Business process analytics using a big data approach

Alejandro Vera-Baquero1, Ricardo Colomo-Palacios1 and Owen Molloy2

1Universidad Carlos III de Madrid

Av. Universidad 30, 28911 Leganes, Madrid Spain

Phone: 0034 91 624 59 58

Fax: 0034 91 624 91 29

averabaq@gmail.com; ricardo.colomo@uc3m.es

2 National University of Ireland, Galway, Ireland

Phone: 00 353 91 493330

Fax: 00 353-91-750501

owen.molloynuigalway.ie

# Abstract

Continuous improvement of business processes is a challenging task that requires complex and robust supporting systems. The use of advanced analytics methods and emerging technologies such as business intelligence systems, business activity monitoring, predictive analytics, behavioural pattern recognition and “what-if” type simulations are essential to assist business users in the continuous improvement of their processes. Nevertheless, the high volumes of event data produced by the execution of processes during the business lifetime prevent business users from accessing analytics data efficiently and on an acceptable response time basis. This paper presents a technological solution using a big data approach to provide business analysts with visibility on distributed process and business performance. The proposed architecture will enable end-users to analyse business performance on highly distributed environments in or near real-time.

# Introduction

As organizations reach higher levels of business process management (BPM) maturity, they often find themselves maintaining very large process model repositories, representing valuable knowledge about their operations [[1](#Dum)]. Business processes (BPs) have become increasingly important in many enterprises since they determine the procedure for developing and distributing value for the customers and are key drivers for the three success criteria: cost, quality, and time [[2](#Ada)]. Several widely used quality models, including ISO 9001 and European Foundation for Quality Management highlight the importance of process orientation. The application of various measurement and analysis techniques on process-related data, using a variety of statistical and artificial intelligence techniques, is often grouped under umbrella terms such as process intelligence, process mining, or process analytics [[3](#Jan)]. According to Van der Aalst et al. [[4](#van)], there are basically three types of business process analysis (BPA): validation, verification and performance analysis. In any case, all these analyses rely on high volumes of process and event data that must be collected and stored. The concept of event plays a crucial role in this scenario. In the context of our work, events represent state changes of objects within the context of a BP [[3](#Jan)]. In spite of the importance of events for event-driven BPM and analysis, no commonly adopted event format for communicating business events between distributed event producers and consumers has emerged [[5](#Bec12)] although BPAF (Business Process Analytics Format) [[6](#zur11)] is one of the most important and adopted proposals. Based in this standard, there are several proposals in the literature devoted to analyze the events and execution outcomes of BPs [[7](#Ver13)].

Nevertheless, due to the incredible growth of process event data, new approaches for BPA must be adopted. One of these approaches is based on big data that, according to [[8](#Zha12)], will be widely leveraged in more and more applications for developing deep business insights. “Big Data” provides new prospects for BPM research given that organizations are recording large amounts of event data and this is great opportunity to promote “evidence-based BPM” [[9](#van12)]. Different areas and paradigms are arising around BPM and Data Mining with the aim at filling the gap between BPM and performance analysis over very large volumes of event data. Process mining intends to connect event data to process models, and thus, in a very large enough scale, it represents the missing link between analysis of big data and BPM [[10](#Van12)]. The purpose of this paper is to introduce an architecture aimed to integrate big data analytics with BPM in a distributed environment in order to provide end-users with a solution to analyze the execution outcomes of business processes.

# Architecture and implementation

This paper proposes a cloud-based infrastructure to support BPA over highly distributed environments. This aims to provide business users with visibility on process and business performance by monitoring BPs from operational systems where their execution data outcomes can be collected, unified and stored in an appropriate structure that enables end-users to measure and analyse such valuable information. The analysis of these event data can assist end-users in gaining an insight into business performance with the aim of achieving organizational effectiveness by improving their BPs as well as providing analysts with a powerful understanding of what happened in the past, to evaluate what happens at present and to predict the behaviour of process instances in the future [[11](#Zur)]. Nevertheless, the effective management of business information is a challenging task that cannot be easily achieved using traditional approaches. Event data integration is essential for analytic applications, but especially hard to achieve on highly distributed environments whose BPs are part of complex supply chains that are normally executed under a variety of diverse heterogeneous systems. Additionally, the continuous execution of distributed BPs produces a vast amount of event data that cannot be efficiently managed by means of traditional systems which are not adequate to manage event data of the order of hundreds of millions of linked records. Likewise, centralized systems are not suitable because they entail a significant latency from the time the event occurs on source to the time the event is recorded in central repositories. These shortcomings prevent existing approaches, such as [[7](#Ver13)], from providing instant business analytics on highly distributed environments. In addition, we are typically dealing with highly distributed supply chains, where individual stakeholders are geographically separate and need a platform to perform BPM in a collaborative fashion, rather than depending on a single centralised process owner to monitor and manage performance at individual supply chain nodes. We therefore propose a fully distributed solution that supports collaborative BPA over highly distributed environments.



Figure 1 - Overall architecture

This approach, based on the framework described in [[7](#Ver13)], proposes an extension based upon a cloud-based infrastructure complemented with a federative approach according to [[12](#Riz11)], in terms of data warehousing and distributed query processing.

This solution provides the capabilities for capturing and integrating event data from operational systems whose BPs flow through a diverse of heterogeneous systems such as business process execution language (BPEL) engines, ERP systems, workflows… as well as storing very large volumes of data in a global, distributed BP execution repository.

Each organizational unit handles its own local *business analytics service unit* (BASU) component, which is attached to every operational business system along with their own local event repository which is built upon Big Data technology. Here, this repository is implemented using the HBase product combined with Apache Hadoop as Big Data storage, and incorporates the Hive project for enabling data warehouse capabilities over Big Data.

These local components enable BPA to be carried out collaboratively in each organization independently by performing distributed queries along the collaborative network. Likewise, the integration of BASU subsystems is required for measuring the performance of cross-functional BPs that are extended beyond the boundaries of organizations.

The Global Business Analytics Service (GBAS), is the entity responsible for integrating BASU components and the core point for providing analytical services to third-party applications.

The overall architecture (Figure 1) has the ability to provide cloud computing services at very low latency response rates. These services can contribute to continuous improvement of BPs through the provision of a rich informative environment that supports BPA and offers clear insights into the efficiency and effectiveness of organizational processes. Furthermore, these services can be leveraged by a wide range of analytical applications such as real-time BI systems, business activity monitoring (BAM), simulation engines, collaborative analytics, etc.

Collaborative BI environments, in terms of business analytics functionality, extend “the decision-making process beyond the company boundaries thanks to cooperation and data sharing with other companies and organizations” [[12](#Riz11)]. In addition, federated data warehouses provide transparent access to the distributed analytical information across different functional organizations, and this can be achieved through the definition of a global schema that represents the common business model of the organization. We therefore require the construction of a generic model with two aims: 1) to represent the business performance of organizations, and 2) to be fully agnostic to any specific business domain.

## 2.1 Event-based model

From a BP perspective, an event model is required in order to provide the framework of a concrete understanding and representation of what needs to be monitored, measured and analysed [[13](#Cla08)]. The event structure must represent the data execution of whatever BP flows through a diverse set of heterogeneous systems and must support the information required to effectively analyze BP performance.

An event model represents actions and events which occur during the execution of a BP. The proposed event model provides the information required to enable the global system to perform analytical processes over them, as well as representing any derived measurement produced during the execution of any BP flow [[7](#Ver13)]. This model is built upon the BPAF standard, specified in [[14](#WfM12)], combined with some important features of the iWISE model widely discussed in [[13](#Cla08)] and [[15](#Mol10)].

BPAF is a standard format to support the analysis of audit data across heterogeneous BPM systems [[14](#WfM12)]. It enables the delivery of basic frequency and timing information to decision makers, such as the cycle times of processes, wait time, etc. This permits host systems to determine what has occurred in the business operations by enabling the collection of audit data to be utilized in both analysis and the derivation of status information [[11](#Zur)].

The primary sources for BPAF data are events streams coming from BPM systems. With regard to the design of a generic process analytics system it provides an event format independent of the underlying process model. This format enables analytic applications and BAM technology to unify criteria and to standardize a state model for event auditing purposes in heterogeneous environments [[11](#Zur)].

This proposed event model, discussed in [[7](#Ver13)], is built upon a BPAF extension to accommodate the event correlation features defined by iWISE. As part of this work, this event format has been modified for supporting distributed storage.

## 2.2 Business Analytics Service Unit

As previously stated, capturing and integrating business events from heterogeneous operational systems and different organizational units is a challenging task, not only from a BAM perspective, but also with respect to the representation of performance information.

The BASU component (see Figure 2) is responsible for achieving such functionality on a local scope whilst the cross-organizational dependencies are managed by the GBAS module which is responsible for integrating an undetermined set of BASU modules across the entire system.



Figure 2 - BASU Architecture

The event publisher is responsible for capturing the events from legacy business systems and publishing them to the network throughout an ActiveMQ message broker instance. The legacy listener transforms event streams into XML messages structured in the extended BPAF format and forwards the enterprise events to a specific JMS queue as they occur.

The event subscriber is continuously listening for incoming events upon a specific JMS queue. Each event is then processed individually by transforming the content of its XML message into a memory representation of an instance in an extended BPAF format. Every instance is then forwarded to the event correlator which is responsible for identifying and setting the correct sequence of the incoming events before they get stored into Big Data tables.

The event correlator module is responsible for identifying the correct sequence of events per process instance or activity. It leverages the extended BPAF data to determine the process instance or activity associated with the event. This is achieved by querying the local event store for the existence of a process instance associated with the correlation data provided. The information retrieval at this stage is critical, as the latency for querying Big Data tables must be minimal in order that the system can provide BAM services in a short-time response basis.

The event store provides a service interface to access the big data store containing the live enterprise event data. The core of this module is basically composed of a set of entity beans that represents the business events in BPAF format, a set of *Spring* components for managing the event data throughout the Java Persistence API (JPA), and another set of *Spring* components that provides the service interface to the data access methods. The use of JPA specification over Big Data tables is performed by an implementation of JPA over HBase which is supported by the DataNucleus project.

An important component of this module is the implementation of the ETL methods for extracting the event information received from the subscriber module, transforming the event data structured in the extended BPAF format into raw BPAF [[7](#Ver13)], and loading the resulting data into the event store.

Whilst the live enterprise data give an insight into the BP execution, they do not provide measurable information about business performance [[7](#Ver13)], therefore the definition of metrics are required in order to provide business analysts with an understanding of the behavioural aspects of BPs.

The event data warehouse module is devised for this purpose, and it is composed of a data repository of metrics combined with a subset of event data that enable end-users to query business events for analytical purposes. The underlying storage system is based upon an HBase instance along the Hive product with the aim at supporting data warehouse capabilities over Big Data.

The proposed system captures and records the timestamp of events locally containing the time at which they occurred on the source system. The analysis of the timestamp of a set of correlated events is leveraged by the event data warehouse module to construct metrics per process instance or activity as the events arrive. This analytical information is derived in a very tight timeframe as events arrive and is fully accessible through a specific-purpose SQL-like query language discussed in [[7](#Ver13)].

Metrics and live event data are jointly stored and managed within a data warehouse implementation. This module implements this component in order to provide analysts with the ability to retrieve and process historical events as well as analyse the behaviour of BPs by the means of a set of proposed metrics widely discussed in [[7](#Ver13)] and [[11](#Zur)].

## 2.3 Global Business Analytics Service

Previous sections have covered how to correlate events per process instance or activity, but not how to identify sequences of inter-related process within a supply chain that are parts of a higher level global BP. As long as a process runs across a diverse set of heterogeneous systems such as BPEL engines, workflows engines, etc. It is necessary to identify the sequence flow of a BP that is running along the involved systems. Such sequence identification is called *instance correlation*.

Instance correlation refers to the way in which messages are uniquely identified across different process instances [[13](#Cla08)] within the context of an upper global BP. From a business analytics perspective, this is extremely important to enable end-users to understand the correlation between business events and to drive automated decision making.

This component is responsible for integrating a set of BASU components and correlating the process instances that are executed across their organizational boundaries. These BASU subsystems are connected through an Enterprise Service Bus (ESB) representing a collaborative network where xml events and metrics data is flowing through.

The GBAS component has the ability to provide analytical services of global processes by itself as it stores information in terms of business performance and live enterprise data from cross-organizational BPs. Likewise, it allows drilling down into multiple levels of detail by performing distributed queries throughout the BASU components along the collaborative network.

Authors collected numerical performance data of live event operations. In a dataset of over 1000000 events in a test environment, authors collect the data under various execution concurrencies. Read operations are performed in the range [0.2-05] milliseconds (average 0.31 and standard deviation of 0.13) while write instructions are performed in the range of [5-9] milliseconds. However, going beyond figures, the most remarkable finding is that times are not growing with the growth of the dataset and there are not statistical significance in such times when comparing, for instance the dataset populated with 700000 events and with 1000000 events.

# Conclusions and Future work

A cloud-computing solution for supporting distributed BPA based on Big Data technology has been presented herein aimed at providing integrated services for monitoring and analysing BP performance over highly distributed environments. A set of BASU components along with the master GBAS component have been devised with the purpose of monitoring operational activities, collecting data originating from distributed heterogeneous enterprises systems, storing the enterprise data leveraging Big Data underlying technology and inferring knowledge from the gathered information. The successful integration of BASU/GBAS components through ESB adapters completes the solution proposed.

Likewise, an event-based model based on BPAF [[14](#WfM12)] and iWISE [[13](#Cla08)] [[15](#Mol10)] has been proposed in [[7](#Ver13)] and extended for fulfilling the aims of this work. Such a model has the capacity to represent measurable data for analysing distributed BPs on distributed environments, as well as enabling capturing systems to interpret and process event streams that are part of a cross-functional BP.

In the absence of standards for querying BPs, a BPEQL (Business Process Execution Query Language) query language has been proposed in [[7](#Ver13)]. The extension of the query engine is also part of future work for fulfilling the objectives presented in this paper in terms of distributed processing and collaborative environments. Furthermore, one of the main objectives of that research work is to enable the systems to access structural and behavioural properties of business properties in real-time by leveraging Big Data technology that is supported on BASU/GBAS components.

One of the major limitations of the current approach is that distributed processing produces a significant overhead in comparison with centralized approach presented in [[7](#Ver13)]. The network latency and processing overhead on GBAS component is greatly increased as the number of nodes grows. Furthermore, process instances correlation considerably impacts on the overall system performance and prevents the systems to respond in an adequate time basis nearly in real-time, especially on large and complex supply chains which are specifically the most interesting cases we are aiming at monitoring and improving. Additionally, gaining accuracy and prediction on system performance in terms of data access on very large volumes of data is one of the main aims for measuring the performance of the system in general and the response time of query processing in particular. In an ideal scenario, BPA techniques will be performed over a very large amount of data, thereby the scalability of the system, in terms of volume of data and workload on business queries, gains an enormous significance in the system evaluation, and thus this will be an important case of study on further research work. In this regard, the use of HDSF clustering capabilities will be key to address potential performance issues on event-correlation due to two main factors: 1) the high dependency of the event correlation mechanism on the data access, and 2) the high event-arrival rates on highly distributed environments.

Other potential further research includes the gradual incorporation of services for supporting advanced functionality that is demanded by emerging technologies such as behavioural pattern recognition or optimization techniques. In addition, the provision of simulation techniques would highly empower the cloud-based functionality since structured data may serve as an input to simulation engines. This will enable business users to anticipate actions by reproducing what-if scenarios, as well as performing predictive analysis over augmented data that constitutes a base of hypothetical information. Likewise, this would enable analysts to reproduce live process instances and re-run event streams in simulation mode for diagnosis purposes and root cause analysis.

Finally, collaborative business analytics is another potential research area to explore. The cooperation and data sharing between different companies or organizations using Big Data would significantly improve, not only the visualization of inter-related business analytical information in real-time, but also to identify and collaboratively perform diagnostics and root-cause analysis on non-compliant situations and bottleneck issues along large and complex BPs that cross-organizational boundaries.

# SIDEBAR

**Apache Hadoop**: Open-source project from Apache that supports intensive processing of large data sets across distributed systems. It is designed to feature high performance and scalability on data-intensive applications, whereby data systems can scale up from a single sever to hundreds or thousands of computing nodes, each offering parallel computation and distributed storage. For further information see <http://hadoop.apache.org>

**HBase**: Open-source distributed database system that is developed as part of the Apache Hadoop project. It is a NoSQL, versioned, column-oriented data storage system that provides random real-time read/write access to big data tables and runs on top of HDFS (Hadoop Distributed Filesystem). More information at <http://hbase.apache.org>

**Hive**: Open-source data warehouse system that is developed as part of the Appache Hadoop project. It provides data summarization, queries, and analysis of large datasets. Likewise, it incorporates a mechanism to feature ad-hoc queries via a general-purpose SQL-like language called HiveQL while maintaining traditional map/reduce on those situations where complex logic is inadequate to be expressed using HiveQL. More information at <http://hive.apache.org>

# Bibliography

|  |  |
| --- | --- |
| [1] | M. Dumas, L. García-Bañuelos, M. La Rosa, and R. Uba, "Fast detection of exact clones in business process model repositories," *Information Systems*, vol. 38, no. 4, pp. 619-633, Jun. 2013. |
| [2] | S. Adam, N. Riegel, J. Doerr, O. Uenalan, and D. Kerkow, "From business processes to software services and vice versa—an improved transition through service-oriented requirements engineering," vol. 24, no. 3, pp. 237-258, 2012. |
| [3] | C. Janiesch, M. Matzner, and O. Müller, "Beyond process monitoring: a proof-of-concept of event-driven business activity management," vol. 18, no. 4, p. 625–643, Jul. 2012. |
| [4] | W. van der Aalst, M. Weske, and G. Wirtz, "Advanced topics in workflow management: Issues, requirements, and solutions," vol. 7, no. 3, pp. 49-77, Jan. 2003. |
| [5] | J. Becker, M. Matzner, O. Müller, and M. Walter, "A Review of Event Formats as Enablers of Event-Driven BPM," in *Business Process Management Workshops*, F. Daniel, K. Barkaoui, and S. Dustdar, Eds. Springer Berlin Heidelberg, Jul. 2012, vol. 99, pp. 433-445. |
| [6] | M. zur Muehlen and K. D. Swenson, "BPAF: A Standard for the Interchange of Process Analytics Data," in *Business Process Management Workshops*, M. zur Muehlen and J. Su, Eds. Springer Berlin Heidelberg, 2011, pp. 170-181. |
| [7] | A. Vera-Baquero and O. Molloy, "A Framework to Support Business Process Analytics," in *Proceedings of the International Conference on Knowledge Management and Information Sharing*, Barcelona, 2013, pp. 321-332. |
| [8] | L. .-J. Zhang, "Editorial: Big Services Era: Global Trends of Cloud Computing and Big Data," *IEEE Transactions on Services Computing*, vol. 5, no. 4, p. 467–468, 2012. |
| [9] | W. M. P. van der Aalst, "A Decade of Business Process Management Conferences: Personal Reflections on a Developing Discipline," in *Business Process Management*, A. Barros, A. Gal, and E. Kindler, Eds. Springer Berlin Heidelberg, 2012, pp. 1-16. |
| [10] | W. Van Der Aalst, "Process mining," *Communications of the ACM*, vol. 55, no. 8, p. 76–83, Aug. 2012. |
| [11] | M. zur Muehlen and R. Shapiro, *Handbook on Business Process Analytics*. Springer Verlag, vol. 2. |
| [12] | S. Rizzi, "Collaborative Business Intelligence," in *First European Summer School (eBISS 2011)*, Paris, 2011, pp. 186-205. |
| [13] | C. Costello, "Incorporating Performance into Process Models to Support Business Activity Monitoring," PhD. Thesis, National Uinversity of Ireland, Galway, 2008. |
| [14] | WfMC. (2012, Feb.) Workflow Management Coalition - Business Process Analytics Format Specification. [Online]. <http://www.wfmc.org/Download-document/Business-Process-Analytics-Format-R1.html> |
| [15] | O. Molloy and C. Sheridan, "A Framework for the use of Business Activity Monitoring in Process Improvement," in *E-Strategies for Resource Management Systems: Planning and Implementation*, E. Alkhalifa, Ed. IGI Global, 2010. |