

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain

A framework for Business Process Data Management based on Big Data Approach

Asma Hassani^{a*}, Sonia Ayachi Gahnouchi^{a,b}

^aLaboratory RIADI-GDL, ENSI, Manouba, Tunisia

^bHigh Institute on Management of Sousse, Tunisia

Abstract

A business process (BP) refers to a set of activities carried out by humans to achieve one or more business goals. BPs are ubiquitous and occur in several sectors: marketing, healthcare, financial management and of course business. BPs generate a significant amount of data known as big data.

In recent years, the management of business process models and data is very challenging. On one hand, business process must be powerful in terms of modeling. On another hand, big data analytics support to find suitable knowledge to enact business process models.

In this paper, we will introduce an overview of our big data process-based approach that places big data and process in the same framework.

© 2017 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Keywords: BPM; Business Process; big data; big data analytics

* * Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .

E-mail address: asmahassani08@yahoo.fr

1. Introduction

A business process is a collection of related, structured activities that produce a specific service or a particular goal for a particular person(s). It can often be viewed as a sequence of activities with decision points based on data in the process. Business processes are inseparable from execution data, artifacts and data generated or exchanged during the execution of the process. Aalst and Ter Hofstede¹ has determined four perspectives to be considered for business process modeling (control flow, data flow, resource and operational perspective). The data flow perspective defines the data manipulated by activities, their structures, their sources, their destinations, and their transformation rules if they are exported to information systems or invoked applications¹. In this context, business process models are integrated with external information systems to ensure their execution by involving actors in the realization of their tasks. Furthermore, business processes become part of a complex area where they store and integrate data. This becomes an essential step for future analytics and decision applications.

On another hand, big data is playing a more and more important role nowadays. While many new technologies have been developed to support big data, it is equally important to explore software engineering technologies with big data. Moreover, Business Process Management (BPM) has gained great importance in the last decade and is increasingly used in several contexts (marketing, E-Commerce, E-Health, E-Learning, E-Government ...).

The usage of big data analytics in order to manage business process has gained more and more importance in recent years. First, BPM provides approaches for management of business processes to achieve business goals. Specifically, BPM includes techniques and tools for design, execution, analysis, and improvement of BP models. Secondly, BPM will have to face a big challenge if we need to design business processes which generate very big data. How to process big data is a key factor for a BPM project.

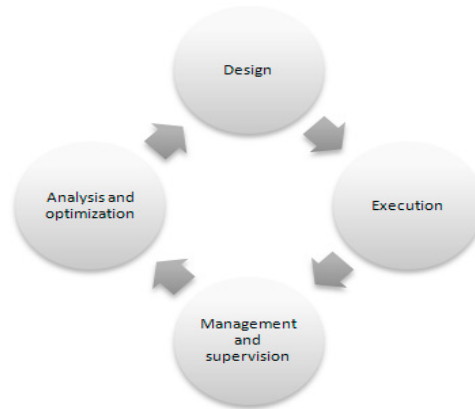
A key contribution of this paper is to propose a framework to facilitate business process improvement based on big data analytics. This framework describes the monitoring of business processes since the modeling phase, the deployment until the analysis of the relevant process-related data by means of big data analytics tools.

This paper is organized as follows. The first section is dedicated to present BPM. The second section is devoted to introduce the big data phenomenon. Big data analytics is presented in the third section. In the fourth section big data analytics and Business process are exposed. Our framework architecture is presented and detailed in the fifth section. Finally, the sixth section draws conclusions and suggests further research.

2. Business process Management

Business Process management (BPM) includes methods, techniques, and software to design, enact, control and analyze operational processes². This approach has received considerable attention in recent years due to its potential for significantly increasing productivity and reducing costs. BPM can be seen as successive steps that form the lifecycle of such an approach. BPM lifecycle model systemizes the steps and activities that should be followed for conducting a BPM project. According to Gillot³, the BPM lifecycle can be decomposed into four steps as follows (fig.1):

- Design: The processes are generally modeled using an understandable and executable graphical tool
- Execution: Once the processes have been designed, documented and simulated, they will be integrated into the information system
- Management and supervision: The processes are deployed in an execution environment and must be managed and monitored.
- Analysis and optimization: After a certain period of operation of the processes, the collected data can be used to analyze their functioning. The analysis will be able to identify the areas of the process that are poor or not performing.

Fig. 1. BPM lifecycle³

Data plays an important role in organizations systems. Nowadays, technological environment involves the interaction between several external and internal information systems (ERP, CRM, Web, etc.). New information and communication technologies contribute to the emergence of data particularly in business environment. Thus, Big Data is a great opportunity for organizations to improve their competitiveness which is based on efficient business processes. Organizations must be aware of new big data challenges and BPM capabilities to transform the massive amount of data to an actionable knowledge.

The big data perspective in BPM approach shows new challenges and opportunities in managing a large number of process data instance, and thus can bring huge value for process decision makers and process actors⁴. Here, Business Process Management Systems (BPMS) must be integrated with new big data infrastructures and have the ability to capture, generate and store vast amounts of data.

This combination of data and process, are fundamental challenges of BPM, including BI, BP integration and interoperation, and holds the promise of simplifying BPM. This approach also provides new ways to think about data. Some researchers⁵ believe that as more database systems become aware of this data-centric approach of BPM, new classes of techniques and approaches will address new challenges to the BPM community.

So, Big Data phenomenon covers two challenges for organizations; on the one hand, this continuous explosion of data and on the other hand the technological capacity to process and analyze this large mass of data in order to benefit from it to enact business process and adjust them with business environment requirements. With Big Data, organizations can now manage and process massive data to extract value, decide and act in real time.

3. Big data

The development of the web and the multiplication of data sources have generated a quantitative explosion of digitally created and shared data. Today, the business environment represents a large space where large amounts of information are added every minute. This evolution gives birth to a new paradigm called big data. Gupta and his colleagues defined big data as data that exceed the processing capacity of conventional database systems⁶. This involved that the data size is too large and data values change too quickly and/or they do not respect the rules of traditional database management systems.

Big data is defined through the 3Vs which are presented by Laney⁷ in 2001 as “high-volume, high velocity and high-variety information assets that demand cost-effective, innovative forms of information handling for improved insight and decision making”. In 2012, Gartner⁸ updated the definition as follows: “Big data is high volume, high velocity, and/or high variety of information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”. Other authors, researchers and engineers have extended the 3Vs to 4Vs and 5Vs by adding Value and Veracity to the definition of big data.

In the concept of "Big Data" it is not really the amount of data that makes the novelty, but rather the combination of 3 Vs. These fundamental characteristics are described as follows⁹:

- Volume: refers to large amounts of any kind of data from any different sources,
- Variety: refers to different types of data collected via sensors, smartphones or social networks, such as videos, images, text, audio, and so on. Moreover, these data can be in structured or unstructured formats.
- Velocity: refers to the speed of data transfers. The data content is constantly changing.

Big Data challenges cover not only the storage and management of a massive amount of data, but also the extraction of consistent knowledge from such data. Consequently, several issues related to acquisition, storage, management and analysis are attracting more and more attention. According to¹⁰, big data system is decomposed into four consecutive phases, including data generation, data acquisition, data storage, and data analytics as shown in figure 2.

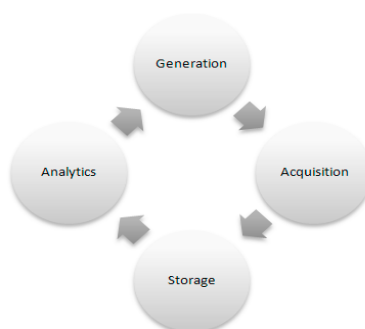


Fig. 2. Big data lifecycle¹⁰

The details of each phase are given as follows:

- Data generation: refers to how data are generated. Data are generated from various data sources such as sensors, video, click streams and several other digital sources.
- Data acquisition: concerns the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing.
 - Data collection: because of the diversity of sources and formats, data collection technologies are dedicated to acquire data from several data environments.
 - Data transmission: after collecting data, a high-speed transmission mechanism is required to transmit data to the storage support.
 - Data pre-processing: data pre-processing operations must be performed to enable future storage and analysis.
- Data storage: concerns persistently storing and managing large-scale datasets.
- Data analytics: provides analytical methods or tools to inspect, transform, and mine data to extract suitable knowledge.

Big data analytics research can be classified into five areas: text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics^{11,12}. In this paper we are interested in big data analytics problem and its integration in Business Process environments.

4. Big Data Analytics Vs Business Process Analysis

In this section we will begin by defining big data analytics (BDA) before presenting the main techniques adopted to analyze big data. First, data analysis is a process that transforms a mass of information into structured information allowing decision-making. This area refers to Business Intelligence and analytics (BI&A) technologies which are mainly employed in data mining and statistical analysis¹³. Otherwise, the existing techniques used by traditional BI&A tools are employed with structured data, but are insufficient with unstructured and semi-structured data^{11,14}. Big data brought with it new analytical challenges. The required type of analysis has to deal with data which are not

only big in term of quantity, but are also generated with various formats and in high speed. So, big data Analytics is when advanced analytic techniques operate on big data¹⁴.

Big data analytics research can be classified into five areas: text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics^{11,12}.

We present bellow a brief review of big data analytics techniques:

- *Text analytics* or text mining: refers to techniques that extract information from textual data. It involves statistical analysis, computational linguistics, and machine learning. This kind of analytics allows organizations to convert large amount of human-generated texts into meaningful summaries that support decision making. Text analytics systems are based on text representation and natural language processing (NLP).
- *Audio analytics or speech analysis*: used to analyze and extract information from unstructured audio data.
- *Video analytics*: Involves several techniques to monitor, analyze and extract meaningful information from video streams.
- *Social media analytics*: is the analysis of structured and unstructured data collected from various social media channels.
- *Predictive analytics*: comprises several techniques for predicting future outcomes based on the past or historical and current data.

Business Intelligence, data mining, machine learning and statistical analysis are becoming more prevalent in organizations interested in extracting information from their databases, in order to explain and predict data. Various data mining algorithms have been developed to solve data analysis problems, such as C45, K-means, KNN(K-Nearest Neighbors)¹³. These algorithms can be regrouped according to their objectives into several classes: classification, clustering, regression, association analysis and network analysis. The basic techniques to proceed the analysis of data are: RDBMS, Nosql databases, data warehousing, ETL, OLAP, etc.

To summarize, we presented big data analytics techniques: text, audio, video, and social media data analytics, as well as predictive analytics.

On another side, Business Process Analysis (BPA) is an important field to improve business processes. BPA aims to provide organizations with suitable information to understand how their processes are currently being performed². This knowledge can then be used to detect gaps between what is determined and what is going on, so that organizations can improve processes and systems in alignment with their determined business objectives. BPA includes a range of different tactics such as simulation and diagnosis, verification, and performance analysis of business processes¹⁵.

BDA and BPA are two research fields and have a great attention from researchers, analytics and decision makers. The majority of research works on business process analytics are concentrated on event data not on execution data. The basis of this paper is to combine these two fields for a better improvement of business processes in the area of big data.

5. Big data analytics and Business process

Data go through several steps before their value becomes observable. Above, we have presented a global overview of different steps in the big data management lifecycle (generation, acquisition, storage and analysis). Big Data relies on several technologies, which are used to exploit the massive data. In this part, we put the focus on big data analytics and its applicable business area. Analytics over big data has recently taken a great attention in several research application areas.

Business process management is based on business processes and refers to a way to enact and improve business efficiency based on data. BPM catches data from the various business applications of the organization¹⁶. In this context, data must be used to continue the execution of the business, as well as provide suitable knowledge to adjust processes with their business requirements and improve the overall business efficiency.

Thus, neither process nor big data is independent one from the other; new big data tools and techniques must be implemented to store, manage and analyze the big amount of process data execution. Besides, processes can benefit from the result of big data analytics; managers and BPM responsables must adjust and perform their processes with big data analytics capabilities to improve, change or update processes.

Data collection, storage, retrieval, analysis, and the application of data are the fundamental research problems dealt with for big data, where many authors and researchers^{17,18,19} claim that these problems can no longer be solved by traditional information processing technologies.

Data in the context of business processes is a valuable source of information for business process management. Nevertheless, traditional data management, basically Relational Database Management Systems (RDBMS) are inadequate to tackle the above mentioned list of major data challenges. In particular, two fundamental aspects qualify the mismatch between traditional RDBMSs and big data paradigm. First, the perspective of data structure; where traditional systems give a little bit attention to semi-structured and unstructured data. They focus and support only on structured data. Secondly, RDBMSs scale up with expensive hardware, which are inadequate to deal with the growing data volume.

Nowadays, new big data technologies in expansion are developed and adopted to support data storage, integration and analysis topics. Thus, these technologies have been developed by making the compromise to abandon some traditional functionalities of relational DBMSs.

Invented since 2009, Nosql databases present a conceptual and technological revolution with regard to relational databases. A panorama of Nosql schema has been created. Key-value, column oriented, document oriented and graph oriented databases are the main Nosql databases. Most NoSQL databases have an important property which is the fact that they are all schema-free. Indeed, the biggest advantage of schema-free databases is that they enable applications to quickly modify the structure of data and does not need to rewrite tables. Moreover, they possess greater flexibility when the structured data is heterogeneously stored¹⁷. The most popular NoSQL databases are Cassandra and MongoDB.

Many organizations give a great importance to Business processes because they allow them to create value. Organizations generally use process intelligence, process mining, or process analytics, and apply several statistical and artificial intelligence techniques to measure and analyze process-related data²⁰.

Big data provides new thought for BPM research. We will present in the next section our general framework architecture where we look for combining BPM and Big data capabilities in the same framework.

6. Framework Architecture

As presented and described in above sections, there are several types of Nosql databases which focus on storing and retrieving knowledge from large amounts of unstructured, semi-structured or even structured data.

Several tools are adopted to acquire and store, manage and analyze process data instances. We choose to adopt MongoDB to store business data because of several reasons. According to the data production speed, variety and volume, RDBMS fails to provide data with those characteristics. So MongoDB can be the suitable system to store and manage big volumes of data. Also, MongoDB allows storing unstructured data. As well, it does not need a predefined schema of data in tables; with MongoDB it is easy to represent open structured schema and store flexible data structure. Equally, document oriented databases are known by their scalability and high performance to treat big data.

We propose in fig. 3 an overview of our framework architecture where we intend to combine the two perspectives of business process and big data.

To make a BPM project successful in the era of big data, several systems need to be brought together. We present in the following the fundamental steps to combine the basic two perspectives.

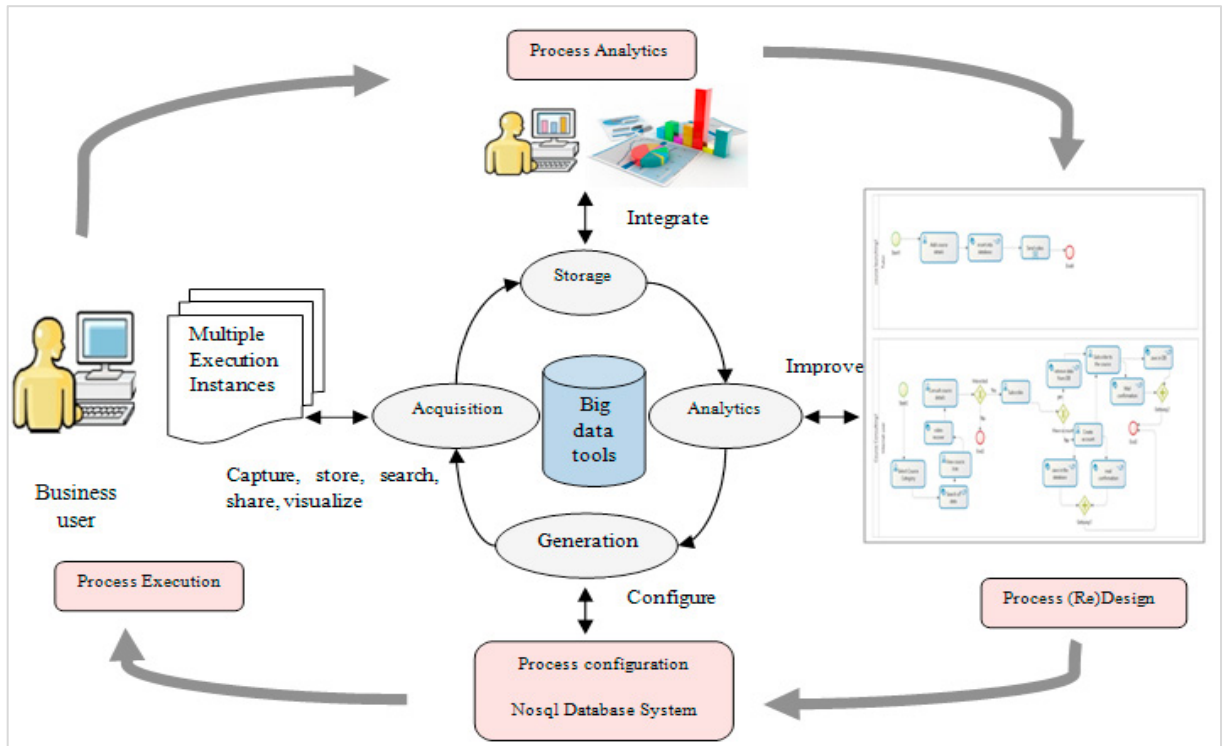


Fig. 3. Overview of the proposed framework

6.1. Process (re)design based on big data analytics

First, business process must be modeled to clearly describe the business scenario. Process modeling involves structuring and representing the scenario of the executing of activities of an organization using graphical notation. Modeling can rely on specialized methods and techniques. Process execution data analytics are considered as the key source of process redesign of existing process models to improve them and to make them more and more efficient.

6.2. Process configuration with big data generation tools

Process configuration refers to customization of the process with external information systems (example: CRM, Nosql DB, etc.) to ensure joint functionality and accessibility to the different services of each system. Data are generated from various data sources so our process model must be configured with all involved systems.

6.3. Process execution data acquisition

Once the process has been designed and configured, it is deployed and must be managed and monitored. The database model is responsible for collecting data from multiple instances of processes and storing them in the appropriate place. During the fulfillment of business activities, business actors will generate, share, use and modify relevant data during multiple instances of the process. All data will be stored in the Nosql database to be then managed and prepared and later processed for analysis.

6.4. Process analytics based on integrated data

After a certain period of process execution, the stored data can be used to analyze the process functioning.

Process analytics requires input data to carry out the needed analysis. So, datasets must be integrated and intended to help decision makers to lead to comprehensive and suitable analysis.

To sum up, the overall framework architecture aims to put together BPM capabilities with the awareness of big data to adapt to large amount of data and changing structure. This combination allows organizations to cover the two aspects, BPM and big data analytics, to ensure the improvement of business process in the era of big data.

7. Conclusion

Nowadays, data in business environments poses some of the most challenging problems of large-scale generation, integration and analysis. Marked by their considerable volume, heterogeneity and high speed of generation, all business data are collected, stored and analyzed to enhance efficiencies and guide decision making.

This paper describes our proposed framework to combine BPM capabilities with big data to adopt an approach based on big data analytics to improve business processes. The overall framework architecture aims to put together BPM means with the awareness of big data to adapt to large amount of data and heterogeneous structure. This combination allows organizations to cover the two aspects: BPM and big data analytics, to ensure an efficient improvement of business process in the era of big data.

Hadoop, Hdfs, mapReduce, Nosql databases and other big data tools revolutionizes the principles of data modeling, storage techniques, integration and analysis methods. As future work, we intend to implement our proposed approach and to apply appropriate big data tools with real process execution data.

References

1. Van Der Aalst WM, Ter Hofstede AH. YAWL: Yet Another Workflow Language. *Information Systems*, 30(4), 2005, p.245-275.
2. Van Der Aalst WM, Ter Hofstede H, Weske M. Business Process Management: A Survey, In: *Business process management*. Springer, 2003, p. 1-12.3.
3. Gillot J N. La gestion des processus métiers. Lulu. Com, 2007.
4. Houy C, Fettke P, Loos P, van der Aalst WM, Krogstie J. BPM-in-the Large–Towards a higher level of abstraction in Business Process Management. In *E- Government, E-Services and Global Processes*. 2010. p. 233-244.
5. Hull R, Su J, Vaculin R. Data management perspectives on business process management: tutorial overview. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 2013. p. 943-948.
6. Gupta R, Gupta H, Mohania M. Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?. In: Srinivasa S, Bhatnagar V, editors. *Big Data Analytics*, Volume 7678 *Lecture Notes in Computer Science*. New York: Springer-Verlag; 2012. p. 42-61.
7. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety, Technical Report, 2001.URL <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-andVariety.pdf>.
8. Beyer MA, Laney D. The Importance of ‘Big Data’: A Definition, Gartner, Stamford, CT. 2012.
9. Bello-Orgaz G, Jung J J, Camacho D. Social big data: Recent achievements and new challenges. *Information Fusion*, vol. 28. 2016. p. 45-59.
10. Hu H, Wen Y, Chua T S, Li X. Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, vol. 2. 2014. p. 652-687.
11. Vashisht P, & Gupta V. Big data analytics techniques: A survey. *International Conference In Green Computing and Internet of Things (ICGCloT)*, 2015. P.264-269.
12. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, vol 35(2), 2015, p.137-144.
13. Chen H , Chiang R H, Storey V C. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, vol 36(4), 2012, p.1165-1188.
14. Russom P. Big data analytics. TDWI best practices report, fourth quarter, 19, 40, 2011.
15. Vergidis K, Tiwari A, Majeed B. Business process analysis and optimization: Beyond reengineering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(1), 2008. p. 69-82.
16. Fang Z, Yin C. BPM architecture design based on cloud computing. *Intelligent Information Management* vol2 (5), 2010.
17. Chen C P, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275,2014, p. 314-347.
18. Moniruzzaman A B M, Hossain S A. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison.2013.
19. Labrinidis A, Jagadish H V. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, vol. 5(12), 2012, p. 2032-2033.
20. Janiesch C, Matzner M, Müller O. Beyond Process Monitoring: A Proof-of-Concept of Event-Driven Business Activity Management, *Business Process Management J.*, vol. 18(4), 2012, p. 625–643.