J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

HBase based Business Process Event Log Schema Design of Hadoop Framework☆

Seonghun Ham[1] Hyun Ahn[1] Kwanghoon Pio Kim[1]

ABSTRACT

Organizations design and operate business process models to achieve their goals efficiently and systematically. With the advancement of IT technology, the number of items that computer systems can participate in and the process becomes huge and complicated. This phenomenon created a more complex and subdivide flow of business process.The process instances that contain workcase and events are larger and have more data. This is an essential resource for process mining and is used directly in model discovery, analysis, and improvement of processes. This event log is getting bigger and broader, which leads to problems such as capacity management and I / O load in management of existing row level program or management through a relational database. In this paper, as the event log becomes big data, we have found the problem of management limit based on the existing original file or relational database. Design and apply schemes to archive and analyze large event logs through Hadoop, an open source distributed file system, and HBase, a NoSQL database system.

☞ keyword : Workflow Process, NoSQL, Process Mining, Event Log, Hadoop, Process Discovery

1. Introduction

Many organizations are designing and operating business processes for efficiently business management. Those business processes can be automated with the helpof information system. This automated system is called (PAIS)Process-Aware In-formation System). Recently, many organizations adopt a BPMN and applicate Process Automation Methodology. When an established process model operatesfor a purpose, a series of processes that occur when defined as an instance orcase. It is also recorded in the event log. The event log contains the process flowand result. The process flow is determined in real time according to the con-text and the control flow of the model, and the result is also influenced by theprocess flow. This is only known when the operation of the process is complete. Event log is an important resource for the organizations because process miningcan extract useful information for process analysis and improvement. Therefore, storage, management, and analysis of logs are critical issues.The technology of software and hardware of IT environment is developedrapidly. Moore's Law predicted an increase in hardware performance. Addition-ally, many experts predict that the amount of data will grow exponentially .Advances in technology have made automation, segmentation and diversifica-tion possible, and the process complicated and huge. As a result, the event logof the process contains more events and attributes. This means that the processevent log is also bog data. The storage medium must accommodate this largeamount of data. The event log of a business process is significant in its ownright, making it difficult to sample or delete some. Also, large amounts of datamust be generated or processed quickly. RDB(Relation Database) required highperformance and high maintenance costs to store and manage large amount ofdata. The event log is not optimized for RDBs that require a fixed format withsemi-structure data. The data sector is using NoSQL(Not Only SQL) as an al-ternative to big data processing. Storage and process big data efficiently throughsimplified design, horizontal scalability, and deregulation.

Hadoop is framework based on distributed file system for handle big data. Itwas developed based on Google's GFS(Google file system) study [1] and MapRe-duce study [2]. This is open source and available from the Apache Software Foundation. The key features of Hadoop are HDFS(Hadoop

1 Div. of Computer Science and Engineering, Kyonggi University, 154-42 Kwangkyosan-ro Youngtong-gu Suwon-si Gyeonggi-do, 16227, Republic of Korea
* Corresponding author (kwang@kgu.ac.kr)
[Received 20 June 2019, Reviewed 15 July 2019(R2 28 August 2019), Accepted 17 September 2019]
☆ A preliminary version of this paper was presented at ICONI-IST 2018.
☆ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant No. 2017R1A2B2010697). This work also was partially supported by Kyonggi University's Graduate Research Assistantship 2019.

Distributed FileSystem) and MapReduce. This consists of one master node and several slavenodes. It utilizes the CPU and storage of slave nodes. In this regard, Hadoopis very easy to scale up horizontally. Because of this, Hadoop does not requirehigh-performance slave nodes and many companies use it as a big data storageand processing framework.

In this paper, we found that as event logs become big data, RDB-basedstorage and management requires excessively high performance and is not opti-mized for semi-structured data. Therefore, we build NoSQL server in distributedenvironment through HBase based on Hadoop framework and design schemastructure for efficient storage and analysis of event log.

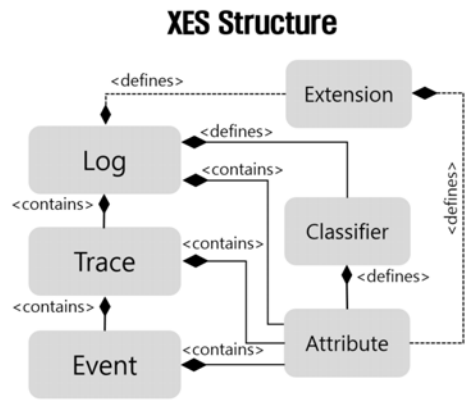
2. Related Work and Scope

2.1 Process Mining

Process mining is the discovery of valuable information from the event log. It isclassified into three types according to the method and purpose. First, discoverymethod is constructing and analyzing a process model with only a log. Second,conformance checking method is comparing and analyzing a model extractedfrom a event log and original process model. Last, enhancement method is findinga process improvement method through event log and original process model.These method are closely related to the processes life cycle. If any issues arisefrom the process, process mining can work as follows. First, the discovery methodwill be used to identify the cause and situation of the issue. The discoveredand analyzed processes are then assessed to conform to existing designs andtargets using conformance checking methods. Once the previous two steps havebeen completed, the enactment method will redesign the process. The redesignedprocess is implemented and operational, again monitored and controlled. If anyissue occurs, repeat the above process. Event logging is a very important resourcein this life cycle.

2.2 Process Event Log

Event logs originating from business process instances are recorded in a spe-cific format. Several organizations have defined the format for event logging.MXML(Mining eXtensible Markup Language) proposed by Eindhoven. It isconsist of four layers: workflow log, process, process instance, and audit trailentry. The lowest audit trail entry has various attribute values for the event.CWAD(Common Workflow Audit Data) offered by the International Organiza-tion for Standardization. In this, prefix information and suffix information havePK(Primary Key), and process instance audit information refers to FK(ForeignKey). XES(eXtensible Event Stream) [3] submitted by IEEE are available. It isconsists of three layers: log, trace, and event. The log layer consists the meta-data of the event log, the trace layers, and the definition of attributes used in the trace and event. Inside the trace layer is the temporal workcase and attributevalues of the business process. temporal workcase represents the beginning tothe end of a completed process and consists of event layers. The event layer isthe lowest layer. The unit of work that occurs in a process. When the process isrunning, events occur according to the control flow. The listing of these eventsconstitutes a temporal workcase. The event log used in this paper is recorded inXES format. Figure 1 shows the structure of the XES log format.



(Figure 1) XES log data structure[3]

2.3 Discover Process Model

The event log consists of workcases, which are time based. The structure of aprocess model composed of several control flows in parallel, selective, and repet-itive. However, the workcase is sequential. It cannot represent a complex flowof control flows. Control flow is cannot recorded accurately in the event log.Therefore, an algorithm for extracting the correct

process model from the logwas studied. The alpha algorithm [4] extracts a Petri-net based process model. The Petri-net model is one of the graphical notations of the process model. Com-paring each workcase reveals the control flow. This notation includes optionalcontrol flows because there is no separate notation for iterative control flows.The sigma algorithm [5] extract a ICN model based process model. The ICNmodel is one of the graphical notations of the process model. This can representparallel, selective, and repeatable control flows. Discover control flow throughworkcase flow.

The previous two algorithms have limitations in handling complex structuresof multiple control flows. In the case of a complex structure in which anothercontrol flow exists inside the control flow, several flows are mixed. In order tosolve such a case, a study [6] of discovering a control flow using a weight of arelationship has been conducted. Sort and classify event relationships within alltraces to reconstruct the model and weight the relationships. The type of controlflow is determined by comparing the input weight and the output weight. Thisinformation is not found on the workcase and is a very accurate discriminatingfactor in complex control flows.

2.4 Hadoop Echo System

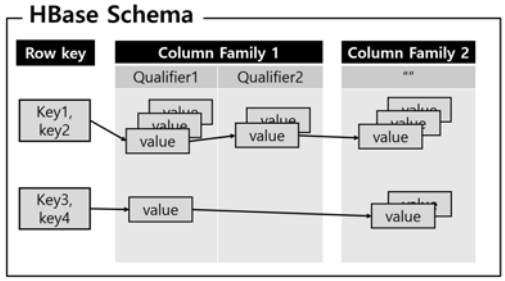
Hadoop is basically made up of HDFS and MapReduce. MapReduce classifiesinto Map function and Reduce function through key-value data. This is a greatway to deal with unstructured big data in distributed file systems. Research[7] has been done on the processing of big data in Hadoop. In the process field, researches[8] on clustering and storing temporal workcases were conducted inMapReduce. Several sub-projects have been underway to use Hadoop more ef-ficiently. Types include distributed database, machine learning, in-memory pro-cessing, data warehouse, interactive query processing, and workflow The Hadoopframework with these features is called the Hadoop ecosystem. In this paper, weuse HBase, a distributed database, for the efficient storage and management oflarge process event logs.

3. HBase-based Event Log data warehouse

HBase is a Hadoop-based columnar NoSQL database. It consists of table, row,column family, column, column qualifier, cell, time stemp, and version. Thecolumn family is static, but the columns in it are not. Therefore, HBase doesnot provide a way to query the list of all columns because each row can havedifferent columns. The fact that each row can have a different column is suitablefor representing the process event log. In the XES log format, the log matadatacontains definitions of attributes and attribute values that can be included intraces and events. However, each trace and event does not always have theseall attribute values only the value that correspond The attribute and attributevalues of an event are determined by the type of event and the progress of theprocess.

3.1 HBase's NoSQL Schema Model

NoSQL has several data models. HBase is a Big Table-style Model. This is a key-value format. Unlike a row-based RDB model, the value is based on a column,and the value itself is constructed as a continuos, multi-dimentional map. First,the row key can consist of a single key, but a composite key is also possible. Thevalue is composed of column, but there can be more than one column. Similarcolumns can be grouped into column families and can be grouped into qualifierswithin them. The qualifier is not essential and can be used as needed. Also, thereis no restriction that each row should have several columns The version of thecolumn is managed through a timestemp, and basically, the value for the latestversion is read. Previous versions are stored without being deleted. Figure 2 isa graphical representation of HBase's data schema.



(Figure 2) HBase Schema

3.2 Row Key in XES Log

In the XES log file format, trace and event have different property values depend-ing on the type and context. However, id and timestemp that uniquely representthe event are included. Figure 3 shows part of the event log in XES log format.”concept: name” represents the unique id of the event. The study[9] of findinga process model from the event log used a sequential sequence of events. Se-quential ordering of events shown in the event log is assumed as the workflow. InFigure 3, we can assume that there is a business flow from event ”Record InvoiceReceipt” to event ”Clear Invoice”. The relationship between these events is themost fundamental unit of process mining. Therefore, a row must keep track ofits own successor events. The process event log is based on the process model.The number and relationship of events in the model is fixed. Temporal workcasescan be organized similar to other temporal workcases. That is, the same eventmay appear. In addition, the same event can appear multiple times through aniterative control flow within a temporal workcase. For this reason, four valuesare required for the row key. event id, successor event id, trace id, timestemp.

```
<event>
    <string key="User" value="user_001"></string>
    <string key="org:resource" value="user_001"></string>
    <string key="concept:name" value="Record Invoice Receipt"></string>
    <float key="Cumulative net worth (EUR)" value="298.0"></float>
    <date key="time:timestamp" value="2018-03-06T07:53:00.000Z"></date>
</event>
<event>
    <string key="User" value="user_002"></string>
    <string key="org:resource" value="user_002"></string>
    <string key="concept:name" value="Clear Invoice"></string>
    <float key="Cumulative net worth (EUR)" value="298.0"></float>
    <date key="time:timestamp" value="2018-03-29T13:06:00.000Z"></date>
</event>
```

(Figure 3) XES Log Format

In HBase, the key configuration order of complex keys is important becauseit directly affects search performance. Placed on the region server based on therow key, all searches are done using the row key. You can also use complexkeys efficiently through partial search of keys. You should also be concernedabout hot spot issues. Due to the nature of the process event log, the value ofone event information is small. When events are gathered together to create aworkflow, the importance of value is increases. Multiple traces are aggregated sothat the weights and control flow propagation rates through relationship countsare important information for discovering and analyzing accurate process mod-els. The study[6] found disjunctive process patterns refinement and probability extraction from workflow logs. The event log is sequential, so the timestempincrements, so hot spot issues are likely to occur. In addition, trace id is neces-sary information to indicate the relationship between events, but since all eventrelationships belonging to the same temporal workcase have the same trace id,search performance is degraded. Therefore, the sequence of row keys is arrangedin order of event id, successor event id, timestemp, and trace id.

Next, we need to construct a column value. Business processes are orga-nized on a workcase basis. It consists of various attributes, such as tasks, roles,performers, data, and applications. OLAP(Online Analytical Processing), whichwas widely used for organizational decision making, performed multidimensionalanalysis while looking at data based on various criteria. In addition, process anal-ysis through social network analysis studies [10] were conducted to discover andanalyze other attribute-based flows out of the existing workcase based on so-cial network techniques. Considering this analytical point of view, we need toconstruct a column family for multidimensional analysis. The value should notlose as much of the workcase information as possible, including predecessor con-trol flow and successor control flow, as well as the attribute values of the event. Therefore, the column family consists of two groups, the relationship betweenthe events and attributes of event. Figure 4 shows the HBase event log schemaproposed in this paper.

	Column Family	
Row Key	**Event Attribute**	**Event Realtion**
EventID : SuccessorEventID: TimeStemp: TraceID	Role, User, Data, ...	Predecessor control flow, Successor control flow, ...

(Figure 4) NoSQL Schema for Event Log

4. Materialization

In the previous chapter, we designed HBase-based NoSQL schema for storing and analyzing large process event logs. In

this chapter, the designed schema is actually built and connected with the process event log analysis tool. The dataset to be used for storage and analysis is “BPI Challenge 2018.xes” [11] provided by 4TU Center for Research Data. This event log is a real life log generated from the EU direct payment application processing process for farmers in the European Agricultural Guarantee Fund. For a total of three years, 43,809 traces and 2,514,266 events were included.

4.1 Building Hadoop Echo System

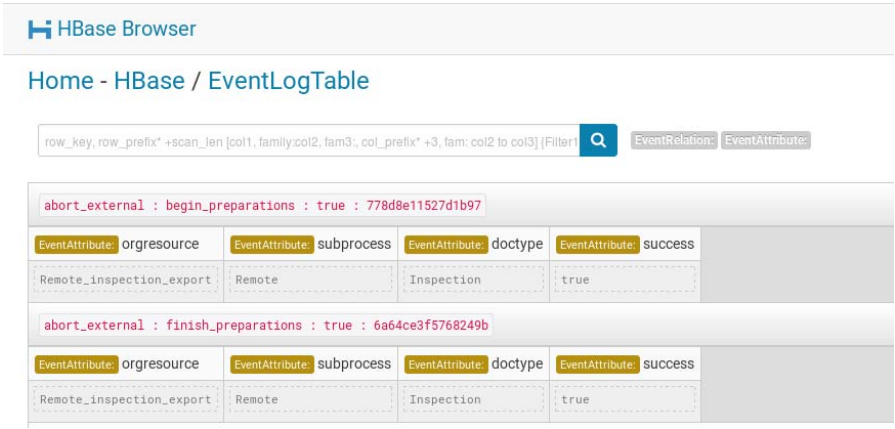
In this study, a fully distributed Hadoop Echo System was constructed throughfive computers. We use Cloudera Manager and CHD(Cloudera Hadoop). CHD version is 5.13.0. CDH is a package system that provides easy management ofHadoop distributed mode and server nodes through web-based control and mon-itoring through the manager provided by Cloudera. We built HDFS, YARN(MapReduce2), HBase, Hive, Hue, and ZooKeeper in the echo system. The op-erating system is Linux Ubuntu 16.04.5LTS, and the specifications of the servercomputer are the same with intel i5, 8GB of RAM and 500GB of hardware, andthe master computer is the same with the rest of 16GB of RAM .

4.2 Preprocessing

From the process event log, we use the algorithm[9] for control-path-basedprocess knowledge analysis to find the relationship and attribute values betweenevents. An event recorded after an event on the log is assumed to be a successorevent in the process model. It also parses the attribute of event based on requiredThis data is stored in the schema we designed earlier. The values stored throughHBase are sorted according to rowkey and assigned to the region server. Inthis paper, we placed timestemp behind the event id and successor event id inthe structure of the complex key. If the timestamp is placed at the beginningaccording to the general method, the sequential time is listed in the nature of theprocess, so that it has a similar key value. This can lead to overloading of I/O bybeing placed on the same region server, and there is a problem of redistributionthrough balancing work in the future. When we use the proposed schema, thesorting is done based on the event id, so it is suitable for the analysis usingthe whole data of the log such as the relation weight and the control flow rate.Because similar types of event relationships are placed in the same region server,data can be physically contiguous for large-scale event log analysis, making itmore efficient, such as clustering with MapReduce. Figure 5 shows the HBaserow and column through Hue monitoring program Hue.

6. Conclusions

This paper presents the need for efficient storage and analysis of large-scaleprocess events. To solve this problem, a NoSQL database in a distributed envi-ronment was



(Figure 5) Schema management via hue

constructed using the Hadoop echo system including HBase. Insidethe database, HBase designed and implemented a schema suitable for event logs.We consider the characteristics of the log, which is a semi-structured data for-mat, and consider the physical efficiency of analysis and large files. The row iscomposed based on the event layer, the lowest level of the XES log format. Inaddition, a successor event is included in the row to preserve flow informationthat is a characteristic of the workcase. Attributes are classified and stored incolumn family according to their type to increase the efficiency of searching.The row key consists of event id, successor event id, timestemp, and trace idto ensure uniqueness of the row and prevent hot spot issues. Hadoop HBase'sschema for large process event logs will help organizations manage and improvetheir process models with low maintenance and high performance for storing andanalyzing.

Acknowledgment. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant No. 2017R1A2B2010697).

References

[1] Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system." 2003. https://ai.google/research/pubs/pub51

[2] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM, Vol. 51, No. 1, pp. 107-113, 2008. http://dx.doi.org/10.1145/1327452.1327492

[3] Günther, Christian W., and Eric Verbeek., "Xes standard definition," Fluxicon Process Laboratories, Vol 13, No. 14, 2009. https://pure.tue.nl/ws/portalfiles/portal/3981980/692728941269079.pdf

[4] W. M. P. van der Aalst, B. F. van Dongena; J. Herbst, L. Marustera, G. Schimm and A. J. M. M. Weijters, "Workflow mining: A survey of issues and approaches," Journal of Data & Knowledge Engineering, Vol. 47, Issue 2, pp. 237-267, 2003. https://doi.org/10.1016/S0169-023X(03)00066-1

[5] Kim, Kwanghoon and Ellis, Clarence A., "σ-Algorithm: Structured Workflow Process Mining Through Amalgamating Temporal Workcases," The Proceedings of PAKDD2007, Advances in Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence, Vol. 4426, pp. 119-130, 2007. https://doi.org/10.1007/978-3-540-71701-0_14

[6] K. im, M. Yeon, B. Jeong, and K. P. Kim, "A Conceptual Approach for Discovering Proportions of Disjunctive Routing Patterns in a Business Process Model," KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS, Vol. 11, No. 2, pp. 1148 - 1161, 2017. https://doi.org/10.3837/tiis.2017.02.030

[7] Patel, Aditya B., Manashvi Birla, and Ushma Nair. "Addressing big data problem using Hadoop and Map Reduce." 2012 Nirma University International Conference on Engineering (NUiCONE). IEEE, 2012. https://ieeexplore.ieee.org/abstract/document/6493198

[8] Minhyuck Jin, and Kwanghoon Pio Kim. "A MapReduce-Based Workflow BIG-Log Clustering Tec," Journal of Internet Computing and Services, Vol. 20, No. 1, pp. 87-96, 2019. https://doi.org/10.7472/jksii.2019.20.1.87

[9] Park, Min-Jae, and Kwang-Hoon Kim. "Control-Path Oriented Workflow Intelligence Analysis and Mining System." 2007 International Conference on Convergence Information Technology (ICCIT 2007). IEEE, 2007. https://ieeexplore.ieee.org/abstract/document/4420383

[10] Kim, Jawon, et al. "An Estimated Closeness Centrality Ranking Algorithm and Its Performance Analysis in Large-Scale Workflow-supported Social Networks," KSII Transactions on Internet & Information Systems, Vol. 10, No. 3, https://doi.org/2016.10.3837/tiis.2016.03.031

[11] BPI Challenge 2018, 4TU.Centre for Research Data, https://data.4tu.nl/repository/collection:event-logs-real.

◐ 저 자 소 개 ◑

Seonghun Ham

2019 B.S. in Computer Science, Kyonggi University
2019~Present, M.S. Student in Computer Science, Kyonggi University
Research Interests : Workflow systems, Hadoop system, discovery control flow, process mining, large-scale log analysis.
E-mail : shham9@kgu.ac.kr

Hyun Ahn

2011 B.S. in Computer Science, Kyonggi University
2013 M.S. in Computer Science, Kyonggi University
2017 Ph.D. in Computer Science, Kyonggi University
2018~Present, Assistant Professor of the Dept. of Computer Science and Engineering at Kyonggi University
Research Interests : Business Process Management, business process intelligence, process mining.
E-mail : hahn@kgu.ac.kr

Kwanghoon Pio Kim

1984 B.S in Computer Science, Kyonggi University
1986 M.S. in Computer Science, Chungang University
1994 M.S. in Computer Science, University of Colorado at Boulder
1998 Ph.D. in Computer Science, University of Colorado at Boulder
1998 ~ Present, Professor of the Dept. of Computer Science and Engineering at Kyonggi University
Research Interests : CSCW, workflow systems, Business Process Management, process mining, enterprise social network analysis.
E-mail : kwang@kgu.ac.kr