A Study on Data Analysis Process Management System in MapReduce using BPM

Yoon-Sik Yoo[1], Jaehak Yu[1], Hyo-Chan Bang[1], Cheong Hee Park[2]

[1] Electronics and Telecommunications Research Institute, 138 Gajeongno, Yuseong-gu, Daejeon, 305-700, Korea
[2] Dept. of Computer Science and Engineering Chungnam National University, 220 Gung-dong, Yuseong-gu, Daejeon, 305-764, Korea
[1]{ys5315, dbzzang, bangs}@etri.re.kr,
[2]cheonghee@cnu.ac.kr

Abstract. MapReduce is a distribution-system-based programming model to process massive data and has been utilized as an analysis model not only in the academic world but also in the industrial fields. However, developers who implement MapReduce have some deficiency in understanding the data analysis, while data analysts have difficulty in programming MapReduce for various analyses by themselves. Hence, it is difficult for developers to provide a demanded analysis output. In order to solve such difficulty between developers of MapReduce and the data analysts, this study proposes a new MapReduce analysis process management system based on BPM (Business Process Management). This system was designed to provide a mutual complimentary intermediary function for MapReduce developers and analysts, and also makes it possible to respond flexibly to any alteration of analysis procedure.

Keywords: MapReduce, BPM, Analysis system, Data processing

1 Introduction

MapReduce is a programming model and an associated implementation for processing and generating large datasets that is amenable to a broad variety of real-world tasks [1], [2]. MapReduce has spread widely through Apache group"s open source project, "Hadoop" [3]. Hadoop was distributed by posting an environment to initiates the MapReduce functions on the HDFS that implemented GFS (Google File System) [4] as an open source [5]. However, it is not easy for data analysts to understand and program such MapReduce framework suitably to their analysis purpose. On the contrary, it is difficult for the program developers to get into the analysis domain because of the difficulties in understanding the data properties fundamentally, utilizing analysis methods efficiently, and interpreting the results. In order to solve such problems, this paper utilizes BPM(Business Process Management)

[2] Corresponding author

where data interworking between systems and flow control according to out-put data results are possible [6], [7].

BPM's process modeling procedure is as follows. First, it defines the application for interworking with the system in advance. Next, a process modeler defines the activities corresponding to each procedure. After that, it performs the work of establishing the procedural flow between activities and mapping the application to each activity suitably. BPM engine is a system that performs the initiation work of the processes defined through such modeling.

MapReduce analysis process management system newly proposed in this paper is so designed as to perform MapReduce job in a BPM application to utilize a BPM engine. In addition, it is designed to make it possible to process an intelligent data analysis by controlling the conditions between diversified MapReduce jobs. The implementation of such architecture enables MapReduce application and analysis process to be loosely coupled so that it can be applied to any flexible alteration of analysis procedure. Efficient data refinement and transmission are also possible by utilizing BPM in order to perform any MapReduce job scheduled.

The paper is organized as follows. In Section 2, we describe MapReduce and the BPM system. In Section 3, the proposed MapReduce analysis process management system is presented. Finally, Section 4 discusses the conclusions and future research directions.

2 Related Work

MapReduce is a programming model and an associated implementation for processing and generating large data sets [1]. MapReduce automatically parallelizes and executes the program on a large cluster of commodity machines. The runtime system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing required inter-machine communication. MapReduce allows programmers with no experience in parallel and distributed systems to easily utilize the resources of a large distributed system. Typical MapReduce computation processes many terabytes of data on hundreds or thousands of machines. Programmers find the system easy to use, and more than 100,000 MapReduce jobs are executed on Google's clusters every day [2].

Conceptually the map and reduce functions supplied by a user have associated types as follows.

$map\ k\ v$

( ) ( )

, → 1 1 2 2

$$reduce\ k(list\ v, \ ) \mapsto (\ )$$

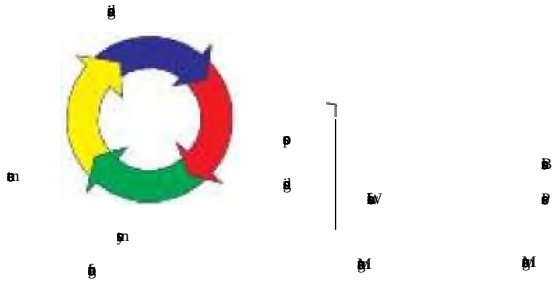
$$_2 \quad _2 \quad _2$$

(1)

That is, the input keys and values are drawn from a different domain than the output keys and values. Furthermore, the intermediate keys and values are from the same domain as the output keys and values.

Many people consider BPM (Business Process Management) to be the "next step' after the workflow wave of the nineties. Therefore, we use workflow terminology to

define BPM. BPM includes methods, techniques, and tools to support the design, enactment, management, and analysis of operational business processes [6], [7], etc. In the last couple of years, many researchers and practitioners started to realize that the traditional focus on enactment is too restrictive. As a result new terms like BPM have been coined. There exist many definitions of BPM, but in most cases it clearly includes WFM (Workflow Management). Note that this definition restricts BPM to operational processes. In other words, processes at the strategic level or processes that cannot be made explicit are excluded. Fig. 1 shows the relationship between WFM and BPM by using the BPM lifecycle [6]. The BPM lifecyle describes the various phases in support of operational business processes. In the configuration phase, designs are implemented by configuring a process aware information system. After configuration, the enactment phase starts where the operational business processes are executed using the system configured. In the diagnosis phase, the operational processes are analyzed to identify problems and to find things that can be improved. The focus of traditional workflow management is on the lower half of the BPM lifecyle. As a result there is little support for the diagnosis phase. Moreover, analysis and real design support are missing. It is remarkable that few WFM systems support simulation, verification, and validation of process designs. It is also remarkable that few systems support the collection and interpretation of real-time data. Note that most WFM systems record data of process tasks. However, no tools to support any form of diagnosis are offered by the traditional systems.



process

Fig. 1. The BPM lifecyle to compare Workflow Management and Business Process Management

3 System Architecture

BPM-based analysis process modeling system consists of 3 layers as shown in Fig. 2.

1) Data Storage Layer: This is a Hadoop-based physical data storage space. It can include a legacy system to provide the initial data of an analysis or store/provide intermediate data.
2) MapReduce Application Layer: In this layer, there exists the implementation of MapReduce job to be performed in the Data Storage Layer. In addition, there exists the implementation of legacy system"s interworking interface.

3) Analysis Process Layer: This is the layer where the process initiates and controls the applications provided in MapReduce Application Layer. This makes it possible to control the conditions by providing variables for each application.

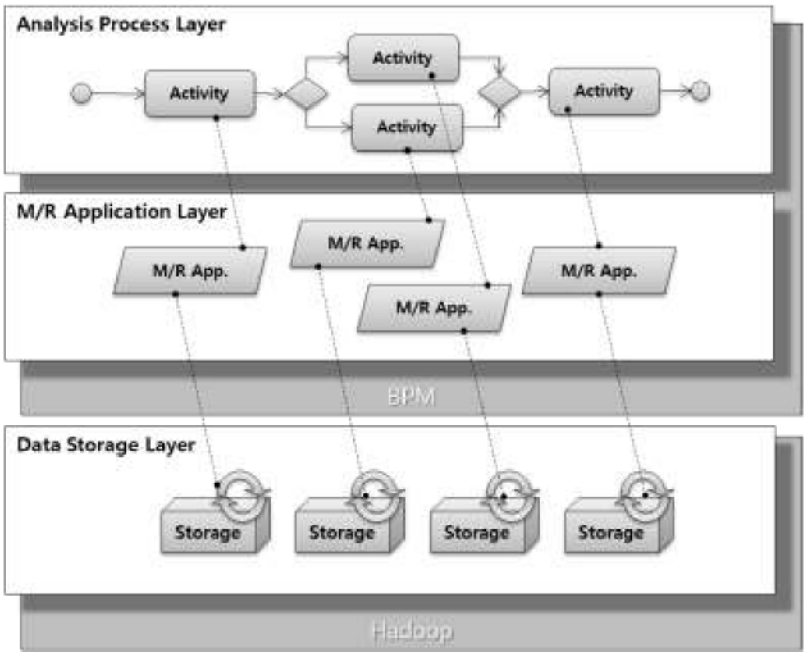


Fig. 2. System Architecture Layer: Data storage layer, MapReduce application layer, Analysis process layer.

The processing procedure for the definition of the MapReduce analysis process is shown in Fig. 3. The detailed execution procedure is as follows: 1) The MapReduce application developer implements various MapReduce functions to provide services. 2) The implemented MapReduce applications shall be registered in BPM‟s MapReduce application repository. 3) Data analysis process modeler finds a suitable MapReduce application and mapping on activities of analysis process. 4) Data process modeler makes a modeling of various analysis processes. 5) Data process modeler registers analysis process definition to the BPM‟s analysis process repository, and makes a scheduling on time to initiate the process or define the rules. 6) BPM produces process instances by initiating the modeled process definitions after defining the input values of the initial variables. 7) Analysis process instances initiate the MapReduce application which has been mapped when modeling. In the application called, an actual analysis work is performed by initiating the defined MapReduce job. The analyzed data shall be utilized as input data for the next analysis stage.

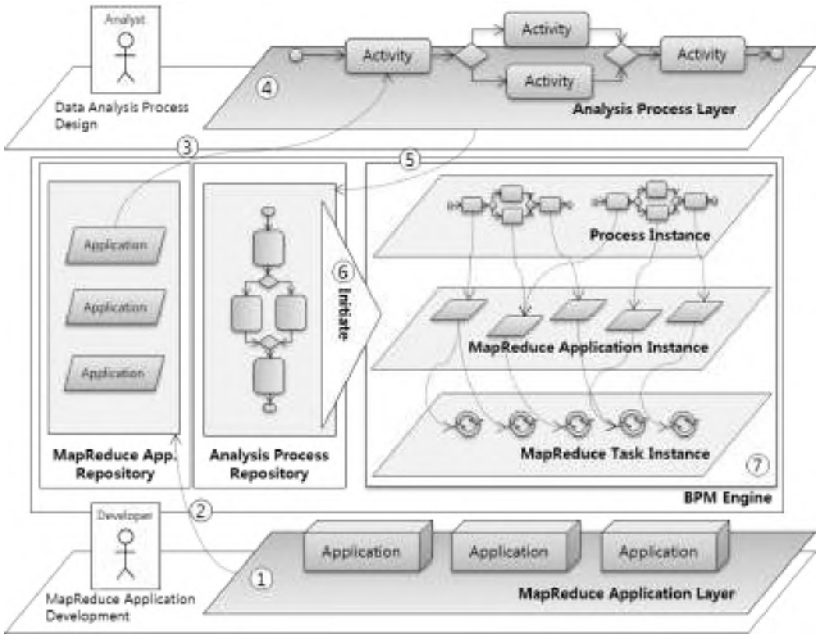


Fig. 3. Define and initiate procedure of MapReduce analysis process: 1) ~ 2) MapReduce application development and register. 3) ~ 5) Analysis process modeling and register. 6) ~ 7) Analysis process initiate and running.

4 Conclusion

In this paper, BPM was applied as a mutual intermediary system between MapReduce developers who have no knowledge about data analysis and data analysts who have no experience of program implementation. MapReduce analysis process management system newly proposed in this paper is so designed as to perform MapReduce job in a BPM application utilizing a BPM engine. In addition, it is designed to make it possible to process an intelligent data analysis by controlling the conditions between diversified MapReduce jobs. The implementation of such architecture enables MapReduce application and analysis process to be loosely coupled so that it can be applied to any flexible alteration of analysis procedure. Utilizing the characteristics of BPM, it is possible to extend to various services, for example, transmitting analysis results to legacy system besides MapReduce, refining data and accumulate them in RDBMS, and interworking with user work system, etc.

In the future, it is intended to open the implementation of MapReduce job in a service format, register its contents and develop more expanded process modeling system by applying the analysis system of SOA/ESB [8] architecture that can be utilized by the combination of analysis service.

Acknowledgments. This work was supported by the IT R&D program of MKE/KEIT (Project No. 10038653-2010-412, Development of Semantic based Open USN Service Platform).

References

1. Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. In: Proceedings of the USENIX Symposium on Operating Systems Design & Implementation (OSDI), pp. 137--147 (2004)
2. Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. J. Comm. of the ACM. 53, 1, 72--77 (2010)
3. White, T.: Hadoop: The Definitive Guide. O'Reilly, Sebastopol (2009)
4. Ghemawat, S., Gobioff, H., Leung, S.: The Google File System. In: Symposium on Operating Systems Principles, pp. 29--43 (2003)
5. Hadoop Distributed file system, http://hadoop.apache.org
6. Wil, M. P., Arthur, H. M., Mathias, W.: Business Process Management: A survey. In: Lecture Notes in Computer Science, LNCS, vol. 2678, pp. 1--12 (2003)
7. Jung, J., Kong, J., Park, J.: Service Integration Toward Ubiquitous Business Process Management. In: IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2008, pp. 1500--1504 (2008)
8. Chappell, D.A.: Enterprise Service Bus. O'Reilly, Sebastopol (2004)