# Business Process Analytics Using a Big Data Approach

**Alejandro Vera-Baquero and Ricardo Colomo-Palacios, Universidad Carlos III de Madrid**

**Owen Molloy, National University of Ireland**

***Business process executions on large and complex supply chains produce high volumes of unstructured event data, so timely data analysis is difficult. An architecture for integrating big data analytics into business performance management helps users analyze and improve business processes performance.***

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] Business processes have become increasingly important in many enterprises, because they determine the procedure for developing value and distributing it to customers. Furthermore, such processes are the key drivers behind three critical success factors—cost, quality, and time.[2]

Several widely used quality models, including ISO 9001 and the European Foundation for Quality Management, highlight the importance of process orientation. Companies often use *process intelligence*, *mining*, or *analytics*,[3] applying a variety of statistical and artificial intelligence techniques to measure and analyze process-related data. According to Will van der Aalst and his colleagues,[4] the three types of business process analysis (BPA) are *validation*, *verification*, and *performance*—all of which require collecting and storing large volumes of process and event data.

Here, we focus on events, which represent state changes in objects in the context of a business process.[3] Despite the importance of events for event-driven BPM and BPA, no commonly adopted format for communicating business events between distributed event producers and consumers has emerged,[5] although BPA solutions often adopt the Business Process Analytics Format (BPAF) standard.[6] Several proposals use BPAF to analyze business process events and execution outcomes.[7]

However, given the recent growth in process event data, new *business intelligence* trends must adopt new BPA approaches, and, according to Liang-Jie Zhang,[8] approaches that apply big data will be widely leveraged in developing deep business insights. Big data provides new prospects for BPM research—especially for *evidence-based BPM*, where research outcomes can be empirically evaluated with real data.[9] Process mining aims to connect event data to process models, and, on a larger scale, act as the missing link between BPM and big data analysis.[10] Our proposed architecture for integrating big data analytics with BPM in a distributed environment will help users analyze business-process execution outcomes in a timely manner.

## Analytics in Distributed Environments

Our cloud-based infrastructure aims to provide business users with greater visibility into process and business performance. It will let them monitor business process executions from operational systems that can collect, unify, and store execution data outcomes in an appropriate structure for later measurement and analysis. By analyzing event data, users can better understand business performance and improve their processes to achieve greater organizational effectiveness. Furthermore, such data helps analysts not only understand what happened in the past but also evaluate what's currently happening and predict the behavior of future process instances.[11]

However, effectively managing business information is challenging and not easily achieved using traditional approaches. Event data integration is essential for analytic applications, but it's difficult to achieve in highly distributed environments, where business processes are part of complex supply chains

that are normally executed under a variety of heterogeneous systems. Additionally, the continuous execution of distributed business processes produces a vast amount of event data that traditional systems can't manage efficiently, because they can't handle the hundreds of millions of linked records. Likewise, centralized systems aren't suitable, because they entail a significant latency between when the event occurs and when it's recorded in central repositories.

These shortcomings prevent existing approaches, such as the Framework for Business Process Analytics (F4BPA),[7] from providing instant business analytics in highly distributed environments. In addition, we're 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 centralized process owner to monitor and manage performance at individual supply chain nodes. So, we propose extending the framework using a cloud-based infrastructure and complementing it with Stefano Rizzi's federative approach,[12] using data warehousing and distributed query processing. This will let the framework capture and integrate event data from operational systems whose business processes flow through a diverse set of systems, such as business process execution language (BPEL) engines and enterprise resource planning systems, as well as store very large volumes of data in a global, distributed business process execution repository.

## Framework Architecture

Each organizational unit handles its own local *business analytics service unit* (BASU) component, which is attached to other operational business systems and to the local event repository built on big data technology. We implemented this repository using Apache Hadoop and HBase, and we further incorporated Hive to enable data warehouse capabilities over big data (see the "Related Open Sources Projects" sidebar for more information).

These local components enable each organization to carry out BPA independently but also collaboratively by performing distributed queries along the network. Likewise, the integration of BASU subsystems lets organizations measure the performance of cross-functional business processes that extend beyond organizational boundaries. The *global business analytics service* (GBAS) integrates the BASU components and acts as the core point for providing analytical services to third-party applications.

The overall architecture (see Figure 1) can provide cloud computing services at very low latency response rates. These services can help continuously improve business processes 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 business intelligence systems, business activity monitoring, simulation engines, and collaborative analytics.

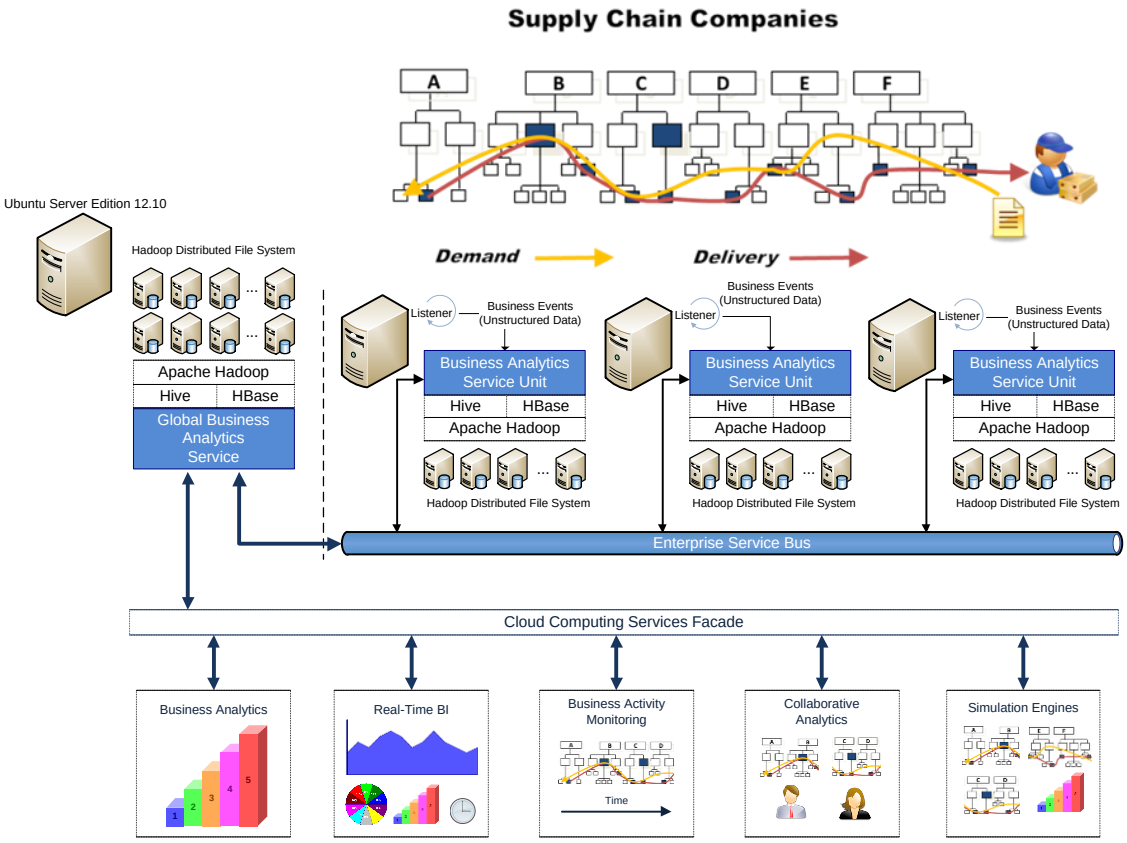


*Figure 1. The distributed business analytics services architecture.*

According to Rizzi, collaborative business intelligence 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] In addition, federated data warehouses provide transparent access to the distributed analytical information across different functional organizations, and this can be achieved by defining a global schema that represents the organization's common business model. We thus must construct a generic model that represents the business performance of organizations yet remains fully agnostic to any specific business domain.

## An Event-Based Model

The framework needs an event model to provide a concrete understanding of what should be monitored, measured, and analyzed.[13] The event structure must represent the data execution of whatever business process flows through a diverse set of heterogeneous systems and must support the information required to effectively analyze business process performance.

An event model represents actions and events that occur during the business process execution. The proposed event model provides the information required to let the global system perform analytical processes over these actions and events, as well as represent any derived measurement produced during business process flow execution.[7] We built this model using the BPAF standard[14] and combined important features from the intelligent Web Services Enterprise Integration Environment (iWise) model.[13,15]

BPAF supports the analysis of audit data across heterogeneous BPM systems.[14] It enables the delivery of basic frequency and timing information to decision makers, such as the cycle times of processes, wait times, and so on. This lets host systems determine what has occurred in the business operations by letting them collect audit data, which users can analyze for status updates and other information.[11]

The primary sources for BPAF data are event streams coming from BPM systems. BPAF provides an

event format independent of the underlying process model, and we leverage this feature to construct a generic process analytics system. This format helps analytic applications and business activity monitoring technology unify criteria and standardize a model for auditing events in heterogeneous environments.[11]

The proposed event model, discussed elsewhere,[7] is built on a BPAF extension to accommodate the event correlation features defined by iWISE. As part of this work, we modified the event format to support distributed storage.

## The Business Analytics Service Unit

The BASU component (see Figure 2) is responsible for local analyses, and the GBAS module manages the cross-organizational dependencies, integrating an undetermined set of BASU modules across the entire system.

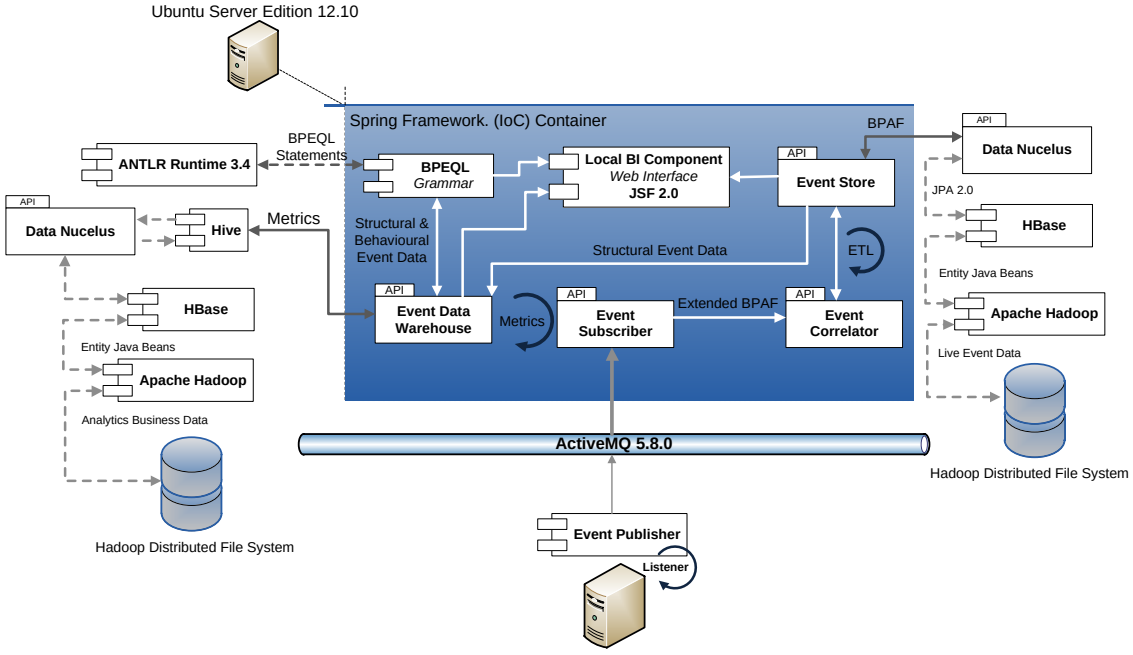


*Figure 2. The architecture of the business analytics service unit (BASU).*

The event publisher captures the events from legacy business systems and publishes 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 Java Message Service (JMS) queue as they occur.

The event subscriber is continuously listening for incoming events in 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 identifies the correct sequence for incoming events before storing them in big data tables.

The event correlator leverages the extended BPAF data to determine the process instance or activity associated with the event by querying the local event store for a process instance associated with the correlation data provided. The information retrieval at this stage is critical, because the latency for querying big data tables must be minimal so the system can provide timely business activity monitoring.

The event store provides a service interface to access the big data store containing the live enterprise event data. The core of this module comprises 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. We used an implementation of JPA over HBase, which the Data Nucleus open source project supports, so we could apply the JPA specification to easily access and manage the big data tables.

An important component of this module is the implementation of the e*xtract, transform, load* 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] and loading the resulting data into the event store. Although the live enterprise data gives insight into the business process execution, it doesn't provide measurable information about business performance,[7] so metrics must be defined to help business analysts understand the processes' behavior. Consequently, the event data warehouse module, composed of a data repository of metrics and a subset of event data, lets users query business events for analytical purposes. The underlying storage system is based on an HBase instance along the Hive product to support data warehouse capabilities over big data.

The proposed system captures and records the timestamp of events locally, noting the time at which the events occurred on the source system. The event data warehouse module analyzes the timestamp of a set of correlated events 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 it's fully accessible through a specific-purpose SQL-like query language.[7]

Metrics and live event data are jointly stored and managed in a data warehouse implementation. This module implements this component to help analysts retrieve and process historical events as well as analyze the business process behavior using a set of proposed metrics.[7,11]

## The Global Business Analytics Service

Now that we've covered how to correlate events per process instance or activity, we turn to identifying sequences of interrelated processes in a supply chain that are parts of a higher-level global business process. As long as a process runs across a diverse set of heterogeneous systems, such as BPEL engines or workflows engines, it's necessary to identify the sequence flow of a business process that's running along the involved systems.

Such sequence identification, called *instance correlation*, refers to the way in which messages are uniquely identified across different process instances[13] in the context of an upper global business process. From a business analytics perspective, this is extremely important, because it lets users understand the correlation between business events to drive automated decision making.

This component integrates a set of BASU components and correlates 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 with XML events and metrics data flowing through.

The GBAS component can provide analytical services of global processes by itself, because it stores information in terms of business performance and live enterprise data from cross-organizational business processes. Likewise, it lets users drill down into multiple levels of detail by performing distributed queries throughout the BASU components along the collaborative network.

We collected numerical performance data of live event operations. In a dataset of over 1,000,000 events in a test environment, we collected the data under various execution concurrencies. Read operations were performed in the range of 0.2 to 0.5 milliseconds (average 0.31 and standard deviation of 0.13), while write instructions were performed in the range of 5 to 9 milliseconds. However, the most remarkable finding was that the times didn't increase with the growth of the dataset, and there wasn't statistical significance in such times when comparing, for example, the dataset populated with 700,000 versus 1,000,000 events.

One of the major limitations of the current approach is that distributed processing produces significant overhead in comparison with a centralized approach.[7] The network latency and processing overhead on the GBAS component increases greatly as the number of nodes grows. Furthermore, process instance correlation considerably affects overall system performance and prevents systems from responding in near real time, especially on large and complex supply chains—precisely the cases we hope to monitor and improve.

Additionally, being able to accurately predict system performance in terms of data access for very large volumes of data is one of the main aims for measuring general system performance as well as response times for query processing. In an ideal scenario, BPA techniques will be performed over a very large amount of data, so system scalability—in terms of the data volume and business queries workload—must be evaluated and thus will be an important case of study in future work. In this regard, using Hadoop

Distributed File System clustering capabilities will be key to addressing potential performance issues for event correlation, owing to two main factors: the high dependency of the event correlation mechanism on the data access, and the high event-arrival rates on highly distributed environments.

Other potential research includes gradually incorporating services to support the advanced functionality that emerging technology demands, such as behavioral pattern recognition or optimization techniques. In addition, including simulation techniques would empower the cloud-based functionality. Structured data could serve as an input to simulation engines, letting business users anticipate actions by reproducing what-if scenarios and performing predictive analysis over augmented data that constitutes a base of hypothetical information. Likewise, this would help analysts reproduce live process instances and rerun event streams in simulation mode for diagnosis purposes and root cause analysis.

Finally, collaborative business analytics is another potential research area. Cooperation and data sharing between different organizations using big data would significantly improve the visualization of interrelated business analytical information in real time. Furthermore, it would help the organizations collaboratively perform diagnostics and root-cause analysis on noncompliant situations and bottleneck issues along large and complex business processes that cross organizational boundaries.

## References


1. M. Dumas et al., "Fast Detection of Exact Clones in Business Process Model Repositories," *Information Systems*, vol. 38, no. 4, 2013, pp. 619–633.
2. S. Adam et al., "From Business Processes to Software Services and Vice Versa—An Improved Transition through Service-Oriented Requirements Engineering," *J. Software: Evolution and Process*, vol. 24, no. 3, 2012, pp. 237–258.
3. C. Janiesch, M. Matzner, and O. Müller, "Beyond Process Monitoring: A Proof-of-Concept of Event-Driven Business Activity Management," *Business Process Management J.*, vol. 18, no. 4, 2012, pp. 625–643.
4. W.M.P. van der Aalst, M. Weske, and G. Wirtz, "Advanced Topics in Workflow Management: Issues, Requirements, and Solutions," *J. Integrated Design and Process Science*, vol. 7, no. 3, 2003, pp. 49–77.
5. J. Becker et al., "A Review of Event Formats as Enablers of Event-Driven BPM," *Business Process Management Workshops*, vol. 99, F. Daniel, K. Barkaoui, and S. Dustdar, eds., Springer Berlin Heidelberg, 2012, pp. 433–445.
6. M. zur Muehlen and K.D. Swenson, "BPAF: A Standard for the Interchange of Process Analytics Data," *Business Process Management Workshops*, M. zur Muehlen and J. Su, eds. Springer, 2011, pp. 170–181.
7. A. Vera-Baquero and O. Molloy, "A Framework to Support Business Process Analytics," *Proc. Int'l Con. Knowledge Management and Information Sharing*, SciTePress, 2013, pp. 321–332.
8. L.-J. Zhang, "Editorial: Big Services Era: Global Trends of Cloud Computing and Big Data," *IEEE Trans. Services Computing*, vol. 5, no. 4, 2012, pp. 467–468.
9. W.M.P. van der Aalst, "A Decade of Business Process Management Conferences: Personal Reflections on a Developing Discipline," *Business Process Management*, A. Barros, A. Gal, and E. Kindler, eds. Springer, 2012, pp. 1–16.
10. W.M.P. van der Aalst, "Process Mining," *Comm. ACM*, vol. 55, no. 8, 2012, pp. 76–83.
11. M. zur Muehlen and R. Shapiro, *Handbook on Business Process Analytics*, Springer, vol. 2.
12. S. Rizzi, "Collaborative Business Intelligence," *Proc. First European Summer School* (eBISS 11), Springer, 2011, pp. 186–205.
14. C. Costello, "Incorporating Performance into Process Models to Support Business Activity Monitoring," doctoral dissertation, Dept. of Information Technology, National Univ. of Ireland, 2008.
15. *Business Process Analytics Format Specification*, Workflow Management Coalition (WfMC), Feb. 2012; www.wfmc.org/Download-document/Business-Process-Analytics-Format-R1.html.
16. O. Molloy and C. Sheridan, "A Framework for the Use of Business Activity Monitoring in Process Improvement," *E-Strategies for Resource Management Systems: Planning and Implementation*, E. Alkhalifa, ed.,

IGI Global, 2010.

***Alejandro Vera-Baquero*** *is a PhD candidate at the Universidad Carlos III de Madrid. His research interests include business process modeling, big data, and business analytics. Vera-Baquero received his MSc in software engineering and database technologies from National University of Ireland, Galway. Contact him at averabaq@gmail.com.*

***Ricardo Colomo-Palacios*** *is an associate professor at the Universidad Carlos III de Madrid. His research interests include applied information systems and software engineering. Colomo-Palacios received his PhD in computer science from Universidad Politécnica of Madrid. Contact him at ricardo.colomo@uc3m.es.*

***Owen Molloy*** *is a lecturer in information technology at the National University of Ireland. His research interests include enterprise computing, intelligent enterprise integration, and business performance management. Molloy received his PhD in industrial engineering from National University of Ireland, Galway. Contact him at owen.molloy@nuigalway.ie.*