activity list.

5) The engine obtains the next activity from the process instance according to the completed task. It will come to the end if it gets the ending activity or it will turn to the 3), and re-execute the instantiation activity until the process is completed or it is aborted from the outside. The specific scheduling process is shown in Figure 1.

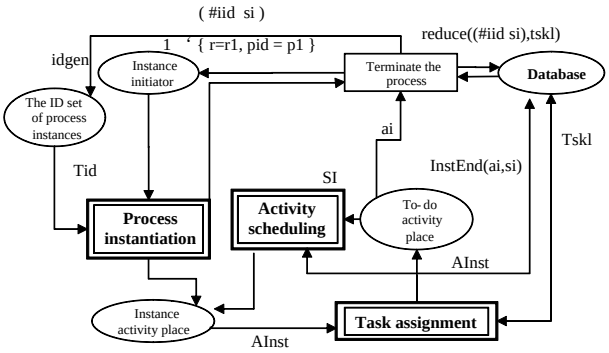

FIGURE 1 The scheduling process workflow engine

3 Colored Petri net

3.1 PRINCIPLES OF THE COLORED PETRI NET

In 1960s, a formalized modeling tool, Petri net, came into being. It achieved visual representations in graphics and got proved strictly by mathematics. However, it was still flawed by defects listed as follows.

1) It contained no data concepts. Under such circumstance, when the data control must be converted to net structure, it would increase the complexity of the model.

2) It contained no hierarchy concepts so that large modules could not be built sub-modules.

These defects restricted the scale of Petri net to be small.

In 1981, a Dane Kurt Jensen put up forward hierarchical colored Petri net (abbreviated as CP net or CPN). It used colors assertion to represent the data type of token, took functions to show the relationship between transitional excitation and colored identity, bound the place with assigned color set and designated what types of resources were to be stored in the place. It took advantage of the strict formalized descriptive method, the graphical visual representation and dynamic simulation, etc. This kind of CPN was widely applied in the modeling of many systems such as concurrence system, communication system and distribution system.

CPN can take any complicated data type as a color set. With such representation capability, it can effectively solve problems listed as follows.

1) The various state spaces generated by the dynamic workflow create excessive application nodes, which limits the computer solidification.

2) During the operation, there are uncertainty about paths generated from several executive activities and application problems about coloring.

3) During the operation, there are token dissipation and chaos caused by the processing of several instances.

CPN are nine tuples [9]. $(\sum, P, T, A, N, C, G, E, I)$,

using the color of token to describe properties of the object space. Among the tuples, *p(s)* indicates the place that is connected with the arc *s*, Var(*exp*) indicates the variable set of the expression *exp*, CMS indicates the multiple set on the set C, and Type(*v*) indicates the type of the variable *v*. The corresponding meanings are all listed in Table 1.

TABLE 1 Nine tuples of the CPN workflow engine model

Title	Tuple	Meaning
∑	Color set	Data type involved in the workflow engine
P	Place set	Workflow engine state
T	Transition set	Workflow engine operation
A	Directed art set	$P \cap T = P \cap A = A \cap T = \phi$ Data direction of the workflow engine
N	Node function	$N: A \to (P \times T) \cup (T \times P)$
C	Data type	$(P \cup T) \to \sum_{XX}$ Data types stored in the place
G	Defense function	To input essential conditions for the engine's operation, except for the parameters, $\forall t \in T$, $[Type(G(t)) = Bool \wedge Type(Var(G(t))) \subseteq \sum$
E	Arc function	workflow engine is the function of I/O, $\forall t \in T$, $[Type(E(a)) = C(p(a))_{MS} \wedge Type(Var(E(a))) \supseteq \sum$
I	Initial function	The initial identity in the operation of the workflow engine, $\forall p \in P$ $[Type(I(p)) = C(p)_{MS} \wedge Var(I(p))) = \phi$

4 CPN triggering analysis

M (p) is the multi-set of different color marks. It is used to represent the token in the place and is explained as that the place P contains two tokens in color <*g*> and three tokens in color <*r*>. The method is as follows.

$$C(m(P)) = \{g, r\} = 2g + 3r. \quad (1)$$

The Equation (1) indicates that the color set *C (P)* of every place defines the token color set that is given the access. The color set of every arc A is included in *C(P)* and the token color belongs to the color set of arc *A*. When the trigger rules and the arc function E decide to carry on the transition, the color transition gets triggered.

$\forall P_j \in t_i$ and $\forall P_k \in t_i$, if $E_f(p_i, t_i) \le m(p_j)$ and $m(p_j)$ is available, t_i is triggered and creates the new mark m as follows:

$$m(p_k) = m(p_k) + E_f(t_i, p_k). \quad (2)$$

$$m(p_j) = m(p_j) - E_f(p_i, t_k). \quad (3)$$

If and only if the importing place p_j of t_i includes as many tokens as the arc function $E_f(p_j, t_i)$ that is relative to the arc $f(p_j, t_i)$ does, the transition t_i can be triggered. When the trigger happens, t_i uses the assigned amount of color tokens from the importing place $E_f(p_j, t_i)$ and store them in the exporting place p_k. That is to say, the arc function $f(p_j, t_i)$ sets the exact number of tokens that are to be retrieved from p_j and the arc function $f(t_i, p_k)$ sets the exact number of tokens that are to be inserted in.

5 Cloud platform realizations

5.1 CLOUD PLATFORM DEPLOYMENT

The algorithm analysis units form sequential combination and achieve the realization of the application and analysis of BI on the cloud system. The cloud platform consists of the following three parts.

1) Web client: the interface used by users;

2) Workflow engine: to realize the process's analysis, distribution, execution and monitoring on the basis of CPN;

3) Cloud platform: to integrate the BI operation analysis algorithm and provide cloud storage and cloud computing services based on Hadoop; to encapsulate all activities nodes of the workflow into the node <invoke> of the Web service; to invoke corresponding abstract type to achieve dynamic mount of different objects and finally complete the whole process.

When users submit their requirements from the Web client, the specific activity on the cloud platform is shown in the Figure 2.

1) Workflow submits transaction requirements to the workflow engine to process; workflow engine then analyzes it as DAG (Directed Acyclic Graph) based on parameters instantiation and saves in the process list;

2) If the users requirements are cloud transaction requirements, then deployment module will analyze and invoke relative BI algorithm and submit to cloud platform;

3) Job Tracker on the cloud platform arranges and executes Job, processes through the calculation mode MapReduce in the Hadoop Distributed File System (or HDFS) and stores the final results in the BigTable of HBase;

4) The system will send back the results to users thus complete the whole process.

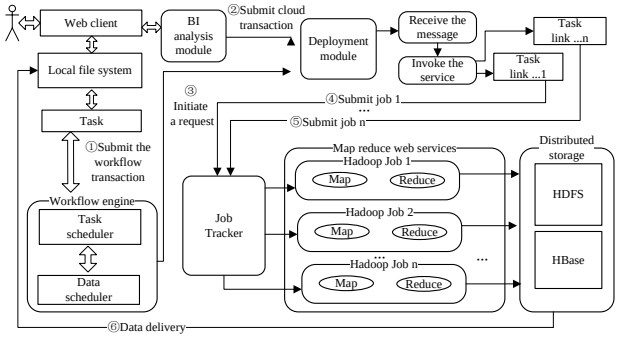

FIGURE 2 The deployment and execution of the cloud platform

5.2 SYSTEM REALIZATION

Combine with CPN the researchers then develop the lightweight workflow middleware of the enterprise's quick information development platform. Such middleware can achieve the graphical process configuration and route control. Process instantiation can dynamically judge executing node and executing route according to parameters and rules. The dark color part indicates executed node and the black part means no need to execute.

Other lighter color node shows that the process has not been completed as shown in the Figure 3.

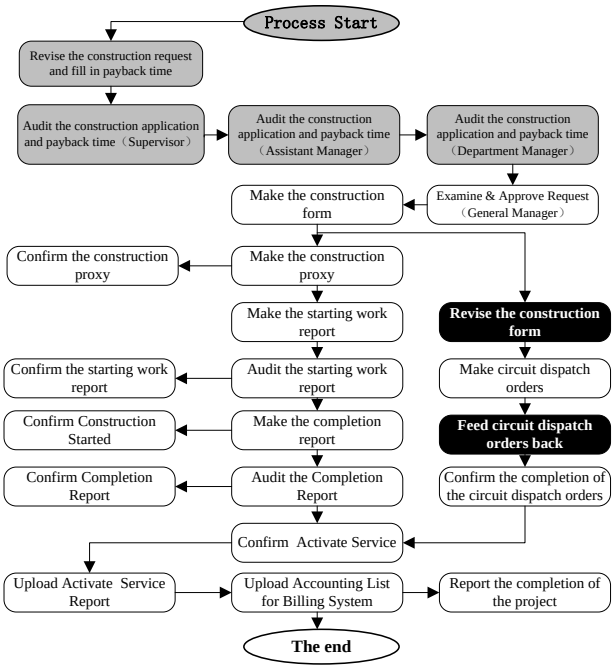

FIGURE 3 The process of an operator's construction project

The cloud service system of an operator's construction project management information integrates the workflow middleware, BI analysis module and the 0.20.2 version Hadoop. The deployment and operation of the system has achieved a success and passed the three-month environment and pressure test.

The operator's construction project includes five categories and 18 sub-categories and achieves the visualized configuration through the workflow middleware. The coloring process and authority of the process nodes correspond to the categories of clients information, get integrated into the seamless communication of all the nodes and deal with personalized requirements (like time limit) just as it is shown by Figure 4.

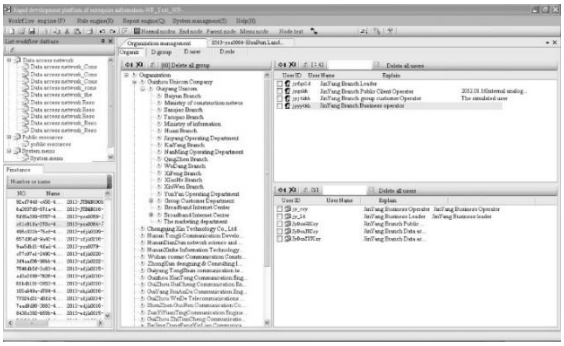

FIGURE 4 The middleware of configuration platform of workflow

In the system, the size of a single construction project ranges from 200MB to 6GB, averaging at about 1.2MB. The projects for the whole province total about 6000 per year, taking up 7.2TB. Considering the requirements of HDFS and the highly fault-tolerant mechanism, the storage space of cloud system is extended with the proportion of 3 to 1, totaling 24TB.

In the environment of cloud platform, an IBM3850 is used as the Web server and database server; an IBM3650 is used as the server of workflow engine; an IBM3650 is used as the master control server for the Hadoop platform; six old servers in the type of Xeon E5405@2GHZ equipped with the two ways, four cores, the capacity of 16 G and the 4TGB hard disk are used as the sub-node for the cloud platform.

In the BI analysis, the cloud platform is subtilized to the node's granularity thus to count the information of different links in the process, and to analyze and process. The information involved are detailed information of any single item, timeout details of multi-latitude projects, the operational conditions of different departments in the process, timeout details of to-do tasks and corresponding evaluation of coloring. Then the platform is able to carry out the analysis of the whole construction and the using proportion of the capital budget according to initiating departments and localization areas. Based on the Hadoop, the cloud platform analysis can provide real-time operation analysis data and support the fair management and quick decision making analysis of the enterprise as it is shown in the Figure 5.

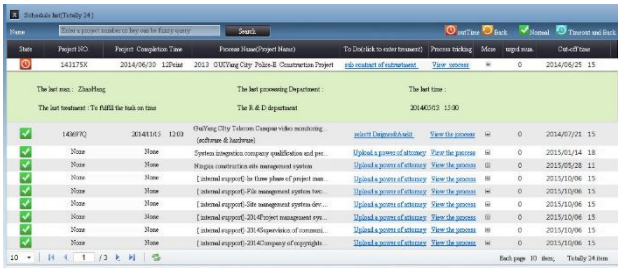

FIGURE 5 The cloud platform of an operator's construction project management

6 Conclusions

Based on CPN, this paper proposes a lightweight cloud platform workflow system and achieves the application of SaaS. The system combines the counting and analysis of parallel BI and integrates the distributed calculation and capacity of storage of the cloud platform thus to provide effective project management and operation analysis application for the operator's construction project management. The application indicates that the system can prop the provincial high concurrency of multi-users, invoke data efficiently to analyze the algorithm and show real-time results of the data analysis.

The next step of research of the cloud platform can focus on: 1) further perfection of the efficiency of the cloud workflow; 2) the establishment of ODS and effective data cleaning; 3) the improvement of BI functions such as OLAP, data mining, and cloudization of the report analysis.

Acknowledgments

National Key Technology R&D Program, Guizhou cultural heritage digitalization protection and development key technology research and application, Project (2014BAH05F01); Ministry of Science and Technology Innovation Fund for Technology Based Firms, Guizhou Communication Technology Public Service Platforms, Project (12C26245206305).

References

[1] Fan Y, Li X, Wang Q 2008 Bottom Level Based Heuristic for Workflow Scheduling in Grids *Chinese Journal Of Computers* **31**(2)
[2] Liang Y, Xu F 2010 Study on modeling of process ontology for enterprise patent resources management *Computer Engineering And Applications* **46**(1)
[3] Wu S 2009 Workflow model of construction projects based on Petrine *Computer Engineering And Applications* **45**(30)
[4] Fan Y 2001 Basic technology of Workflow Management System [M] *Beijing: Tsinghua University Press*
[5] Yu L, Zhao S, Zhang Y, et al. 2013 The application of cloud workflow technology in BlSaaS *Computer integrated manufacturing system* **19**(9)
[6] Yan G, Yu J, Yang X 2013 Scientific two-phase workflow scheduling strategy nder cloud computing environment *Computer application* **33**(4)
[7] Yu L, Zheng J, Wu B 2012 BC-PDM: data mining, social network analysis and text mining system based on cloud computing *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York N.Y.USAACM* 1496-9
[8] Jensen K 1994 An introduction to the theoretical aspect of colored Petri nets *A decade of concurrency lecture notes in Computer Science* 803 230-72
[9] Li B, Zhang L, Ren L 2012 Typical characteristics, technologies and applications of cloud manufacturing of cloud computing *Computer Integrated Manufacturing Systems* **18**(7)

Authors

Hou Qing, 1980, Tianjin, China.

Current position, grades: senior engineer and doctoral candidate, the vice president of Guizhou Planning & Design Institute of Posts & Telecommunications.
Scientific interest: computer networks, computer integrated manufacturing system (CIMS).

Xie Qingsheng, 1954, Guiyang, China.

Current position, grades: professor and tutor of doctoral candidate.
Scientific interest: network manufacture, CIMS and computer aided innovation design.

Li Shaobo, 1973, Shaoyang, Hunan Province, China.

Current position, grades: professor and tutor of doctoral candidate.
Scientific interest: information management and information system, cyber information security, E-Government, computer aided innovation.