International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Big Data Quality: Early Detection of Errors in Process Flow using Alignments and Compliance Rules

Melody Wadzanayi Murakwani[1], Manoj Sethi[2]

[1]Delhi Technological University, Department of Computer Engineering, Delhi-110042, India

[2]Delhi Technological University, Department of Computer Engineering, Delhi-110042, New Delhi, India

Abstract: ***In our Big Data era, data is being produced at scale, in motion, and in heterogeneous forms. Uncertainty is another significant attribute exhibited by this data and hence there is need to comprehend and (perhaps) repair erroneous data timely. Due to heterogeneity of data source and usage, data quality rules are contextual; hence we require data management solutions that acknowledge these varied uses and incorporate them to determine the required level of quality and standardization. Today, there is a wide range of process mining techniques that are able to uncover the reality of processes through a systemic analysis of event data. These techniques are being applied in this work with the aim to isolate the source of the introduction of data flaws to fix the process instead of correcting the data. This paper employs the Heuristic Miner algorithm for process discovery, Petri nets with data (DPN nets) and conformance checking using alignments and compliance rules. We showed that alignments between event logs and the discovered Petri Net from process discovery algorithms reveal frequent occurring deviations and compliance rules are an effective data management solution. Insights into these deviations are then exploited to repair and enhance the original process models. Our novel diagnostic data-aware process discovery technique is applied on a real-life event log and evaluated for its success in providing new and valuable insights and failure in other areas of performance.***

Keywords: Big Data Quality, Process Mining, Alignments, Compliance Rules

1. Introduction

While the potential and promise of Big Data is real, for instance, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 [1], recent studies show that poor quality data issues widen the gap between its potential and its realization. Erroneous data is estimated to cost US businesses 600 billion dollars annually [2]. The level of inaccurate data is a cause for concern when one considers how much research and industry are relying on information for analytics and business. There is an increasing demand to improve data quality so as to add accuracy and value to research and industry alike.

The thrust for this work is to isolate the source of the introduction of data flaws to fix the process instead of correcting the data, thereby reducing data cleaning costs. Big data characteristics include heterogeneity and veracity, and veracity is a big data characteristic which directly refers to inconsistency and data quality problems [3]-[6]. Process mining as a new and promising research field has shown strong capabilities to mine knowledge from data as it applies both process modeling and data mining techniques to discover models from the event logs. By leveraging IT footprints, process mining attempts to create a realistic picture of the process as it actually takes place, and as a consequence enables targeted adjustments to improve the performance or compliance of the process. The gained transparency of what is actually going on is a huge value in itself. Moreover, knowledge of the current status is also a prerequisite for any improvement actions, because one can only improve what they can measure.

The greater chunk of research in the field of process mining focuses on the control-flow aspects of a process and the data-flow perspective is not awarded much attention. The control-flow perspective presents the order of process' activities, but ignores data movements within the process. This research adopts data-aware process mining introduced in [14] to show that the data-flow perspective coupled with resource network analysis and conformance checking using compliance rules may bring improvements to the process.

This data driven fault detection enabled by use of process mining techniques will enable us to do diagnostic and enhancement of existing processes towards data quality improvement. This work is inspired by Massimiliano de Leoni et al in [14] where they use recent advances in conformance checking using alignments, and they also employed the use of Petri Nets with Data, in their paper: Data-Aware Processing Mining: Discovering Decisions in Processes Using Alignments.

2. Literature Review

The literature accounts for different ways to capture the data-flow perspective of a process, but most of ways employ a data modeling approach such as ERDs as opposed to our process data perspective, not all approaches discover the model from event logs. In addition, none of the existing approaches combine both alignments and compliance rules for error detection in big data. This section reviews previous research that approached the process data perspective.

Our review of related work focuses on the approaches directly relevant to this paper, particularly those in which both the control-flow and data-flow are considered for analysis. The significance of data-flow verification in process workflows was introduced in [7], [8]. In [7], several possible errors in the data-flow are identified, such as the

missing and redundant data error, but the means to check these errors is not provided. Afterward, [9] gave a method to check the errors from [7] using UML diagrams, and gave supporting verification algorithms. None of these approaches consider control-flow properties.

In [10], a model that uses dual workflow nets is proposed, that can describe both the data-flow and the control-flow. However, no explicit data correctness properties are considered. In [11], model checking is used to verify business workflows, from both control- and data-flow perspective. The underlying workflow language is UML diagrams as opposed to the Petri net approach taken in this paper. Only a few data correctness properties are identified and no systematic classification is presented. [12], presented an approach to detect data errors in workflow using a systematic graph traversal approach. The work closest to this work is in [13], and [14]. In [13], Nikola Trocka et al presented an analysis approach that uses so-called "anti-patterns" expressed in terms of a temporal logic to discover data-flow errors in workflows. In [14] Massimiliano de Leoni et al use recent advances in conformance checking using alignments, and they also employed the use of Petri Nets with Data.

3. Implementation

Our proposed approach aims to use Petri nets with data (DPN-net) for modeling workflows, followed by conformance checking using compliance rules and alignments. Most analysis in this work is performed using existing and dedicated plug-ins within the open-source process mining toolkit, ProM (version 6.4), performance related experimentation was conducted in DISCO and additional result analysis was done in MS Excel.

This work is performed on a real life event log in XES format obtained from 3TU.Datacentrum. 3TU.Datacentrum is an educational repository for archiving scientific data at the University of Twente, Netherlands. The BPI Challenge 2012 dataset and its metadata were sourced at http://data.3tu.nl/repository/uuid:3926db30-f712-4394-aebc-75976070e91f.

Table 1: Metadata for event log

Name	*Description*
Doi	10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f
Name	BPI Challenge 2012.xes.gz
Description	Event log of a loan application process
Language	Dutch
log type	Real-life
process-type	Explicitly structured
source institute	Eindhoven University of Technology
rights type	Public
# of traces	13087
# of events	262200

3.1 Discovering the control-flow perspective with Heuristic Miner Algorithm

The control-flow perspective of a process establishes the dependencies among its tasks, such as which tasks precede which other ones or whether there are any loops in the log. In this work we employ the use of Petri Nets for model representation. A Petri net consists of places represented using circles with a start and an end point and transitions represented using rectangles. Transitions may be connected to places and places may be connected to transitions, as shown in the Petri net for our data set in Figure 1 below

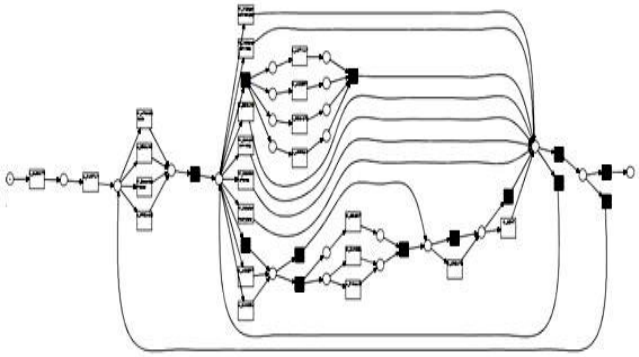

Figure 1: Petri Net Observed from the event log

3.2 Compliance Requirements

Compliance requirements are employed to ensure that all necessary data governance requirements are met without making the system vulnerable through unnecessary duplication of activities and effort from resources. They control one or several process perspectives such as the data flow, process time, control flow, or organizational aspects. Restrictions may be imposed for individual cases or groups of cases; they can prescribe properties of process executions or process design [15].

Data-Aware and Resource-Aware Compliance Rules

Elham Ramezani et al in their work [15] compiled a list of compliance rules, which they also adopted for conformance checking. Below is a collection of compliance rules employed in this work from the compilation in [15].

Table 2: An extract of data-aware and resource-aware compliance rules

Compliance requirement	Adopted example
Four-eye principle: A security principle that segregates privileges and associates the execution of critical tasks to groups of users.	The person dealing with offer processing of a loan should not be the one who approves the same.
Authorization (Access control): A security principle that ensures only authorized individuals perform certain activities or access certain data objects.	Only a 2nd level Administrator or supervisor can approve a loan.
Two (three)-way match: This is an accounting rule which states that the value of two different related data objects should match.	All customer invoices for purchases should be matched with respective purchase order lines

Activity L may only be executed if the value of attribute P is greater than or equal to v	An application must not be approved for processing in case risk is high.

3.3 Conformance Checking (Aligning event logs and process models)

In this step we align the event log and control-flow process model, i.e., events in the log need to be related to transition executions.

Conformance Analysis Result

The conformance analysis plug-in is employed for this purpose, the plug-in takes two inputs: - the discovered model which is a Petri Net in this work and the event log and returns a Petri Net showing statistics of
1) synchronous moves between the log and the model,
2) moves on the model only
3) moves that appear in the log only
4) frequency of moves in particular events
5) frequency of moves between particular events

3.4 Discovering the Data-Flow Perspective

The data flow perspective takes a Petri net (P, T, F) and applies read/write variables captured at each activity using the Data-flow Discovery plug-in in ProM. The plug-in takes two objects as input: a Petri net, the read/write variables captured by each activity. It returns a Petri net with data where the read and write operations show the data flowing through the process.

The outcome is a Petri net with data N = (P; T; F; V; U; R; W; G), where our technique mines V; U; R; W and G. In the remainder, we say an event, $(t,\phi) \in L$ if there exists a trace $\sigma \in L$ such that $(t,\phi) \in \sigma$. We reasonably assume based on the work done in [14] that the set of variables of N are the set of variables defined in the event logs. For each event we assumed write and read operations given the data objects captured by the system. We will start by defining the variables through detailing the data objects captured by the system at various events.

The variables indicated in Table 3 are a subset of variables actively read or written to in the process under investigation.

Table 3: Definition of variables

Variable	*Type*
Amount	Non-negative number
Status	Boolean
valid_Until	Date
Decision	Boolean

Table 4: Read/Write Operations

Transition	Variable Read	Variables Written
A_ACCEPTED	amount, valid_Until	Status
A_APPROVED	valid_Until,status,amount	Decision
A_CANCELLED	valid_Until,status,amount	Decision
A_DECLINED	valid_Until,status,amount	Decision
O_CANCELLED	status, amount	Decision
O_ACCEPTED	Amount	valid_Until, status

3.5 Petri Net with Data

Figure 8 below shows a screenshot of the Petri net with data: the output of the Data-Flow Discovery plug-in in ProM. The white and black rectangles identify the visible and invisible transitions. The yellow & rounded circles represent the variables defined in the process data-flow. The dotted/faint arrows going in and out the yellow circles describe the write and read operations.

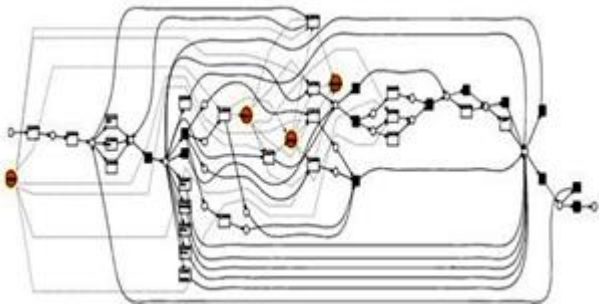

Figure 2: Petri Net with data

4. Results and Analysis

4.1 Observations

1. 6,8% of the activities are performed by an unknown resource.
2. Resource named 112 is an automated resource because
   i. Only resource 112 performs initial activities such as A_SUBMITTED and A_PARTLYSUBMITTED in all 13087 cases.
   ii. And this resource does not handle work or offer (manual) activities outside of initiating application activities.
3. Resources named 10138, 10609, 10809, 10972 and 10629 seem as though they are super user resources, since they are active in approval and/or cancellation of applications.

4.2 Analyzing the Resource Perspective

This step involves the analysis of the resource perspective against the defined compliance rules. The event log contains 69 distinct resources who have worked on at least one case in the log. 68 resources have a resource id while the 69th resource is an unknown resource.

Resource-Aware Compliance Checking

Result and observation - Rule 1: Four-eye principle: *A security principle that segregates privileges and associates the execution of critical tasks to groups of users.*

Resource 10138, 10609, 10809, 10972 and 10629 are violating the four-eye principle, i.e these resources are involved in almost all loan handling activities from request to approval in the loan application process.

Analysis – *Rule 1: Violation of the Four-eye principle*
This is a strong indication that there is no strict separation of roles and consistent adherence to the four-eye principle to ensure compliance. The four-eye principle is a core perspective of Information governance. Information

Governance is a key internal control measures to mitigate introduction of error.

Result and Observation: ***Rule 2- Authorization (Access control):*** *A security principle that ensures only authorized individuals perform certain activities or access certain data objects.*

Observation 1: Resource 112 violates Rule 2 as it is an automated response or a system but it has approved 3 loans, as shown by Table 5 below

Table 5: Resource against frequency of their decision on loan applications

Resource	A_APPROVED	A_CANCELLED	A_DECLINED
10138	681	5	156
10609	335	5	206
10629	359	1	119
10809	271	1	87
10972	518	3	106
112	**3**	1004	3429
11289	68	3	55

Observation 2: Missing resource information. An unknown resource performs several work related items in a number of cases.

Analysis: ***Rule 2- Authorization (Access control):*** *A security principle that ensures only authorized individuals perform certain activities or access certain data objects.*
In both Observation 1 and Observation 2 resources violate the principle of least privilege in the access control policy. The principle of least privilege refers to the practice of granting subjects access only to what they need to perform their jobs and no more. Security violations that threaten access control such as the one exhibited by resource 112 and the unknown resource may result in Trojan Horses or viruses being implanted whose activity may not be triggered long after the original event.

4.3 Analyzing the data-flow perspective

Analyzing the data-flow perspective considers the milestones in our data-aware diagnostic model. The analysis is a triple feature initiative which incorporates:-
1) A snapshot of the performance analysis result (figure 3),
2) The alignment result from conformance analysis in the top right corner (snapshot of result in figure 4).
3) The Petri net with data (figure 2).

Observation 1 – Transition frequency: - In the results captured in figure 3 above, event *W_Nabellen offertes*(start and complete) which involves following up after transmitting offers to qualified applicants is the most frequently traversed transition.

Analysis- Transition frequency
The high traffic between W_Nabellen offertes start and W_Nabellen offertes complete might have a negative impact as resources may evade decision-support defences that the system provides during eventful periods. If the existing process is not flexible for offer processing for instance, it often causes adjustments in workflows. These alternate workflows introduce challenges because they increase deviation from routine sequences.

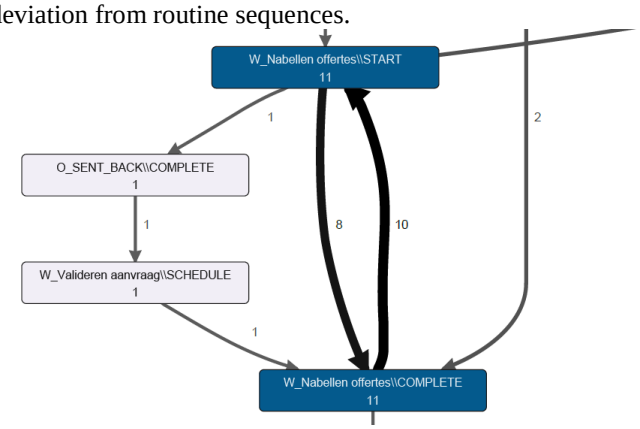

Figure 3: Activity performance snapshot diagram

Observation 3- Non-alignment of events – In the alignment result zoomed in figure 4 below we observe non-conformance at event O_ACCEPTED, the red border on O_ACCEPTED indicates that out of the 2248 moves recorded in that event, 3 moves that were seen in the model were not present in the log.

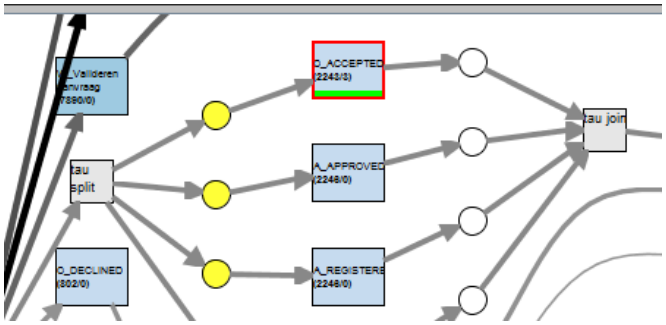

Figure 4: Snapshot of Conformance Analysis result

Analysis – Non-Conformance
Non conformance is failure to adhere to requirements which directly impacts internal and external costs of a system. The system data used in this work shows only internal costs. Internal failure shows through deficiencies which may be caused by inefficiencies in processes. The internal costs of non-conformance include repeating work, delays, re-doing designs, shortages, failure investigation, verification, inflexibility and malleability.

5. Validation

Using our data-aware diagnostic method for early fault detection, we have successfully noted entry points of dirt in data as influenced by the workflow. Next, we evaluate the approach using performance and conformance analysis statistics before applying the method and after applying the method to the real-life data log. To validate our method we used information observed in the results above to clean, filter and improve the data and we compared the result to the initial raw data result.

5.1 Conformance analysis statistics

The statistics obtained from conformance analysis furnish us with details regarding the raw fitness cost of the alignment, calculation time in milliseconds, move-model fitness, trace fitness, move-log fitness, and trace length. Below is a comparison of these results followed by an evaluation of a filtered log against the raw log.

a. Calculation Time Statistics

These statistics capture the performance of the conformance checker algorithm in terms of time taken to replay the log on the model. In general, calculation time for the filtered log is lower than calculation time for the original log. However, as shown in figure 5 below the average calculation time for the filtered log is 43.4% higher than that of the original log. This is an indication that the algorithm performed poorly on the filtered log compared to the original log. It is possible that our filtering has brought about complications in the control flow that make average calculation time per case high in the filtered log.

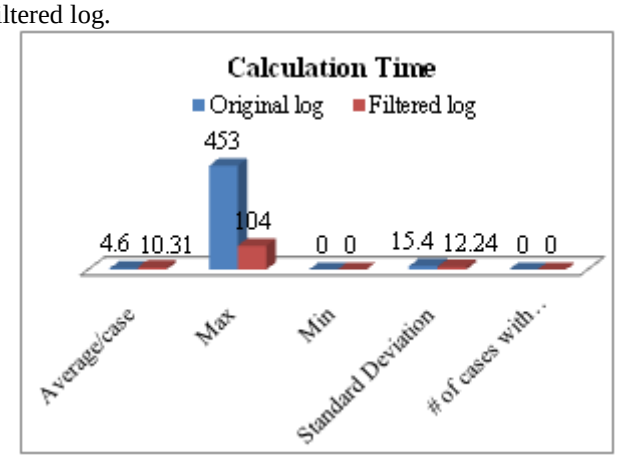

Figure 5: Calculation Time Statistics

b. Trace Fitness

Trace fitness is investigating whether a process model is able to reproduce all execution sequences that are in the log at the case level, i.e., the fraction of traces in the log which can be replayed fully. The filtered log shows lower trace fitness compared to the original log for all 3,429 traces, it is possible that the changes made to the log had a negative impact on the control flow of events.

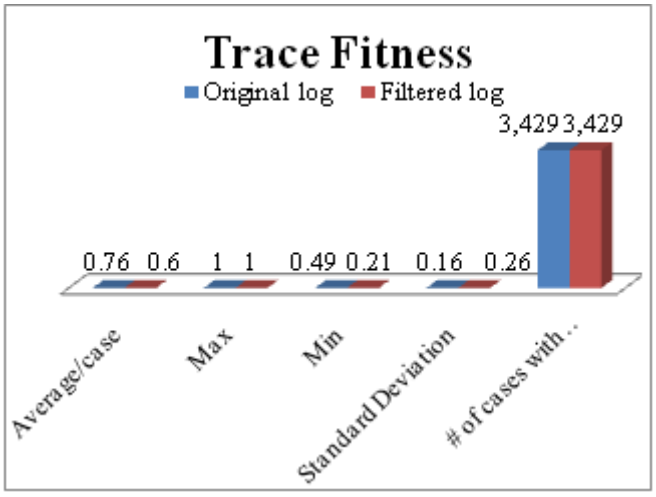

Figure 6: Trace fitness statistics

c. Move-log fitness

In move-log fitness the algorithm moves a step in the log while the model remains in position and shows if there is disparity between the experiential activities in the traces and the achievable activities in the log. For the 3,429 cases, the original log has higher fitness compared to the filtered log.

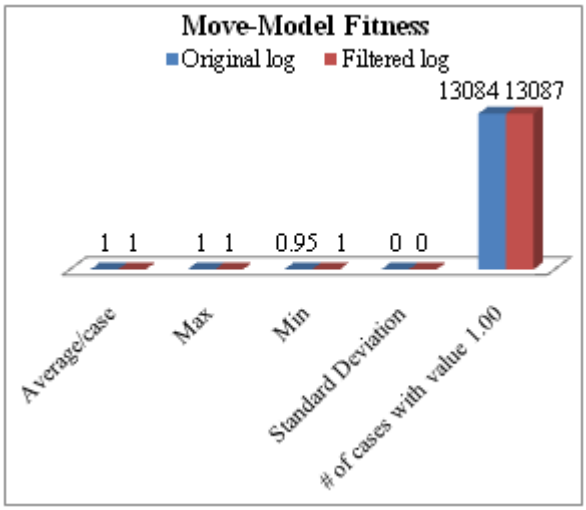

Figure 7: Move-Model fitness statistics

d. Overall Conformance Analysis statistics

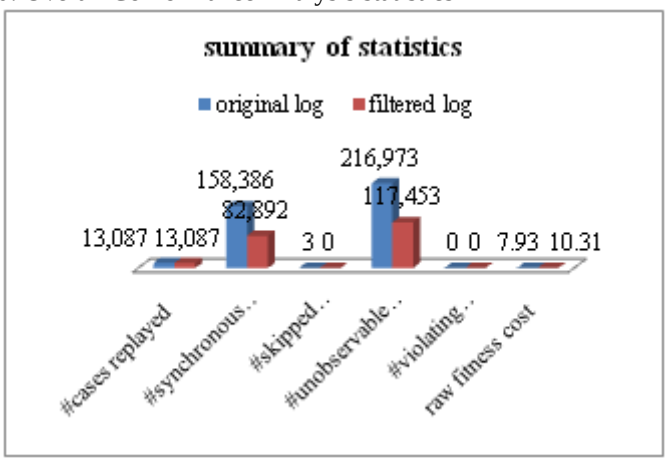

Figure 8: Summary of statistics

Figure 8 above shows a summary of statistics evaluated by conformance analysis. The attributes captured for evaluation are explained below:

1) *Synchronous event class (log + model)* -this means both an activity in the model as well as an activity in the current trace can be 'moved' without misalignment.
2) For the 13,087 cases replayed for both the original log and the filtered log, the original log has a higher synchronous event class (log+model) fitness value compared to our filtered log which shows that the overall alignment of the original log is better compared to the filtered log.
3) *Skipped events classes* – the original log recorded a total of 3 event classes seen in the model but not captured in the log; in the filtered log no event classes were skipped.
4) *Violating synchronous event class* – no log+ model activities were noted in both logs that occur in the log without taking any obligations (i.e. tokens) in the model.
5) *Raw fitness cost statistics* - A model with good fitness captures most of the behavior observed in the event log. Fitness is expressed as a value between 0 which represents

very poor fitness and 1 for perfect fitness. Our result showed that the average fitness/case increased and standard deviation also increased to show an increase in the variation of cases in the log.

6. Conclusion and Future Work

In this paper, we applied data aware-process mining as a process-data centric approach to improve data quality in the early life cycle stages of big data applications by remodeling the process to capture data more accurately. Data-aware process mining is an approach presented by M. de Leoni at al. in [14]. We applied their approach and improved it by adding compliance rules for conformance checking.

The proposed method has been implemented using existing plug-ins in ProM and evaluated on a real-life event log obtained from the Dutch Financial Institute. The evaluation of our method for early detection of errors in process flow using Petri nets with data was successful in providing new and valuable insights in detecting introductory points of error. However, our filtering method also failed in other areas of performance as indicated by calculation time statistics, and synchronous event classes' statistics.

In future work, this team would like to implement the suggested the diagnostic data-aware detection algorithm as a Prom plug-in which is an integrated algorithm for checking the alignment of a Petri Net with data and finally generating a verification report for the workflow designer.

References

[1] Alexandros Labrinidis, H.V Jagadish, "Challenges and Opportunities with Big Data", Proceedings of the VLDB Endowment, Istanbul, Turkey, 2012.
[2] Eckerson W, "Data quality and the bottom line: Achieving business success through a commitment to high quality data", The Data Warehousing Institute, 2002.
[3] Barna Saha, Divesh Srivastava, "Data Quality: The other Face of Big Data", ICDE Conference, 2014.
[4] Mark Madsen, "The Challenges of Big Data & Approaches to Data Quality", Third Nature Inc, 2013.
[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey Global Institute, 2011.
[6] Michael Schroeck, et al, "Analytics: The real world use of big data", 2012. Available: http://www-ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf [Accessed: January 21, 2015].
[7] S.W. Sadiq, M.E. Orlowska, W.Sadiq, and C. Foulger, "Data Flow and Validation in Workflow Modelling", In Fifteenth Australasian Database Conference (ADC), Dunedin, New Zealand, volume 27 of CRPIT, pp 207–214, Australian Computer Society, 2004.
[8] S.X. Sun, J.L. Zhao, J.F. Nunamaker, and O.R. Liu Sheng, "Formulating the Data Flow Perspective for Business Process Management", Information Systems Research, 17(4):374–391, 2006.
[9] S.W. Sadiq, M.E. Orlowska, W.Sadiq, and C. Foulger, "Data Flow and Validation in Workflow Modelling", In Fifteenth Australasian Database Conference (ADC), Dunedin, New Zealand, volume 27 of CRPIT, pp 207–214, Australian Computer Society, 2004.
[10] S. Fan, W.C. Dou, and J. Chen, "Dual Workflow Nets: Mixed Control/Data-Flow Representation for Workflow Modeling and Verification", In Advances in Web and Network Technologies, and Information Management (APWeb/WAIM 2007 Workshops), volume 4537 of Lecture Notes in Computer Science, pp 433–444, Springer-Verlag, Berlin, 2007.
[11] W.M.P. van der Aalst, "The Application of Petri Nets to Workflow Management", The Journal of Circuits, Systems and Computers, 8(1):21–66, 1998.
[12] M.H. Sundari, A.K. Sen, and A. Bagchi, "Detecting Data Flow Errors in Work-flows: A Systematic Graph Traversal Approach", In 17th Workshop on Information Technology & Systems (WITS-2007), Montreal, 2007.
[13] Nikola Trocka, Wil M.P. van der Aalst, and Natalia Sidorova, "Data-Flow Anti-Patterns: Discovering Dataflow Errors in Workflows", LNCS 5565, 2009.
[14] Massimiliano de Leoni and Wil M.P. van der Aalst, "Data-Aware Process Mining: Discovering Decisions in Processes Using Alignments", ACM 978-1-4503-1656-9, 2013.
[15] Elham Ramezani, Vladimir Gromov, Dirk Fahland, and Wil M. P. van der Aalst, "Compliance Checking of Data-Aware and Resource-Aware Compliance Requirements", On the Move to Meaningful Internet Systems: OTM 2014 Conferences Lecture Notes in Computer Science Volume 8841, pp 237-257, 2014.
[16] Wil van der Aalst, W.M.P., Adriansyah, A., van Dongen, B.F, "Replaying history on process models for conformance checking and performance analysis", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(2), pp 182–192, 2012.
[17] M. de Leoni, W. M. P. van der Aalst, B. F. Van Dongen, "Data- and Resource-Aware Conformance Checking of Business Processes", In 15th International Conference on Business Information Systems, volume 117 of LNBIP, pp 48-59, Springer Verlag, 2012.