A System Architecture for Manufacturing Process Analysis based on Big Data and Process Mining Techniques

Hanna Yang, Minjeong Park, Minsu Cho, and Minseok Song
School of Business Administration
Ulsan National Institute of Science and Technology
Ulsan, South Korea
{hnyang, pmj9959, mcho, msong}@unist.ac.kr

Seongjoo Kim
Cyberdigm Co.
Seoul, South Korea
sjkim@cyberdigm.co.kr

Abstract— **Interests in manufacturing process management and analysis are increasing, but it is difficult to conduct process analysis due to the increase of manufacturing data. Therefore, we suggest a manufacturing data analysis system that collects event logs from so-called big data and analyzes the collected logs with process mining. There are two kinds of big data generated from manufacturing processes, structured data and unstructured data. Usually, manufacturing process analysis is conducted by using only structured data, however the proposed system uses both structured and unstructured data for enhancing the process analysis results. The system automatically discovers a process model and conducts various performance analysis on the manufacturing processes.**

Keywords—big data; process mining; text mining; Hadoop distributed file system (HDFS)

I. INTRODUCTION

Recently, several companies have been trying to manage and analyze big data generated from information systems, in order to optimize business processes [1-3]. In the manufacturing industry, process analysis receives attention from companies, since process optimization is directly connected to a decrease in production costs [4, 5]. There are two types of manufacturing data: structured data and unstructured data. Structured data has a pre-defined data format and usually comes from systems such as ERPs (Enterprise Resource Planning), HMIs (Human Machine Interface), MESs (Manufacturing Execution System), and various sensors in a factory. Data generated without a predefined format, and collected from systems such as e-mail systems, document management systems, etc. is unstructured data. While both structured and unstructured data contains important information about manufacturing processes, current researches on manufacturing process analysis usually use structured data [4, 5, 26], since it is difficult to extract and analyze process information from unstructured data. Furthermore, even though only structured data is considered for the process analysis, the increase of the size of data requires the capabilities to handle big data in the analysis.

In this paper, we propose a process mining based manufacturing data analysis system incorporated with big data. The proposed system extracts event logs from both structured and unstructured data, and analyzes the extracted logs with process mining. By utilizing process mining techniques, manufacturing processes can be analyzed in several ways. For example, process discovery techniques are used for discovering an actual manufacturing process model. Alpha algorithm [18], heuristic algorithm [23], and fuzzy algorithm [24] are frequently used process discovery techniques. Performance analysis techniques in process mining help analyzers detect bottlenecks of a process and assess the performance of activities, workers, and machines [17, 26-28]. Organizational mining techniques are used to derive a handover-of-work network for machines and a network for workers. In order to handle big data, the proposed architecture consists of the-state-of-the-art big data techniques such as NoSQL (Not only Structured Query Language), a Hadoop distributed file system, and distributed database technologies including HBase, MongoDB, etc.

The remainder of this paper is organized as follows. Section 2 describes related works, and we explain the proposed system architecture in Section 3. In Section 4, a prototype system based on the proposed system architecture is introduced. Finally, we conclude the paper in Section 5.

II. RELATED WORKS

A. Big Data

These days, big data handling techniques are evolving rapidly due to the growth of interests in big data and its value [2, 8]. Even though there are various definitions of big data, Garner and many others use the “3Vs” model [1, 3, 6, 7]. The 3Vs model explains big data with volume, velocity, and variety of data. According to the model, volume means the big scale of the data, velocity shows the rapid generation of the data, as well as the needs of the rapid data analysis, and variety represents the various types of the generated data. Therefore data with huge volume, high velocity, and great variety can be called big data [3, 6].

Among many techniques for handling big data, the proposed system is developed based on Hadoop, which is one of the most widely used big data techniques for distributed computation [9] in many industries [10, 11] and in academic fields [12]. Bahga and Madisetti [11] proposed a Hadoop based machine sensor data analysis framework for environment monitoring and failure forecasting. A Hadoop based system is used for genome data analysis and chemical structure data analysis in [12].

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

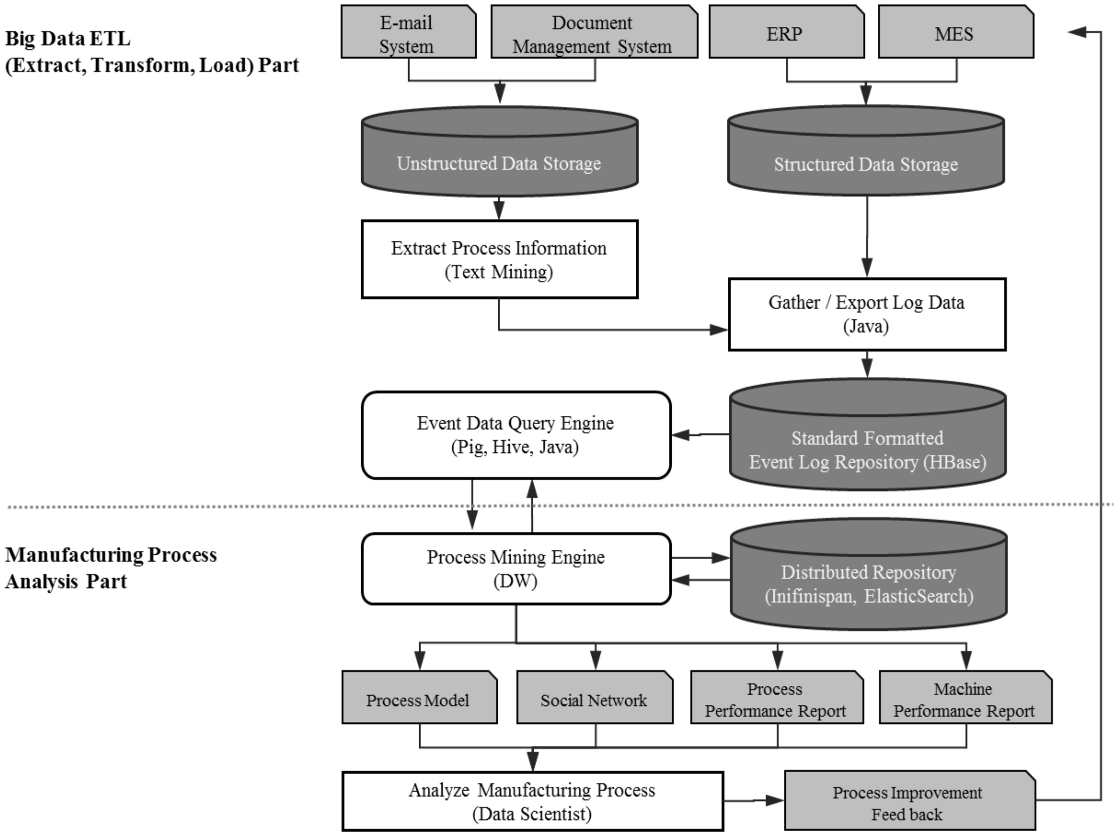

Big Data ETL
(Extract, Transform, Load) Part
E-mail System
Document Management System
ERP
MES
Unstructured Data Storage
Structured Data Storage
Extract Process Information (Text Mining)
Gather / Export Log Data (Java)
Event Data Query Engine (Pig, Hive, Java)
Standard Formatted Event Log Repository (HBase)
Manufacturing Process Analysis Part
Process Mining Engine (DW)
Distributed Repository (Inifinispan, ElasticSearch)
Process Model
Social Network
Process Performance Report
Machine Performance Report
Analyze Manufacturing Process (Data Scientist)
Process Improvement Feed back

Figure 1. A manufacturing data analysis system architecture

Database technology plays a key role in big data utilization [9], and the proposed system used HBase to build a log repository. HBase [15] is one of NoSQL databases that have become popular for big data handling. NoSQL databases are distributed key value databases which have advantages in managing and utilizing big data due to their flexible features [9]. Several NoSQL storage systems such as MongoDB [13], Dynamo system [14], Bigtable [16] and HBase have been developed and utilized in many big data management systems.

B. Process Mining

Process mining extracts meaningful information and knowledge from event logs which contain actual business process information [17-20]. By applying process mining, companies can discover useful information such as actual process models, handover-of-work networks among resources, performance measures of the processes, etc. [18, 21, 22]. Process model discovery is one of the most important process mining techniques [17], and numerous algorithms have been developed. Among them, alpha algorithm [18], heuristic algorithm [23] and fuzzy algorithm [24] are frequently used in practice. Organizational mining for discovering handover-of-work networks among resources is introduced in [25]. From the networks, process analysts can understand working relationships among resources, and figure out resources which play important roles in process executions. Performance analysis is conducted in several case studies [26-28]. In those case studies, performances (e.g. working and waiting time) of activities, workers and machines are measured and used for detecting problems of processes.

Process mining can analyze highly complex processes, so it is useful to analyze event logs from manufacturing processes which are usually complex because of many involving activities and resources. A few related case studies on manufacturing process analysis are introduced by Son *et al.* [4] and Rozinat *et al.* [5]. However, analyzing big size logs with process mining is computationally extensive and time consuming. In this context, we studied to integrate big data technologies with process mining for manufacturing process analysis. A system that we propose is presented in the next section.

III. Manufacturing Data Analysis System Based on Process Mining

In this section, we explain a manufacturing data analysis system based on process mining. The proposed system architecture is shown in Fig. 1. The system is divided into two parts: a big data ETL (Extract, Transform, Load) and a manufacturing process analysis. The big data ETL part extracts data from both structured and unstructured data sources and derives process related information. Then it transforms the information to event logs and stores them in an event data repository. The manufacturing process analysis part analyzes the extracted logs with process mining. The rest of this section describes each part in detail.

A. Big Data ETL (Extract, Transform, Load)

The big data ETL part extracts event logs from manufacturing data generated from legacy systems such as e-mail, document management, MESs, ERPs, etc., and it accumulates the logs in data storage. This part utilizes the high-capacity data storage structure of the Hadoop distributed file system and a distributed database (i.e. HBase and MongoDB) for supporting a parallel and distributed processing of big data (such as GB, TB data).

The process information extraction and the transformation of the extracted information are conducted separately for structured and unstructured data, because different techniques are used for different type of data.

Extracting process information from unstructured data is a complex process, and several modules are required for each step. First, we need a module for retrieving data (e.g. senders, recipients, the body of an e-mail) from the data storage and accumulating retrieved data into an analysis repository. In the repository, we need a module for recognizing important data properties which have to be transformed to features of event logs. These recognized properties become rules for the process information extraction.

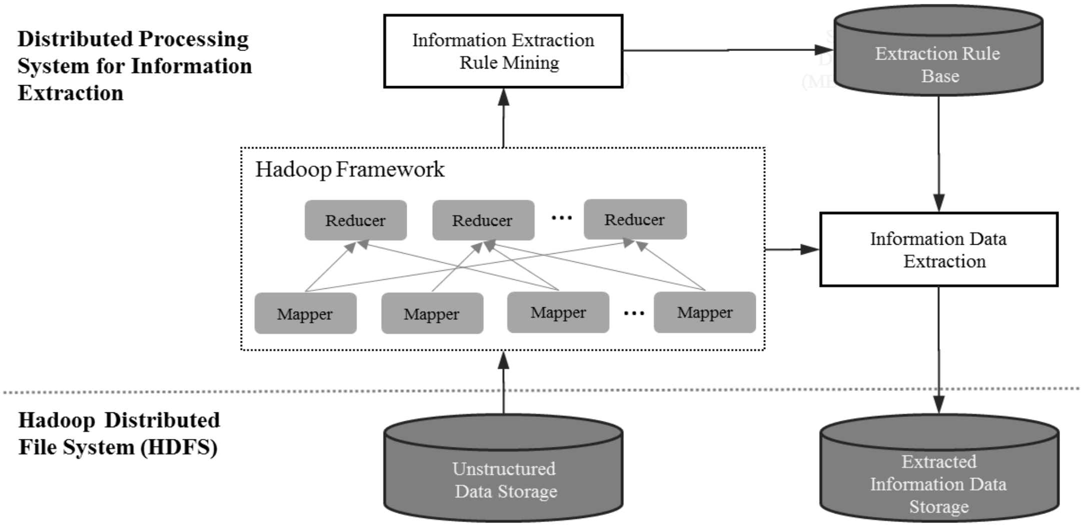
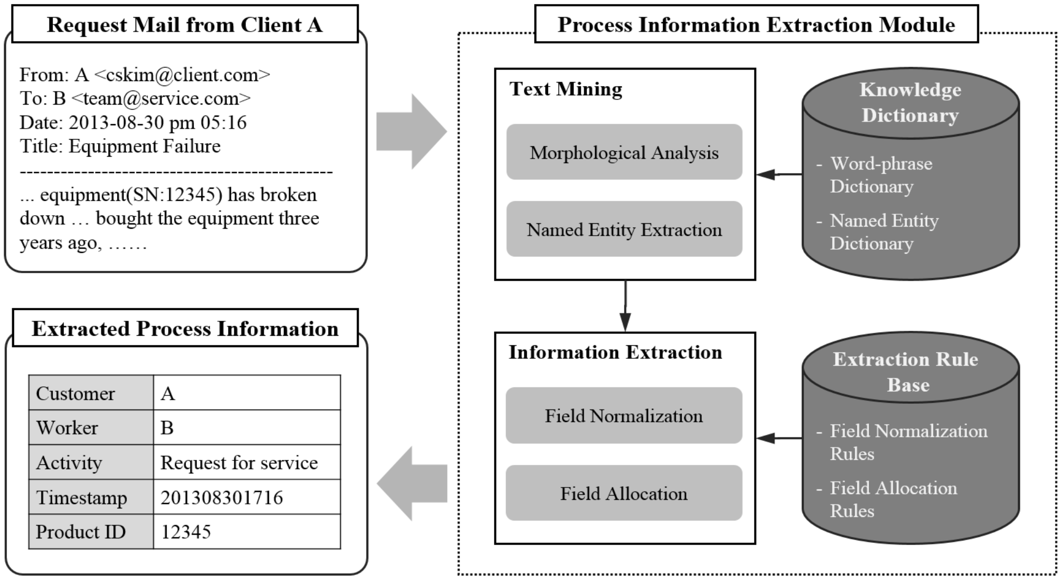

Distributed Processing System for Information Extraction
Information Extraction Rule Mining
Extraction Rule Base
Hadoop Framework
Reducer
Reducer
…
Reducer
Mapper
Mapper
Mapper
…
Mapper
Information Data Extraction
Hadoop Distributed File System (HDFS)
Unstructured Data Storage
Extracted Information Data Storage

Figure 2. A parallel and distributed processing framework for text mining



From: A <cskim@client.com>
To: B <team@service.com>
Date: 2013-08-30 pm 05:16
Title: Equipment Failure

... equipment(SN:12345) has broken down … bought the equipment three years ago, ……

Customer	A
Worker	B
Activity	Request for service
Timestamp	201308301716
Product ID	12345

Figure 3. An example of the process information extraction using text mining

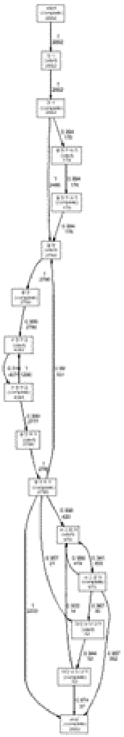


Figure 4. An example of a process model discovered by heuristic algorithm

In the proposed system, text mining is utilized with a distributed and parallel processing framework based on Hadoop, as shown in Fig. 2. The distributed and parallel processing system for extracting and integrating the process information uses Mapper and Reducer of the Hadoop framework. The Mapper is used for extracting the process information, and the Reducer is used for integrating the information. The collected process information is stored into the distributed file system of the Hadoop framework, then the stored data is transformed to event logs.

For text mining, a morphological analyzer that can recognize information, and a named entity recognizer that tags semantic information, are utilized. The morphological analyzer consists of three modules: (1) a word segmentation module which divides words into morphemes (2) a pre-analyzed word processing module which contains pre-analyzed results of frequently used words, and (3) a word-phrase dictionary which is used for analysis and combining of morphemes. The named entity recognizer also consists of three modules: (1) a LSP (lexico-semantic pattern) generation module which generates patterns for named entity recognition by tagging morphological analysis results, (2) a center extraction module which derive centers from sentences, and (3) a pattern checking module which improves grammar accuracy with defined patterns.

An example of the process information extraction through text mining is shown in Fig. 3. The data in the upper-left of Fig. 3 is an example of a repair request e-mail, and the data in the lower-left of the figure is extracted process information. A customer is the sender of the e-mail and a worker is the receiver. An activity name is determined using the information extracted by text mining, and a timestamp is same as a time and date that the customer sent the e-mail. Note that other information, such as a product ID, can be extracted and transformed as attributes of the event logs for additional process analysis.

In order to extract process information from structured data and transform the information to event logs, an event log collector is essential. The event log collector requires readable and meaningful logging policies to collect appropriate event logs which contain appropriate process information for applying process mining. The policies are decided by assessing the type and the level of event logs that can be extracted from structured data. The log collector of the proposed system uses a centralized collecting method. It means that the event logs are collected from the data of each information system (e.g. MES, ERP, HMI) through a batch job scheduled by the log collector server. Then, the collected logs are stored in the standard formatted event log repository as integrated event logs with the logs extracted from unstructured data.

B. *Manufacturing Process Analysis*

The manufacturing process analysis part contains a process mining module which analyzes the integrated event logs collected by big data ETL part. In this part, analysis is conducted in a distributed file system based repository to deal with the big size logs. The analysis repository is constructed by using technologies from ElasticSearch [29] and Infinispan [30]. ElaticSearch is an open-source distributed text search engine, and Infinispan is an open-source in-memory data store and in-memory transactional grid platform. Furthermore, a web-based user interface is required for data scientists who use the management data analysis system, select process mining techniques according to the purpose of analysis, and interpret process mining results.

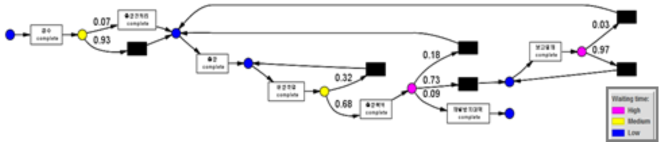


Figure 5. An example of bottleneck analysis result

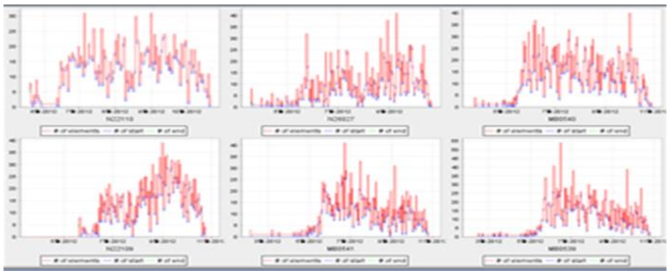


Figure 6. An example of operation rate analysis results

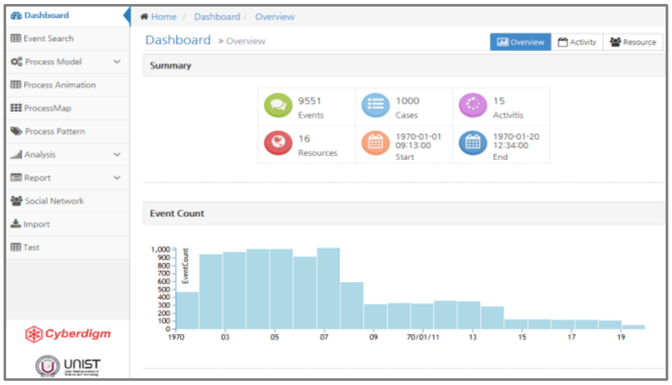
Figure 7. An example of a handover-of-works network

Various mining techniques are applicable for manufacturing process analysis. First, a manufacturing process model can be discovered from the integrated event logs. Several process model discovery algorithms, including fuzzy algorithm, heuristic algorithm and α-algorithm, are available. Therefore, a data analyst needs to understand the logs and select modeling algorithms which are most appropriate for the analysis purpose. An example of a process model discovered by using heuristic algorithm is illustrated in Fig. 4. By observing the process model, we can understand the actual manufacturing process. Moreover, we can detect abnormal working behaviors by comparing the model with the reference process model (i.e. the pre-existing process models that are supposed to be followed by resources).

Activity performance, including the manufacturing time, waiting time and lead time of each activity, can be assessed by using the performance analysis techniques of process mining. We can detect problems in the process, such as activities with long waiting times or working times, and conduct cause analysis for these.

Based on the process model and activity performance analysis results, we can conduct a bottleneck analysis. We assess the average sojourn time of each place of a Petri net, and then compare it to a threshold which needs to be decided by the data analyst. The average sojourn time of a place is the average waiting time of all following activities connected to the place. The bottlenecks are expressed by color. Fig. 5. shows an example of a bottleneck analysis result, and the red-colored places indicate bottlenecks in the process.

Operation rate analysis for resources is available. It measures the workload of resources (i.e. machines and workers) in units of time (e.g. an hour), and analyzes whether the resources work enough, and whether or not they are efficient. We can compare the operation rates of resources which work for the same manufacturing activity to optimize resource allocation. Fig. 6. shows an example of results of operation rate analysis for machines, and each small graph shows an operation rate of a corresponding machine.



Summary

9551 Events

1000 Cases

15 Activitis

16 Resources

1970-01-01 09:13:00 Start

1970-01-20 12:34:00 End

Figure 8. A screenshot of the prototype main page

With organizational analysis, we can derive handover-of-work networks of resources. The network helps analysts understand relations among resources, and detect the important resources which involve many types of work and connects with other resources. Fig. 7. is an example of a handover-of-work network of manufacturing machines.

IV. PROTOTYPE

The manufacturing data analysis system proposed in the previous section is currently under development. It is in a prototype state, which will be briefly introduced here. Fig. 8 is a screenshot of the prototype main page.

When an event log is uploaded into the system, the prototype automatically provides basic log information, such as an event log summary and event count statistic (i.e. the number of events, cases and activities, a start timestamp of the first case, and a complete timestamp of the last case). Algorithms for discovering a manufacturing process model and playing a process animation were implemented in the prototype. The animation is played based on the discovered process model. Each case is expressed as a dot which moves through the process model, and the timeframe of the animation can be adjusted by users. We also implemented an activity-specific performance analysis for measuring the working and waiting time of each manufacturing activity, and a case-specific performance analysis for measuring the total execution time of each case. In addition, an organizational analysis for discovering a handover-of-work network is available.

V. CONCLUSION

In this paper, we proposed a manufacturing data analysis system that extracts event logs from different types of data and analyzes them with process mining. To handle big data, the proposed system utilizes several big data technologies. The system also utilizes several process mining techniques to analyze manufacturing processes. To validate our approach, a prototype system is implemented. For future works, the system should be fully implemented to validate our approach. Furthermore a case study with real life data should be performed. In this paper, we only considered five process mining techniques for manufacturing process analysis. Thus it is necessary to investigate the applicability of other process mining techniques to analyze logs from manufacturing companies. More sophisticated process mining techniques specialized for analyzing manufacturing processes should be developed.

ACKNOWLEDGMENT

This work was supported by the Industrial Core Technology Development program (10045047) and Technology Diffusion Platform for C-MES project (10033159) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

REFERENCES

[1] P. Zikopoulos, and C. Eaton, “Understanding big data: analytics for enterprise class hadoop and streaming data,” McGraw-Hill Osborne Media, October 2011.

[2] J. Manyika, McKinsey Global Institute, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big data: the next frontier for innovation, competition, and productivity,” McKinsey Global Institute, May 2011.

[3] D. Laney, “3-d data management: controlling data volume, velocity and variety,” META Group Research Note, February 2001.

[4] S. Son, B. N. Yahya, M. Song, S. Choi, J. Hyeon, B. Lee, Y. Jang, and N. Sung, "Process Mining for Manufacturing Process Analysis: A Case Study," Asia Pacific Business Process Management Conference (APBPM2014), Brisbane, Australia, July 2014.

[5] A. Rozinat, I. S. M. de Jong, C. W. Günther, and W. M. P. van der Aalst, “Process Mining Applied to the Test Process of Wafer Scanners in ASML,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (TSMC), vol 39, no. 4, July 2009, pp. 474-479, doi: 10.1109/TSMCC.2009.2014169.

[6] M. Beyer, “Gartner says solving big data challenge involves more than just managing volumes of data,” Gartner, 2011.

[7] E. Meijer, “The world according to linq,” Communications of the ACM, vol. 54, no. 10, August 2011, pp. 45–51.

[8] S. Lohr, “The age of big data,” New York Times, February 2012, pp. 11.

[9] M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” Mobile Networks and Applications, vol. 19, no. 3, Aprill 2014, pp. 171–209, doi: 10.1007/s11036-013-0489-0.

[10] J. Dean, and S. Ghemawat, “Mapreduce: simplified data processing on large clusters”. Communications of the ACM, vol. 51, no. 1, January 2008, pp. 107–113, doi: 10.1145/1327452.1327492.

[11] A. Bahga, and V. K. Madisetti, “Analyzing massive machine maintenance data in a computing cloud,” IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 10, October 2012, pp. 1831–1843, doi: 10.1109/TPDS.2011.306.

[12] T. Gunarathne, T-L. Wu, J. Y. Choi, S. H. Bae, and J. Qiu, “Cloud computing paradigms for pleasingly parallel biomedical applications,” Concurrency and Computation Practice and Experience, vol. 23, no. 17, June 2011, pp. 2338–2354, doi: 10.1002/cpe.1780.

[13] K. Chodorow, “MongoDB: the definitive guide,” O’Reilly Media Inc, 2013.

[14] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, “Dynamo: amazon’s highly available key-value store,” ACM SIGOPS Operating Systems Review - SOSP '07, vol. 41, no. 6, December 2007, pp. 205–220, doi: 10.1145/1323293.1294281.

[15] L. George, “HBase: the definitive guide,” O’Reilly Media Inc, August 2011.

[16] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, “Bigtable: a distributed storage system for structured data,” ACM Transaction on Computer Systems (TOCS), vol. 26, no. 2, June 2008, pp. 4:1-4:26, doi: 10.1145/1365815.1365816.

[17] W. M. P. van der Aalst, “Process mining: overview and opportunities,” ACM Transaction Management Information System (TMIS), vol. 3, no. 2, July 2012, pp. 7:1-7:17, doi: 10.1145/2229156.2229157

[18] W. M. P. van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. van Dongen, A. K. A. de Medeiros, and M. Song, “Business process mining: an industrial application,” Information Systems, vol. 32, no. 5, July 2007, pp. 713–732, doi: 10.1016/j.is.2006.05.003.

[19] W. M. P. van der Aalst, A. J. M. M. Weijters, and L. Maruster, “Workflow mining: Discovering process models from event logs,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 9, September 2004, pp. 1128–1142, doi: 10.1109/TKDE.2004.47.

[20] IEEE Task Force on Process Mining, “Process Mining Manifesto,” Proc. Business Process Management Workshops (BPM 2011), Springer-Verlag GmbH Berlin Heidelberg, vol. 99 of Lecture Notes in Business Information Processing, September 2007, 2012, pp 169-194, doi: 10.1007/978-3-642-28108-2_19.

[21] M. Bozkaya, J.M.A.M. Gabriels, and J.M.E.M. van der Werf, “Process Diagnostics: A Method Based on Process Mining,” Proc. International Conference on Information, Process, and Knowledge Management (eKNOW), IEEE Press, February 2009, pp. 22-27, doi: 10.1109/eKNOW.2009.29.

[22] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker, “Process mining in healthcare – a case study,” Proc. international conference on health informatics (HEALTHINF’08), INSTICC Press, January 2008, pp. 118–125.

[23] A. J. M. M. Weijters, and W. M. P. van der Aalst, “Rediscovering workflow models from event-based data using little thumb,” Integrated Computer-Aided Engineering, vol. 10, no. 2, April 2003, pp. 151-162.

[24] C. W. Günther, and W. M. P. van der Aalst, “Fuzzy mining - adaptive process simplification based on multi-perspective metrics,” Proc. International Conference on Business Process Management (BPM 2007), Springer-Verlag GmbH Berlin Heidelberg, vol. 4714 of Lecture Notes in Computer Science, September 2007, pp. 328-343, doi: 10.1007/978-3-540-75183-0_24.

[25] W. M. P. van der Aalst, H.A. Reijers, and M. Song, "Discovering Social Networks from Event Logs," Computer Supported Cooperative Work, vol. 14, no. 6, December 2005, pp. 549-593, doi: 10.1007/s10606-005-9005-9.

[26] S. Lee, K. Ryu, and M. Song, "Process Improvement for PDM/PLM Systems by Using Process Mining," Transactions of the Society of CAD/CAM Engineers, vol. 17, no. 4, August 2012, pp. 294-302, doi: 10.7315/CADCAM.2012.294.

[27] E. Kim, S. Kim, M. Song, S. Kim, D. Yoo, H. Hwang, and S. Yoo, “Discovery of Outpatient Care Process of a Tertiary University Hospital Using Process Mining,” Healthcare Informatics Research, vol. 19, no. 1, March 2013, pp. 42-49, doi: 10.4258/hir.2013.19.1.42.

[28] D. Jeon, B. N. Yahya, H. Bae, M. Song, S. Sul, and R. A. Sutrisnowati, "Conceptual Framework for Container-handling Process Analytics," ICIC Express Letters, vol. 7, no. 6, June 2013, pp.1919 - 1924.

[29] O. Kononenko, O. Baysal, R. Holmes, and M. W. Godfrey, “Mining modern repositories with elasticsearch,” Proc. Working Conference on Mining Software Repositories (MSR 2014), ACM New York, May 2014, pp. 328-331, doi: 10.1145/2597073.2597091.

[30] F. Marchioni and M. Surtani, “Infinispan Data Grid Platform,” Packt Publishing, August 2012.