

Business Provenance – A Technology to Increase Traceability of End-to-End Operations

Francisco Curbera, Yurdaer Doganata, Axel Martens, Nirmal K. Mukhi,
and Aleksander Slominski

IBM T J Watson Research Center, 19 Skyline Drive, Hawthorne NY 10532
{curbera,yurdaer,amarten,nmukhi,aslom}@us.ibm.com

Abstract. Today's enterprise applications span multiple systems and organizations, integrating legacy and newly developed software components to deliver value to business operations. Often business processes rely on human activities that may not be predicted in advance, and information exchange is heavily based on e-mails or attachments where the content is unstructured and needs discovery. Visibility of such end-to-end operations is required to manage compliance and business performance. Hence, it becomes necessary to develop techniques for tracking and correlating the relevant aspects of business operations as needed without the cost and overhead of a fully fledged data and process reengineering. Our business provenance solution provides a generic data model and middleware infrastructure to collect and correlate information about how data was produced, what resources were involved and which tasks were executed. Business provenance gives the flexibility to selectively capture information required to address a specific compliance or performance goal. Additionally, a powerful correlation mechanism yields a representation of the end-to-end operation that puts each business artifact into the right context, for example, to detect situations of compliance violations and find their root causes.

1 Introduction

In today's complex business environment, actual business operations often differ from their original design resulting in business integrity lapses and compliance failures with significant penalties. The cost of compliance with regulatory mandates such as HIPAA or the Sarbanes-Oxley Act has been higher than most companies expected. According to a survey, an average Fortune 1000 company spent more than \$2 million and logged more than 10,000 hours of work in 2005 [1]. In order to reduce the cost of compliance assurance, companies are seeking to automate manual process controls and reduce the amount of internal and consulting labor [2]. There is a general consensus that compliance solutions must be an integral part of organization's business process and enable a proactive approach to reduce risk. Such a solution cannot rely merely on the business models but should be based on the actual execution trace of end to end business operations. This way, operational aspects of the enterprise are captured, operational risks are measured, compliance to business rules and regulations can be assured, risk points are identified and actions are taken for remediation.

Tracking provenance as part of business process management is particularly important in the area of compliance, where the majority of spending goes to the labor of auditors and consultants to document and track the lineage of business tasks and items. Business provenance, presented in this article, is a technology developed to increase the traceability of end-to-end business operation in a flexible and cost effective way. Capturing and managing provenance systematically is recognized as a problem with significant impact [3], [4], [5]. Provenance technologies help to understand what actually happened during the lifecycle of a process by examining how data is produced, what resources are involved and which tasks are invoked. Accurate tracking of the lineage of the business process executions is essential to determine the root cause of compliance failures, but as the computers get faster and applications become more complex, tracking and processing large volumes of business data is an expensive proposal. Fortunately, in case of a specific compliance problem or to achieve a particular performance goal, it is not necessary to track all the events. The provenance of relevant business data can be identified and tracked selectively in order to reduce the complexity of the solution.

We define business provenance as capturing and managing the lineage of business artifacts to discover functional, organizational, data and resource aspects of a business. Examining business provenance data gives insight into the chain of cause and effect relations and facilitates understanding the root causes of the resultant event. Thus, business provenance technology is related to Business Process Management (BPM) and Business Activity Monitoring (BAM). While Business Activity Monitoring is mostly focused on real time access to business performance indicators, including interactive and real time dashboards and proactive alert generation, Business Provenance Technology adds a historical perspective to BAM that enables root cause analysis, process discovery and more. On the other hand, business provenance inherits BPM's process centric view and extends it to end-to-end business operations across BPM and non-BPM processes. Fig. 1 depicts the scope of Business Provenance Systems.

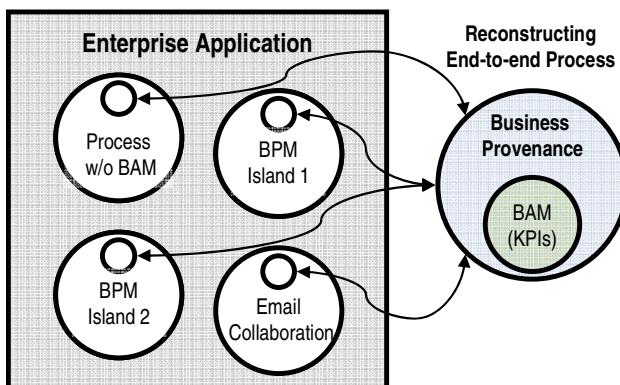


Fig. 1. Business Provenance solution

Business provenance technology focuses on automatic discovery of what actually has happened during the business process execution by collecting, correlating and analyzing operational data. Steps of the approach can be summarized as follows:

- Identifying data collection points that would generate data salient to operational aspect of the business is the first step of the solution. This requires understanding the business context. Information and documentation about the business operations, process execution platforms, and business models help determine the relevant probing points.
- A generic data model that supports different aspects of business must be in place to in order to utilize the operational data.
- Finally, the collected data must be correlated and put into the business context in order to have an integrated view.

The following sections provide the technical details of the steps outlined above and cover the implementation architecture.

The section below provides an overview of related work. The following section describes a compliance scenario that demonstrates how regulatory compliance is checked in a typical enterprise application; this scenario is references throughout the paper. Then, the technical approach of the business provenance solution is detailed which is followed by the description of some architectural components. Data collection, storage, data enrichment and data access processes are outlined. Towards the end of the paper, the generalization of the solution to support root cause analysis, continuous monitoring and proactive prevention of compliance violations are discussed. The article concludes with future work and concluding remarks.

2 Related Work

The problem of utilizing information technologies to increase productivity and reduce the cost of labor through automation has been around since companies started using computers to automate their business processes. Information systems were major enablers of automating business within and across organizational borders. In the 90s, the concept of business process reengineering (BPR) was introduced to improve performance measures such as cost, quality, service and speed by radically redesigning the existing business processes [7]. BPR created a lot of discussions and significantly influenced the way people think about their business, but did not stay long. Its rejection was mainly due to its radical approach to reengineering and its disregard for people within organization, resulting in a lot of criticism of the BPR concept. After some failures and misuses, the excitement around BPR has faded. In early 2000s, business process management (BPM) started to gain a major acceptance in the corporate world as the successor of BPR. BPM is also driven by using information technology to increase efficiency, productivity and manageability by following a more formal but less radical approach [8]. BPM encompasses design, modeling, execution, monitoring, analysis and optimization aspects of business processes. Today, suites of BPM tools and products are available from different vendors to help automate business processes and enhance productivity. These include IBM's Websphere Business Integration and

Filenet Business Process Manager, BEA's AquaLogic BPM Suite, Accents's AgilePoint, Adobe's LiveCycle Worflow, Oracle's BPEL Process Manager, PagaSystems's SmartBPM Suite, Ultimus's BPM Suite, ECM's Documentum, etc.

Our work in this paper is a contribution to the monitoring, analysis and optimization aspects of BPM discipline using provenance technology. Business provenance technology introduced here traces end-to-end business operations, based on specific instrumentation, to extract relevant information. Process mining [9], [17] is another technique that allows for process analysis based on event logs. Our approach differs from process mining technique in two ways. Firstly, enterprise applications generally span multiple systems and runtimes. Hence, single system traces may only partially capture information because they are not aware of the entirety of the business solution. As a result, data and event logs without proper correlation using the apriori knowledge of the business model may not capture all the salient features. Secondly, process mining does not treat data items as first class entities. In reality, the lifecycle of a data item may outrun the life cycle of the process. This may cause a discontinuity in the lineage information of the data item. Assurance of business integrity may require personalized and focused event and data probing. Events or relations that are important for a particular rule may not be captured by the logs. The logs produced by information systems are in general transparent to the business rules with which the business solution must comply. Their scope is limited to the process for which the log is generated. Since the design of a particular software component is generally made without taking a specific business rule into account, event logs may not always be useful in identifying a business integrity lapse. In particular, correlating log data across runtimes and platforms will not be possible by using traditional mining techniques when correlation may be crucial to detect integrity lapses.

Provenance has been an important subject for scientific work as establishing provenance can be decisive factor when determining validity of scientific results and assuring that they can be reproduced. With advances in availability of computing and data resources it is natural that responsibility to record provenance moves to software and becomes integral part of computer systems used to run scientific experiments. A survey of scientific provenance solutions can be found in [11]. Chimera [15] is an example of such system. In the life sciences area ^{my}Grid [12] is an example of a system that allows biologists to develop and execute *in silico* experiments. Unfortunately there is no standard representation or an interchange format available for provenance data between scientific projects [13] although the value of such standardization is recognized in scientific community. For example OPM, The Open Provenance Model [10], was proposed as a result of Provenance Challenge Workshops. OPM aims to define a common model for provenance - an annotated causality graph (a directed acyclic graph) that contains nodes to represent artifacts, processes, and agents. The execution of a scientific experiment is captured as a graph that connects generated artifacts, processes that work on them and agents that are responsible for process execution. OPM defines a set of standard relations between nodes such as “used” and “wasDerivedFrom”. OPM does not define any serialization, and some specification needs to be defined in future to allow exchange of provenance data.

There are similarities between OPM and our work. The main difference is that in a business environment, it is rare to have only one process (scientific experiment) whose provenance needs to be captured; instead there are many interconnected processes with different degrees of recording availability. Therefore the main challenge is not only to capture business provenance, but to correlate related pieces of information and construct end-to-end process traces. Correlation and relationship discovery are challenges in constructing provenance within a business environment. As an example, deciding the identity relation between multiple data pieces, such as records about customers, is a non trivial task that requires infrastructure, such as the Provenance Store we describe later, that can dynamically analyze data. The IBM Identity Resolution product is one solution proposed in this area [14]. It captures data into a common XML format and then utilizes a set of sophisticated analytics to discover relations: deterministic matching rules and probabilistic-emulating thresholds, relationship scoring, conditions under which to issue intelligence alerts, etc.

3 A Compliance Scenario

Regulatory compliance requires businesses to become aware of and take steps to comply with relevant laws and regulations. Especially in the area of financial accounting, violations of regulations such as the Sarbanes-Oxley Act may cause severe penalties. To illustrate a typical, cross-platform business transaction and its orthogonal compliance requirements, we take a closer look at the variable compensation of sales employees. Our example represents a simplified version of an actual process seen in a customer engagement. The process can be described succinctly as follows: A sales employee receives commissions for the generated revenue or profit as a variable part of his income. Managers create challenges that align these commission incentives specifically to the line of business, geography, and individual situation of the employee. A challenge is a document that describes in detail each sales target and the associated compensation. If an employee is able to provide evidence about the achievement of a particular challenge, the commission is added to his next pay statement as an incentive.

From a modeling point of view there is one end-to-end process that spans all activities from the creation of a particular challenge to the issuance of the corresponding payment statement. In practice, various distributed systems are involved in the execution of the process. These systems process structured and unstructured documents and run formal sub processes as well as ad-hoc tasks, increasing overall operational complexity. Fig. 2 illustrates the scenario.

In the first step, the manager creates the challenge (1) using a Web-front-end to the central Record Management System. This task triggers an automated email informing the employee about the challenge. To claim the achievement, the employee has to provide evidence (2) – which can take various forms: a contract or receipt, a fax from the sales customer, a pointer to a different revenue database, etc. Typically, the evidence is available electronically and it is attached to an e-mail sent to the manager by the employee. Upon reviewing the evidence, the manager evaluates the challenge and in case of achievement marks its status as “achieved” (3). Periodically, the latest

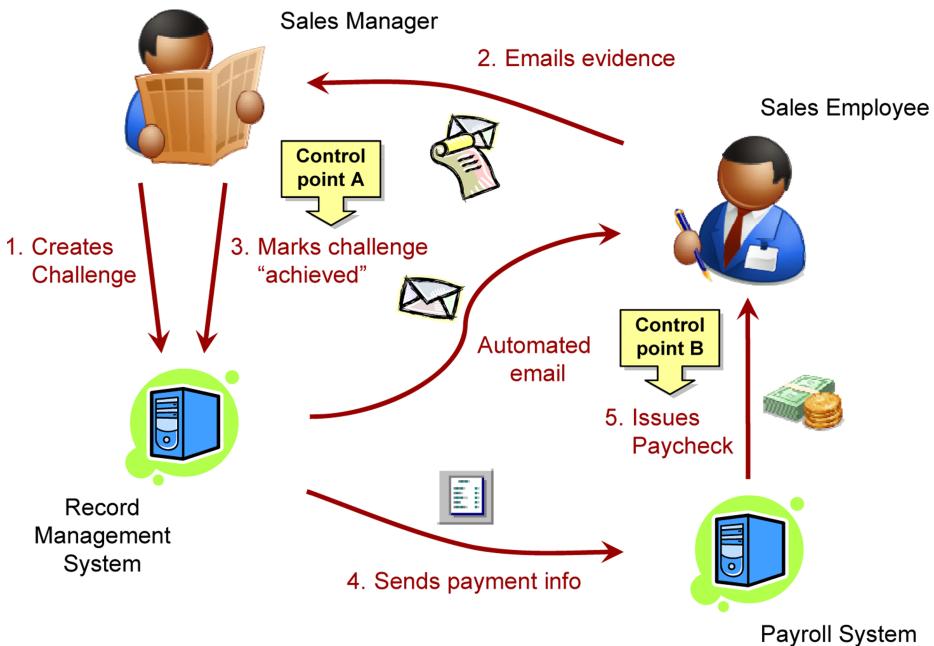


Fig. 2. Compliance scenario illustrating end-to-end business operation

achievement data is collected and fed into the Payroll System. Finally, the paycheck is issued to the employee.

In order to assure compliance of the overall process with legal accounting regulations, various control points are introduced. Each control point reflects one locally verifiable requirement that is currently manually performed for a small number of sampled transactions by internal and/or external auditors. Typically, control points are established for the interaction of various systems, and verification of the control point requires the correlation of structured and/or unstructured data. **Fig. 2.** shows two control points. Control point A requires the manager to obtain, evaluate carefully, and maintain the evidence of any achieved challenge. Control point B requires the paycheck to reflect the accumulated commissions correctly. To verify control point A, an auditor selects an achieved challenge, requests the evidence, and compares the sales targets with the documented achievements. This seemingly simple task has proven to be quite complicated in practice. Firstly, the evidence is not directly linked to the challenge. In some cases it is not even stored in a central repository but kept locally by the manager. The auditor therefore has to contact the manager, and the manager has to find the right documents. Our observations have shown compliance failure rate of 70%, largely because the evidence could not be located. Also, we have observed lengthy email exchanges between an auditor and a manager until the correct evidence could be identified. As a result, only a small fraction of the total number of transactions can be sampled and a high number of undetected questionable situations and possibly fraud may occur. In addition, there has been no support available to track down the root-cause when a questionable situation was detected. This is a major

drawback of the existing auditing method. To enable businesses to prevent future wrongdoing, or simply to detect a pattern of fraudulent behavior, it is essential to answer the causality question: “Why did this happen?” Our proposed business provenance technology targets exactly this question.

In the given example, one might argue that the process is not well designed. But regardless how carefully an application is architected, there will always be gaps between the different systems involved, there will always be data that does not fit into predefined forms, and there will always be exceptions in the execution. Rather than requiring a full scale, heavyweight data integration, our approach focuses on recording meta-data concerning relevant objects and events into a centralized and easily accessible store with links into the original systems; automating correlation of the meta-data to establish execution traces, versioning histories, and other relevant relations; and finally deep analysis to detect situations after the fact, raise alerts while monitoring continuously, and even interfere with execution to prevent compliance violations.

4 Technical Approach

Ensuring business integrity requires understanding how business operations are executed, who are the people involved, and where supporting documents and evidences are found. Thus, probing into different perspectives of the business operations may be necessary to validate compliance with operational rules and regulations. This way a more specialized and effective diagnostics will be possible. At a high level, our technical approach consists of the following steps: (1) Identifying the control points, relevant business artifacts and required correlations; (2) probing the actual execution of the business process to collect data; (3) correlating and enriching the collected data and the relations among them to create a *Provenance Graph*; (4) analyzing aggregated information to enable business activity monitoring or to interfere with the execution by generating alerts; (5) providing access to information stored in the graph for detailed investigation and root cause analysis.

Fig. 3 shows the components of the technical approach. The management component supports the specification of the provenance data model, i.e. the list of business objects to be captured and the level of details. Recording components are used to capture, process, and reformat application events of the underlying information system and record the meta-data of business operations into the provenance store. The analytics component links and enriches the collected data to produce the provenance graph (that will be discussed in greater detail in the later sections). To do so, the analytics components have access to the content of the provenance store as well as the original business data. The enriched business data is accessed through a query interface and analyzed to verify business control points in two different styles. Firstly, a query can be deployed into the provenance store to emit results in real-time, feeding existing dashboard systems to display key performance indicators for example. Secondly, a query frontend enables visualization and navigation through the provenance graph from the outside.

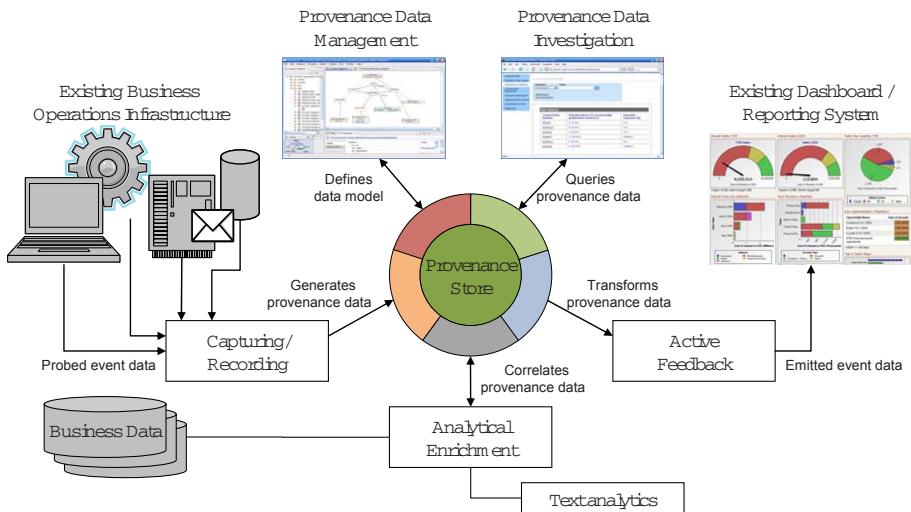


Fig. 3. Architecture of technical approach

The following sections explain how we set business control points, model business data, create a provenance graph that depicts various aspects of the operations, represent control points, enrich the graph and extract information from the graph.

4.1 Compliance Scenario Walkthrough

Business integrity requires that tasks and activities executed to achieve a business goal within the company are compliant with internal or external rules or regulations. Compliance goals are identified by examining business rules and deciding what action steps are needed. In other words, from the business rules expressed in the language of business people, compliance goals are identified. Identifying compliance goals lays the ground work for setting up IT rules for compliance. Once the compliance goals are identified, tasks, activities, resources, artifacts and relations relevant to the identified goal are determined. Business control points are expressed in terms of these artifacts. The control points provide a bridge between various components of the business operations and the actual data that could be consumed by the IT system.

Fig. 4 depicts the flow of steps for checking compliance. It starts by converting business rules and regulations into compliance goals and creating a data model to capture various aspects of the business. The relevant business aspects are captured as provenance data by recording probes. The compliance goal's scope is gathered by correlation rules to create the control point pattern. For example:

Business Rule: “A sales challenge can be considered achieved after the supporting evidence is received and evaluated by the first line manager and found satisfactory”. The business rule defines what it means to achieve a challenge in the language of business people.

Compliance Goal: “Ensure that there exist documents that support achieved challenges”. The compliance goal explains what needs to be done to comply with the rule. The data model specifies the required information of sales challenges and supporting documents, managers and employees. This information is derived from the compliance goal and broken down into recording configurations and correlation rules. The provenance store materializes the data records of challenges, manager, and documents at runtime, linked appropriately by the correlation rule. In this particular case, the data model must support employee id, claim and challenge statements, manager ids, status of challenges, etc.

Business Control Point: “Request documentation about the employee’s achievement with respect to the challenge from the manager; check if the challenge has been achieved.” The control point pattern in the provenance store links the records of the challenge, the manager and employee, and the documentation (cf. Fig. 2). Analytics can then be developed to process operational data to determine potential violation and generate alerts. In the example given above, integrity is broken if the challenge is marked achieved without supporting evidence. Auditors look for the supporting document to check whether the business rule is satisfied. They also look for information traces of tasks executed across runtimes and processes and links between business artifacts, tasks and the people who executed the tasks. They need to know when a challenged is marked achieved, who approved it, who claimed it and where the supporting documents are. Capturing the execution trace helps the auditors to answer the questions related to a particular business control point. The next section illustrates the provenance data model using the initial compliance scenario.

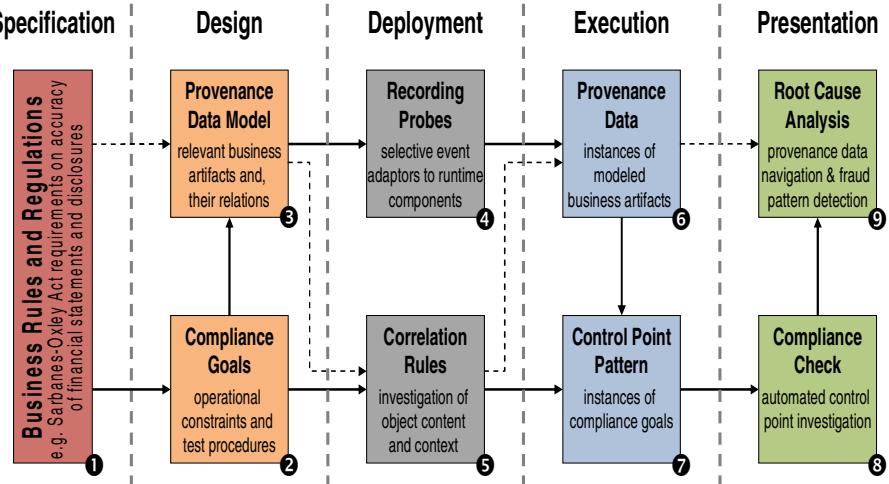


Fig. 4. Steps for compliance checking

4.2 Creating the Provenance Data Model

Ensuring compliance through information system requires laying out a data model that covers the relevant aspects of the business operations. Creating a data model is

the first step to bridge business operations to information systems. The data model should support relevant and salient aspects of the business. Our business provenance solution comes ready with a comprehensive, generic data model that can be extended to meet the domain specific needs. To illustrate the data model Fig. 5 provides a first view on the relevant business artifacts and their relations with respect to the previously explained compliance scenario.

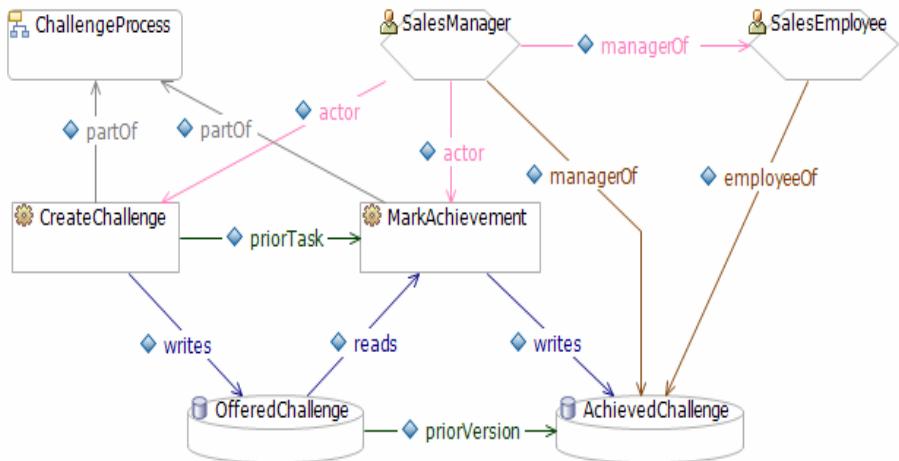


Fig. 5. Provenance data for compliance scenario

Analogous to the five dimensions of business process modeling with ARIS [16], the business artifacts stored in the provenance store falls into one of the following five dimensions:

Data Records: A data record (shown as a cylinder) represents a business artifact that was produced or changed during execution of a business processes. Typically those artifacts include documents, e-mails, and database records. The provenance store represents each version of such an artifact separately. Thus, two data records for the same challenge document can be seen in Fig. 5: one represents the challenge in the state “offered” and another in the later state “achieved”.

Task Records: A task record (shown as a rectangle) is the representation of the execution of one particular task. Such task might be part of a formally defined business process or being stand-alone, might be fully automated or manual. In Fig. 5, both task records are part of the challenge process. As a task manipulates data, a relationship is created between the corresponding task record and the affected data records.

Process Records: A process record (shown as a rounded rectangle) represents one instance of a process. In automated business management systems, tasks are executed by processes. Hence, each task record associated with the corresponding process record.

Resource Records: A resource record (shown as a hexagon) represents a person, a runtime or a different kind of resource relevant to the selected scope of business provenance; an actor of a particular task is an example of such a record. In Fig. 5, an employee and her manager are represented, both related to the achieved challenge.

Custom Records: Custom records provide an extension point to capture domain specific, mostly virtual artifacts like compliance goals, alerts and checkpoints. The next section will provide greater details.

Nodes of the provenance graph represent these five classes of records. Edges between nodes are made based on correlation between two records, *Relation Records* represent the edges. These are the records generally produced as a result of relation analysis among the collected records. We only consider binary relations between records. However, relations between relation records are possible and such higher degree relation could be expressed. Some relations are rather basic on the IT level, like the read and write between tasks and data. Other relations are derived from the context, like that between manager and achieved challenge.

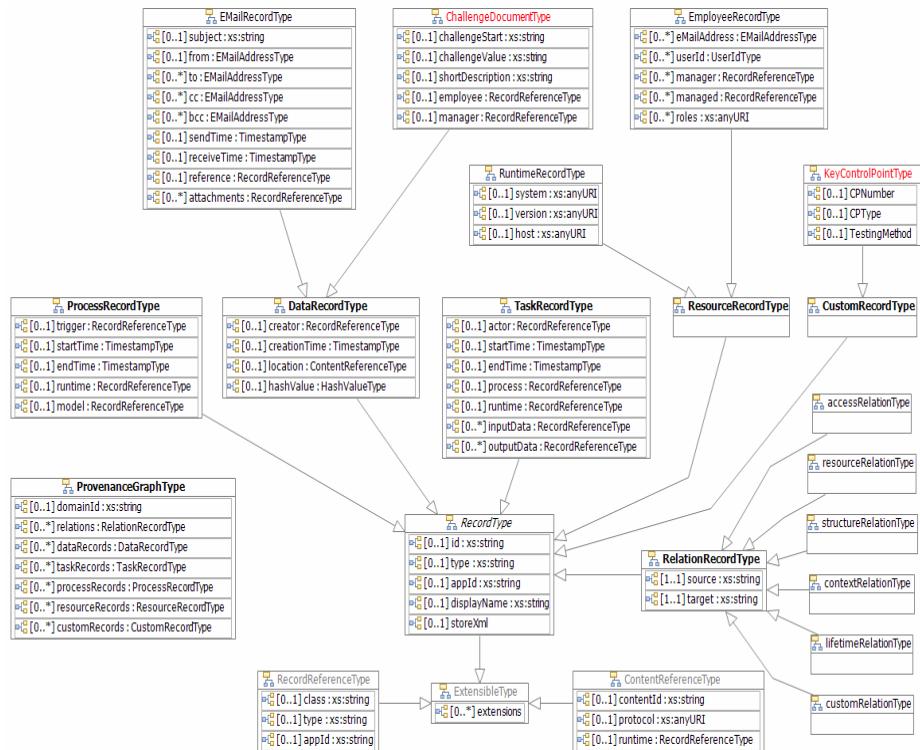


Fig. 6. Class diagram of basic provenance data model

As mentioned before, our business provenance solution comes with a generic data model that can be extended to meet the application domain specific needs. Figure 6 shows an extract of that data model. Basically, the provenance graphs consists of six different sets of records, while each record is an extensible xml data structure and all records share common attributes: *id* and *type* are used to identify and classify the record within the graph; the *appId* (application specific id) and *displayName* refer to

characteristics of the corresponding business artifact. Data, task and process records are added to the provenance graph as the business operations are executed. Resource and custom records are often added after the fact by analytics. Details on the creation of a resource record follow in the next section. Figure 6 shows several specializations of the basic record types. Some of them are of generic nature, too. The challenge document and key control point type, however, are specifically added for the compliance scenario.

4.3 Generation of Provenance Data

Provenance records reflecting data objects, tasks or processes are created by the recording probes that are small adapter components that are able to listen to event mechanisms of the underlying information systems. Recreating the end-to-end process based on provenance records requires that those records are connected together. As described above, this naturally translates into creating edges in the provenance graph by adding relation records. This process may require multiple steps. Basic relations between a task and the manipulated data or a task and the corresponding process instance can be established based on the information the task record holds. We use the concept of record references to materialize this information upon creating the task record. For example when the “create challenge” task (cf. Fig. 5. Provenance data for compliance scenario) is executed, its recording probe adds to the task record a reference to the offered challenge based on the best available application specific identification, for example the *challenge id*. The analytics component tries to locate the correct data record in the provenance graph, and creates a relation between those two records if successful.

Other relations are established by utilizing data outside of provenance graph, such as data stored in content repositories. Additionally text analytics may be used to process emails, to categorize them, and to extract from the emails data that is then stored as new entries in provenance graph. Other agents may run process mining and discovery algorithms to create new relations that identify potential processes. Creation of relations and entries may trigger other agents that may add new entries that may be later processed. We call that incremental process of graph building ‘enrichment’ and it is the key capability needed trace end-to-end processes. The underlying provenance graph gets continuously enriched, as the creation of some relations may trigger execution of other enrichment rules. As relations between provenance records are established, the hyperlinked structure provides for each such record a context that describes the lineage of its existence, a path into related events that had occurred prior to its existence, and related events that had happened later.

Recording and linking provenance records into a centralized repository is a prerequisite for compliance operations. This data becomes meaningful information only when placed in the appropriate context, such as the business transaction that it is part of. Business provenance is architected to allow the specification of rules that enrich the provenance. As described above, enrichment can take the form of the execution of a trivial rule that creates a relation based on knowledge of the business data formats,

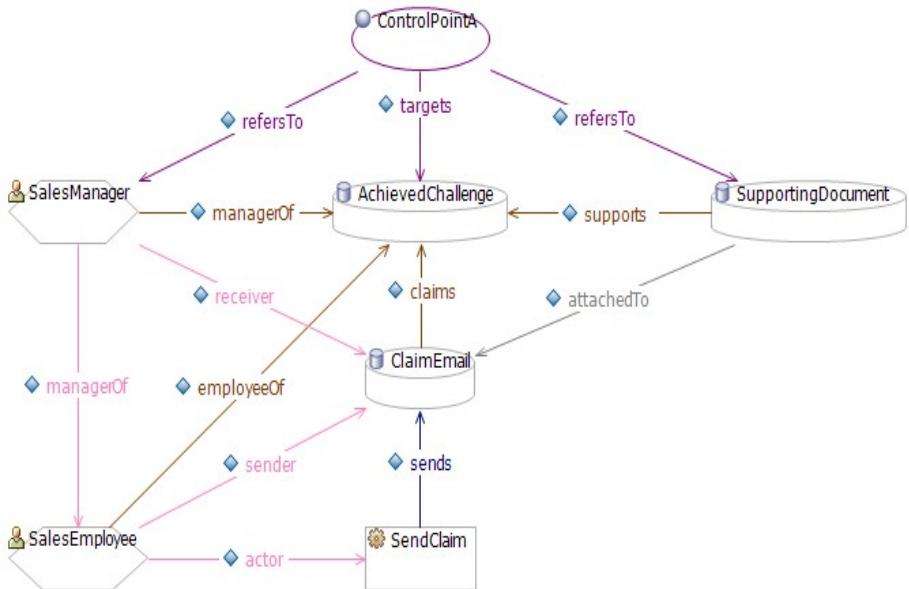


Fig. 7. Depiction of a business control point

or a non-trivial execution of a text analytics component that, for example, relates emails to workflow tasks or application events.

4.4 Representing Business Control Points

The resource, task and data records represent organizational, functional and data aspects of the business as defined above. A relation between a resource record and a task record shows who was involved in executing that particular task. A relation between a data record and a task record reflects the effects of the task on the business artifact or the task dependency on its availability. An integrity goal often combines those aspects by describing which tasks should be performed, when, how and who should be involved. A business control point can be created using this description in terms of provenance graph entries. A business control point is satisfied if certain vertices and edges exist in the provenance graph. Hence, it is possible to claim that a business control point is a sub graph of the provenance graph.

Looking back on the compliance scenario, a business control point required the conclusive supporting documentation to be present at the time the manager marks the challenge achieved. This control point is materialized as a sub graph or pattern of the provenance graph as shown in Fig. 7. For each data record that represents an achieved challenge, a control point node (an extended custom record) is added to the graph and linked to the achieved challenge. The properties of the challenge recorded into the provenance graph include the identifier for the manager and employee, so it is straightforward to build relations between this control point instance and the corresponding manager and employee nodes using a simple analytics rule.

Sometimes relations are built on the basis of more sophisticated analytics; for example, to build the *supports* relation between a SupportingDocument and AchievedChallenge, the analytics component needs to discover an email that references this particular challenge and included this document as an attachment, which requires the use of text analytics. Such rules help complete the provenance graph and are conditionally triggered whenever the graph undergoes a change via the addition of a new record.

Having a physical representation of the control point materialized in the provenance graph, it becomes much easier to check. If there is one control point node without a link to a supporting document the business rule is clearly violated. If there is a supporting document found, an auditor can easily retrieve its content as well as the original challenge's content to verify the evidence. To some extent text analytics might provide red flags in advance to direct the auditor's attention.

4.5 Extracting Information from Provenance Data

Extracting information about the business operations from the Provenance Graph is equivalent to extracting the properties of the graph or discovering various sub-graph patterns. The nodes, associated edges and sub-graphs reveal various aspects of the business. Our business provenance solution provides two different ways to extract information. Firstly, the analytics component provides as mentioned before the capability to detect patterns in the graph. Besides creating new provenance records and relations, the detection of a pattern can trigger propagation of an alert to an existing dashboard or monitoring system. For example: recording an achieved challenge triggers the creation of control point pattern as explained before. A control point node that has no link to a supporting document represents a potential violation of the business rule. Hence, an alert is raised. The rich contextual information our business provenance data provides supports more complex and more subtle alert situations than traditional business activity monitoring. Secondly, provenance data enables extensive root cause analysis once an alert or a situation of interest has been identified. Users can navigate directly on the graph to query the information, zooming in and out, switching pivot and perspective, walking and aggregating links, etc. Alternatively, the user can interact with an application domain specific Web front-end that presents the most relevant records and relations. We have successfully tested this approach in our customer engagement, and envision having this front-end generated automatically based on the provenance data model.

5 Data Collection and Storage

Fig. 8 shows the connection between recorder clients and the provenance store. The recorder client is an important component of the overall business provenance architecture. It processes application events and transforms them into provenance events. The responsibilities of a recorder client are to access application data, create provenance events, and record provenance events in a provenance store.

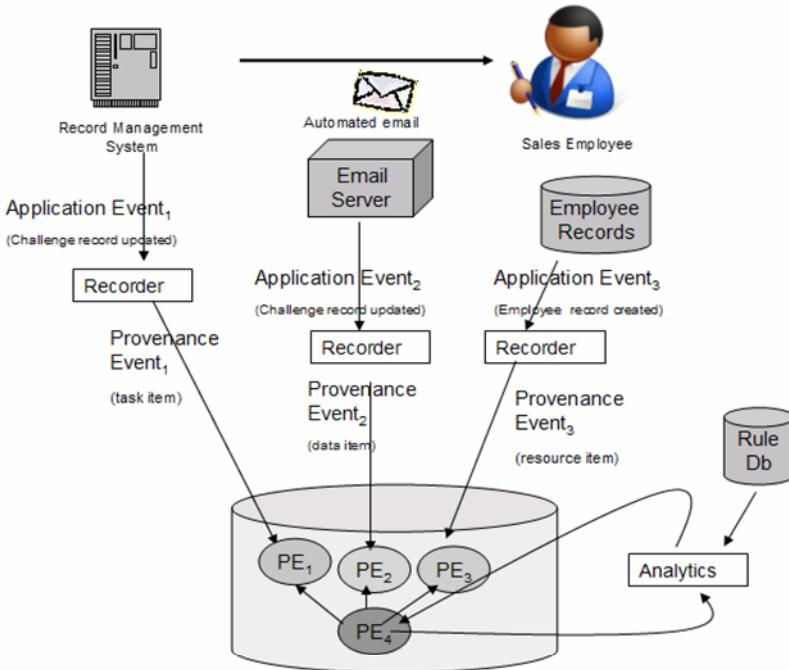


Fig. 8. Data collection and storage

5.1 Access Application Data

This may require access to underlying system logs, email repositories, or user specified directories where documents are stored. If applications are already producing events (for example using IBM's CEI [19] infrastructure or event reporting provided by other middleware products) then the recorder's job is easier as it just needs to capture those events. Fig. 8 shows various recorder clients collecting events from different applications.

5.2 Create Provenance Events

A provenance event should contain a subset of application data that needs to be stored as business provenance. The captured data must be relevant and specific to business control points. To avoid redundancy and possible exposure of sensitive data, recorder clients should not copy all application data, particularly if the application data is stored in another repository. For different kind of applications and different parts of customer organizations, there may already exist content repositories with specific retention and access control policies. If such content repositories exist then they should be referenced and provenance events should only contain location references

and extracted metadata that is relevant to fulfill business goals, such as tracing end-to-end process execution. Identifiers that can be used later to discover how recorded events relate process instances are crucial to record as part of provenance events. Fig. 8 shows that three types of provenance events, namely, task, data and resource events are created from various application events.

5.3 Record Provenance Events in Provenance Store

To facilitate recording of provenance events in a distributed environment and enable scalability, Web service APIs to the provenance store, such as a SOAP endpoint, are provided.

5.4 Storage of Provenance Events

The provenance store is a repository for the provenance events created by recording clients. There are many possible representations for events with many ways to store them in the repository. We choose to use XML elements for event representation and store them directly in database tables. In the example depicted by Fig. 6, the recorded data may be stored in two tables of an SQL database as shown below:

Table 1. Content of database table *Entry*

ID	CLASS	APPID	XML
PE1	task	challengeDoc23	<recordUpdate ps:id="PE1" ps:class="task"><documentId>challengeDoc23</documentId><status>achieved</status></recordUpdate>
PE2	data	Emai20320932	<email><subject>Challenge Achieved</subject><id>Emai20320932</id></email>
PE3	resource	Joedoe	<person><name>Joe Doe</name><userid>joedoe</userid></person>
PE4	relation		<controlPoint><description>...

Table 2. Content of database table *Relation*

ID	SOURCEID	TARGETID	XML
PE5	PE4	PE1	<relatedTask/>
PE6	PE4	PE2	<supportingDocuments/>

The Entry table stores the content of recorded provenance events as XML and extracts from events common attributes such as class types and IDs. The Relation table stores relations between entries and make it easier to traverse the graph of interconnected entries. Those can be used to make queries execute faster. If the database system does not support the XML query language then additional tables and processing may be needed to “shred” the XML document tree into relational data.

While recorders are capable of actually capturing the relevant artifacts themselves, generally we expect that only metadata will be persisted into the provenance store; the actual artifact will be linked from the provenance record so it can be recovered as needed.

6 Application to Compliance Problems

In the section 3 we identified a set of features that a compliance solution should have. Here we describe how business provenance technology provides these features, and can thus be the technical basis for a compliance solution. In the next section, we will describe other uses of business provenance.

One of the core parts of the compliance process is testing operational data to validate compliance against a set of policies and procedures. Testing thus depends on the ready availability of operational data. Business provenance is designed to capture relevant business data and events through appropriate configuration of recorders as described in section 4.

6.1 Support for Root Cause Analysis

Current compliance processes tend to have many manual tests and a few automated tests. Root cause analysis is one aspect of the process that is always manual. This is also amongst the most time-consuming parts of the process since it often involves contacting many people, looking at various pieces of data and evidence and so on. Business provenance, through enrichment of the provenance graph, provides many views onto the provenance items and relations. Through appropriate filtering, it can provide users with a time series view of events, a look at a single user's actions, and other navigational views over semantic relations. One view that was useful to auditors for incentives and commissions was the end-to-end process view that related offered challenges to achieved challenges, associated supporting documentation for claims, and finally related the achieved challenge to the payment statement issued to the employee. This gave auditors the ability to understand *when* a particular defect (such as missing supporting documentation) took place, *who* is responsible for it, and finally to trace *why* it happened by looking at the pattern of behavior for the responsible sales manager. This ready support for root cause analysis without necessitating the cooperation of the sales people in the field resulted in a more efficient and effective compliance process.

6.2 Continuous Monitoring

Business provenance supports continuous monitoring using rules. We have already discussed rules that enrich the provenance graph through the addition of relations. Additionally, we have another application for rules: a second set of rules authored specifically to look for patterns within the structure of the graph that might reflect potential compliance violations or other concerns. Such rules highlight of the presence of a pattern through the creation of a special *alert* node within the provenance graph, and also the creation of relations between the alert node and relevant information. For example, in incentives and commissions, we authored a specific rule to look

for a graph pattern consisting of an achieved challenge and corresponding supporting documentation. This rule was triggered shortly after a challenge was marked achieved. In the absence of related supporting documentation, the rule created an alert node and linked it to the offending challenge.

Such alert nodes are continuously produced as the process unfolds, and can be exploited via an appropriate user interface to alert the user of a potential compliance violation. In our example, we designed a web-based user interface for auditors that highlighted the current alert situation up front so that they could be addressed immediately. Other possible modes of alerting users of the provenance system would be through email or text messages, both of which would be simple additions to our system. Once the alert is known, users can use the different navigable views to examine its cause and determine the most effective way of addressing the situation.

6.3 Proactive Prevention of Compliance Violations

We have described how provenance can detect compliance violations on a continuous basis through the creation of alert situations, and how manual investigation of alert situations can be aided using the provenance graph. The provenance store is already connected to IT systems running business operations through recorders. This connection as we have described thus far is passive, wherein recorders merely send relevant application events to the store as provenance items. A possible use of provenance is to make this connection active, and allow a provenance component to affect the operation. Many business applications offer management interfaces to interrupt the application, suspend the handling of a particular transaction and so on. Most workflow engines also offer this kind of support. Using recorded provenance information, it might be possible to *predict* a future failure, based on knowledge of the process or on statistical data that allows us to make reasonable predictions. If the probability of failure is high, the active provenance component connected to the IT system can take the appropriate actions to affect the operations and highlight the situation for human investigation through an alert. We did not implement prevention of violations in our system, but this is one of the research areas we will be focusing on this year.

The next section discusses other application of business provenance, going beyond support for the compliance process.

7 Concluding Remarks and Future Work

This paper introduces business provenance as a new technology to trace end-to-end business operations. Today's enterprise applications that are often built on top of a heterogeneous and distributed IT infrastructure that often comprise both structured and unstructured processes and data, and that are often hard or costly to change or re-engineer. Legal and market requirements affecting these applications change frequently. To effectively deal with this environment, we have proposed business provenance as a light-weight approach to monitor the business operations. The core infrastructure consists of a DB2 database and a set of adapters interfacing with existing information systems.

Our approach is goal driven: Depending on the requirement – e.g. to assure compliance with respect to legal regulations, to monitor key performance indicators of end-to-end processes, or to discover process dependencies, etc. – our approach gathers and correlates only the information needed. This information defines the lineage, context and dependencies of business artifacts. It can be used for real time monitoring, after the fact investigations or even interference with the actual business operation.

We provide a generic provenance data model that can be extended and tailored to specifics of the application domain. Both, our middleware infrastructure and our provenance data model were put into a first practice test during a customer engagement. There the increased visibility of end-to-end operations helped to identify previously hard to spot situations of potential fraud and thus immediately reduced the risk of Sarbanes-Oxley compliance violations.

Our current research efforts focus on three main areas. Firstly, there is the need for tool support of the process for extending the provenance data model and deploying the business provenance solution. Such an integrated development environment for provenance will be able to generate configuration files, analytics rules and Web front-ends and thus coordinate recording, enrichment and query of business provenance. The second research focus targets a tight integration of business provenance with standard middleware products, observing and interfering, and thus extending the range of business activity monitoring to end-to-end business operations. Finally, we are working to improve the interaction with the provenance data on a semantic level by using natural language processing such that domain specialist can browse and query the data using their own vocabulary and business language.

References

1. Greengard, S.: Compliance Software's Bonus Benefits. *Business Finance Magazine* (February 2004)
2. Gartner, Simplifying Compliance: Best Practices and Technology, French Caldwell, 6/6/05 (Business Process Management Summit 2005)
3. Freire, J., Koop, D., Santos, E., Silva, C.T.: Provenance for Computational Tasks: A Survey. *IEEE Computing is Science & Engineering* 10(3), 11–21 (2008)
4. Simmhan, Y.L., Plale, B., Gannon, D.: A Survey of Data Provenance in E-Science. *SIGMOD Record* 34(3), 31–36 (2005)
5. Bose, R., Frew, J.: Lineage Retrieval for Scientific Data Processing: A Survey. *ACM Computing Surveys* 37(1), 1–28 (2005)
6. Miers, D., Harmon, P., Hall, C.: The 2007 BPM Suites Report,
http://www.bptrends.com/reports_toc_01.cfm
7. Hammer, M.: Reengineering Work: Don't automate, obliterate, *Harvard Business Review*, 104–112 (July/August 1990)
8. Harvey, M.: Essential Business Process Modelling ISBN 0-596-00843-0
9. Aalst, W., van der Weijters, A., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering* 16(9), 1128–1142 (2004)
10. Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J., Paulson, P.: The Open Provenance Model Technical Report, *Provenance in Scientific Computing* (2007),
<http://eprints.ecs.soton.ac.uk/14979/>

11. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-Science. ACM SIGMOD Record 34(3) (September 2005)
12. Lord, P., Alper, P., Wroe, C., Stevens, R., Goble, C., Zhao, J., Hull, D., Greenwood, M.: The Semantic Web: Service discovery and provenance in my-Grid. In: W3C Workshop on Semantic Web for Life Sciences (2004)
13. Second Provenance Challenge Workshop (SPC) in High-Performance Distributed Computing (2007), <http://twiki.ipaw.info/bin/view/Challenge>
14. Jonas, J.: Threat and Fraud Intelligence, Las Vegas Style, Security & Privacy, vol. 4(6), pp. 28–34. IEEE, Los Alamitos (2006),
<http://jeffjonas.typepad.com/IEEE.Identity.Resolution.pdf>
15. Foster, I., Vöckler, J., Wilde, M., Zhao, Y.: Chimera: A Virtual Data System For Representing, Querying, and Automating Data Derivation. In: Conference on Scientific and Statistical Database Management (2002)
16. Scheer, A.-W., Nüttgens, M.: ARIS Architecture and Reference Models for Business Process Management Institut für Wirtschaftsinformatik, Universität des Saarlandes, Im Stadtwald Geb. 14.1, D-66123 Saarbrücken
17. Rozinat, A., van der Aalst, W.: Conformance checking of processes based on monitoring real behavior. Information Systems 33(1), 64–95 (2008)
18. Fu, S.S., Chieu, T.C., Yih, J.-S., Kumaran, S.: An Intelligent Event Adaptation Mechanism for Business Performance Monitoring. In: IEEE International Conference on e-Business Engineering (ICEBE 2005), pp. 558–563 (2005)