Document Provenance in the Cloud: Constraints and Challenges

Mohamed Amin Sakka,[1,2] Bruno Defude[1] and Jorge Tellez[2]

[1] Novapost, Novapost R&D, 13, Boulevard de Rochechouart 75009 Paris-France
{amin.sakka, jorge.tellez}@novapost.fr
[2] TELECOM& Management SudParis, CNRS UMR Samovar, 9, Rue Charles Fourrier 91011 Evry cedex-France
{mohamed_amin.sakka,bruno.defude}@it-sudparis.eu

Abstract. The amounts of digital information are growing in size and complexity. With the emergence of distributed services over internet and the booming of electronic exchanges, the need to identify information origins and its lifecycle history becomes essential. Essential because it's the only factor ensuring information integrity and probative value. That's why in different areas like government, commerce, medical and science, tracking data origins is essential and can serve for informational, quality, forensics, regulatory compliance, rights protection and intellectual property purposes. Managing information provenance is a complex task and it has been extensively treated in databases, file system and scientific workflows. However, provenance in the cloud is a more challenging task due to the traditional problems in provenance management in the mentioned domains add to that the specifies related to the cloud.

1 Introduction

With the advent of cloud computing and the emergence of web 2.0 technologies, the traditional computing paradigm is shifting new challenges. This new paradigm brings issues related to the trustworthiness of data as well as computations performed in third-party clouds. Certainly, the digital form of electronic data has many advantages of being easily accessible, accurate and more useful. However, the mobile and modifiable nature of digital entities requires specific metadata and techniques ensuring information probative value. More specifically, in a cloud-computing scenario, most of the data will reside in data clouds, while the applications will run in the cloud. In this case, users require additional insurances regarding the confidentiality, privacy and integrity of information. Required metadata is provenance and can be defined as information about the origin, context or history of the data. Depending on the domain, provenance is useful for many purposes and its systems can support different types of usage. According to studies on information provenance [9, 5, 7], it can serve for audit trail and justification, regulatory compliance and forensics, replication recipes, attribution and copyright, data quality and informational purposes. Despite it

seems obvious, information provenance issues introduces hard challenges. Provenance is not obvious because it overtakes widely recording the whole history of the data [14]. Provenance is listed as a hard problem in some information science recent surveys like the UK Computing Society's Grand Challenges in Computing [10] and the InfoSec Council's Hard Problems List. Provenance is challenging because it's difficult to collect and is often incomplete, it comes from untrusted and unreliable sources, it's heterogeneous and non-portable and it requires confidentiality and privacy insurance. In this paper, we present information provenance challenges in the cloud and how it can serve to enhance information integrity. Especially, we focus on services for electronic documents and more precisely storage and archival services. Traditionally organizations and companies buy storage systems for each of its sites or acquire equipment for its data center locations. Generally, this represents enormous costs especially when we take into account the cost of managing, supporting, and maintaining these systems. For these reasons, outsourcing storage and archival to external clouds is a good alternative to reduce costs. It permits also to discard the hard task of locally managing and maintaining storage systems according to legal compliance and security recommendations [11]. However, this leverages many issues about probative value and implies an immediate need to information provenance and lifecycle.

This paper is organized as follows. Section 2 presents the case study illustrating provenance needs and challenges for documentary clouds. Section 3 generalizes these constraints and categorize them. The fourth part presents provenance in the cloud challenges. Section 5 presents provenance management approaches, some provenance related works and their lacks compared to the expected solution. Section 6 shows ours solution for the mentioned case study and section 7 concludes.

2 A bank record use case

2.1 Context

We are going to present a case study illustrating information probative value and provenance challenges for electronic documents. This case study comes from Novapost [3] which is a French company specialized on providing collect, distribution and archiving services for electronic documents (especially human resources documents like pay slips, or personal confidential documents like bank records). Novapost aims to consolidate the existing services by providing dematerialization and long term archiving solutions allowing to:

- Improve archive management: formats migration, information integrity, document lifecycle management, interoperability and reversibility issues.
- Ensure compliance with regulations relative to the probative value of electronic documents.

[3] Novapost: www.novapost.fr

– Seamless integration with customers through SaaS service provided in partnership with IBM and based on innovative technology and highly secure infrastructure.
– Ensure continuity and reliability in the management of information lifecycle and its provenance.

The case study scenario is the following: a customer of Novapost services (a bank) sends its customers bank reports to Novapost in AFP [4]format. Novapost documents service (provided as a cloud service) applies a specific processing depending on the received document type (splitting, templating, copying), then generates individual bank report in PDF/A [5] format for archiving, interacts with external trust authority (signature service composed by a signature server and a timestamp server), communicates with IBM cloud for legal archiving to archive them in IBM's datacenters. The archived files are accessible to bank customers via web portal or a widget plugged on the bank website (cf. Fig 1).

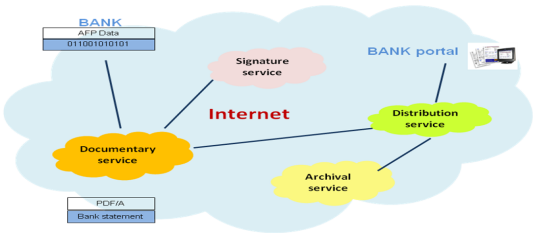

Fig. 1. Case study scenario

2.2 Case study constraints

As we have mentioned before, the heart of Novaposts solutions is document lifecycle management and documentary workflows requiring high probative value of the processed documents. The purpose of the R&D department in Novapost is to provide solutions to address the problems confronting dematerialization actors. We address the following issues:

– How to ensure the probative value of documents?
– How to ensure full document lifecycle traceability?
– How to guarantee the readability and the sustainability of processing and exchanges history for long term?

[4] AFP is a document format originally defined by IBM to drive its printers and support the typical form printing on laser printers.
[5] PDF/A: http://fr.wikipedia.org/wiki/PDF/A-1

In France, the probative value of electronic documents is subject to compliance with the conditions requested by the Article 1316-1 of the Civil Code mandating to identify the emitter persons/entities and to provide an integrity insurance through the whole document lifecycle. To ensure documents probative value and to perpetuate the informational capital of electronic archives, we need to keep a plausible trustworthy history of their provenance. This explains the strong requirements for traceability mandated by record management standards [15, 17, 19]. If we consider the standard lifecycle of a document, it includes steps of generation, additional processing, transfer and archiving. This means that document's technical integrity could change while its legal integrity is the same. At present, traceability management solutions for informational flows, especially documents are based on an accumulation of log files, coming from heterogeneous systems and multiple remote legal entities. Document provenance information's are distilled in multiple files that we need to cross to retrace its history. Add to that the sustainability of the responsible entities for the conservation of these trace elements is also problematic. In this context, the main difficulties in the reconstruction of the history of a document or any information flow are:

- The identification of log files about the processing of the document known that document identifiers are not the same across different systems. These track elements are distributed across multiple systems and managed by multiple legal entities.
- The heterogeneity of formats and data structures that can affect the readability of the trace of the document due to the heterogeneity of systems or to confidentiality constraints.
- The heterogeneity of the different information management policies in each participant system.
- The trust and the reliability about track elements provided by external systems/services during their creation or their transfer.

3 A generalization of provenance constraints

Information provenance in the cloud is considered as a transverse problem, at the crossroad between different types of constraints. According to our study, provenance constraints can be categorized into legal, business and technical constraints:

- Legal constraints relatives to the regulatory compliance about conservation duration of personal information as well as compliance with legislation governing information privacy like CNIL recommendation in France [6] or the CDT in USA [7]. Add to that the recent international law and the principle of territoriality which is not always in harmony with the distributed nature of the clouds over countries.

[6] CNIL: National Commission for Computing and Liberties. www.cnil.fr
[7] CDT: Center for Democracy and Technology: www.cdt.org

- Business constraints that are relevant to the definition of the suitable provenance collection and management policy. It impacts also the granularity of provenance depending on the specificity of each cloud. It involves also the definition of accessibility rules and profiles to provenance.
- Technical constraints that are relevant to provenance collection. It involves also interoperability at the semantic and syntactic level between provenance files and records generated by heterogeneous systems, security issues including provenance integrity, access control, reliability and privacy. Performance and scalability constraints are also important because we are in a cloud context mandating from a provenance service to be pervasive and scalable.

4 Provenance in the cloud challenges

As illustrated by the case study, provenance becomes more and more important and challenging. The need for a provenance technology provided as an end-to-end service to guarantee information integrity and probative value is accentuated. Principally, the challenges can be divided into two categories:

- Known challenges but whose nature changes according to the cloud specificity. It involves object identification and coupling with provenance, provenance reliability and confidentiality. It contains also the level of genericity of such a service.
- New challenges which occur only in a cloud context and that are essentially relevant to performance, availability and scalability. Clouds are characterized by their ability to scale dynamically on demand. They can scale up and down according to the workload. So, how can we define a provenance service in the cloud and what are its characteristics? Would it be possible to use existent cloud computing frameworks like Hadoop [8]or more high level frameworks like Chukwa [9] or Pig [10] to answer to high scalability and fault tolerance characteristics?

4.1 Identification and mapping digital objects with their provenance

The first question is how to identify objects flowing between different clouds? The problem is that every cloud has its own policy to identify objects and that these policies are generally confidential. The challenge is that the technical integrity of a numeric object can change whereas its legal integrity remains the same. For these reason, traditional identification techniques based on hash computing are not suitable and cannot provide a unique object identifier. The second question is how to keep the link between a digital object and its provenance information along the different lifecycle phases and how to ensure provenance persistence? Persistence is the ability to provide an object provenance even if the

[8] Hadoop project: http://hadoop.apache.org
[9] Chukwa project: http://hadoop.apache.org/chukwa
[10] Pig project: http://hadoop.apache.org/pig

object was removed. This data independent persistence is mandatory because in many cases where the ancestor object is removed, the provenance graph will be disconnected and the descendents objects provenance will be undetermined. So how provenance-data mapping should be managed? Would it be possible to consider provenance as an intrinsic metadata of each object or it will be better to consider the document and its provenance separately and just keep a descriptor referencing the object and its provenance.

4.2 Trust and reliability

In the cloud, information crosses the traditional boundaries and passes through untrusted clouds. Generally, clouds are untrusted since the guarantee provided regarding the origin and the transformations are minimal, unclear or unreliable. So provenance information can be forged or tampered. Having a tempered or low quality provenance can be worse than having no provenance at all. To ensure trustworthy provenance, we need to provide the mandated security elements [15, 19]:

- Integrity: the assurance that provenance data is consistent and not tampered. So, any forgery or tampering in provenance records will be detected.
- Availability: an auditor can check the integrity and the correctness of provenance information.
- Confidentiality: provenance information should not be public, only authorized authorities can access to it. This means that systems must have the ability to restrict views of data and its provenance according to different confidentiality levels.
- Atomicity: at storage time, both provenance and data should be stored or neither should be stored.
- Consistence: at retrieval time, data returned should be consistent with provenance.

4.3 Heterogeneity and granularity

As illustrated by the use case, provenance is generated and managed by different entities according to different policies. It is often heterogeneous and not interoperable. Depending on the context, provenance needs and capabilities differ from one cloud to another and can't be exactly the same. The problem is that there is no portable and standard plausible "one format fits all" provenance that can be wired into general-purpose systems. Provenance should be tracked at different levels of granularity by the use of a flexible granularity approach. Such an approach should be based on a standardized provenance format that can be extended to a coarse-grained format or reduced to a fine-grained format. In this scope, a standard format for provenance called "Open Provenance Model" (OPM) was proposed to the scientific community [16]. This model standardizes provenance syntax and proposes a vocabulary for provenance, but there is still a lot of work in OPM model to be adapted as a provenance standard applicable for the cloud without any incompatibility or misinterpretation.

4.4 Cloud constraints: extensibility, availability and scalability

Cloud is designed to be scalable and available on demand. Existing provenance solutions in other environment (workflow, database...) dont consider the availability or scalability in their design. Also, cloud is not extensible and textcomp couldn't be modified or extended. In addition, the availability of the provenance service has to match the high ability of the cloud. If it's not the case, the overall availability is reduced to the limited availability of the provenance service. Regarding scalability, the use of a database can make provenance queriable but not scalable. This limitation is due to updates synchronization between clients. These updates can cause a distributed service lock introducing a distributed deadlock or a sclability bottleneck due to a single lock. In this case, the need is to communicate with a cloud database because the use of a parallel database is hard to maintain and is in contradiction with clouds model. Storing the provenance in a separate service introduces the issue of coordinating updates between object store (documents) service and the database service .

5 Provenance approaches and related works

5.1 Provenance approaches

To address the aforementioned challenges, we present in this section approaches for provenance management. We have identified three types of approaches for provenance collection and management:

– Centralized approach (cf. Fig 2): it consists on the definition of a centralized policy for collecting and managing provenance from end-to-end. In this case, a unique trusted authority provides a provenance service. This approach permits to alleviate data structure heterogeneity by using a unique standardized provenance format. However, we have to imagine that this authority can access external clouds through specific API. This authority should provide strong authentication techniques permitting to its clients to trust this centralized service and to communicate without revealing critical information.

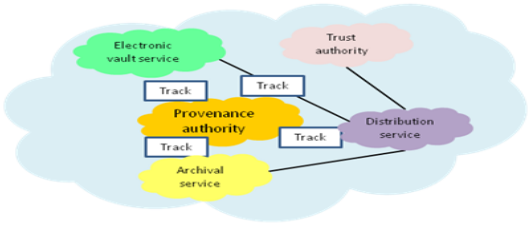

Fig. 2. Managing provenance within a centralized approach

– Hybrid approach (with import/export feature): this approach allows the import and the export of provenance between a centralized provenance service and external clouds (cf. Fig 3). The import feature allows retrieving the requested information without an additional management effort because it's supposed that the imported information is directly exploitable. In the same way, the centralized authority can directly export exploitable provenance to external services requesting it. Within this approach, provenance should define a management policy to alleviate syntaxic and semantic interoperability issues.

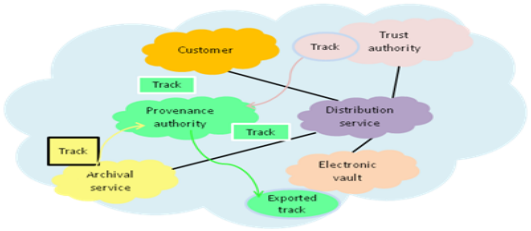

Fig. 3. Managing provenance within a hybrid approach

– Federated approach (cf. Fig 4): in this approach provenance management responsibilities are delegated to each participant cloud which should handle it autonomously. Provenance is distilled through multiple files that we need to cross to produce the provenance chain. Within this approach, limits are heterogeneity between provenance formats coming from different heterogeneous entities implying semantic and syntactic interoperability issues. To have a probative value, it is necessary that provenance elements can be folded and be consistent at both ends of a transaction. Since one of the parties dont meet these conditions, the whole provenance chain becomes suspicious. Also, the accuracy of information implies a need for corroborating provenance elements between two clouds. This limits evidence character to transactions where only two players are concerned.

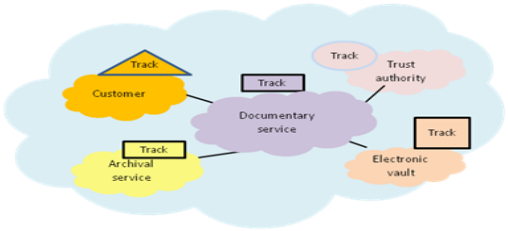

Fig. 4. Managing provenance within a hybrid approach

After the identification of the different constraints and the presentation of the approaches, we have achieved the following classification according to the management approach (cf. Table 1). This table illustrates that the challenge level about trust and reliability is high for the three approaches because provenance access control is always challenging since traditional access control techniques are not relevant for provenance DAG [3] and that provenance tempering is possible even within centralized approach. In the same way, managing cloud constraints has always the level (H) in the three approaches.

Constraint/ Approach	Identification Mapping	Trust / reliability	Heterogeneity / granularity	Cloud specific constraints
Centralized	L	H	L	H
Hybrid	H	H	M	H
Federated	H	H	H	H

Challenge level : - High : H
- Medium : M
- Low : L

Table 1. Classification of provenance constraints

5.2 Related works

Previous researches on provenance tackles principally the problem from one of two sides: database and files management or workflow and grid environments. Works on database and files provenance [2, 4, 21] focused on data- driven business processes and relies on the ability to trace information flows through SQL manipulations or by increasing file headers and tagging them [1, 20]. On the other side, workflow provenance research [8, 9] tackles scientific processing and deals with less transparent but predefined process executions. Many other works on provenance addresses specific problems like provenance in software development known as versionning systems which manages distributed intra-files provenance but are not able to manage inter-files provenance. We also mention provenance in collaborative environments for scientific communities (biologists, earth scientists) or securing provenance from tempering based on cryptographic mechanisms [12]. The particularity of the provenance needs in our work is the superposition of a set of traditional provenance constraints with cloud specific constraints. Most of the aformentioned works assumes the ability to alter the underlying system components, and dont propose a native provenance infrastructure addressing the problem as an end-to end property corresponding to all the aforementioned constraints.

6 Achieved work and perspectives

For the presented use case, we proposed a centralized provenance service model in an intra-cloud context(cf. Fig 5). Since we hope to define a generic service

that is able to track different types of digital objects, we have defined a sort of "business dictionaries". To create these dictionaries, we have proposed AXTL (Advanced XML Tracking Language) which is a markup language based on XML. The core module of AXTL specifies the form of requests, responses and tracks. It provides an abstraction of application domain and accepts proprietary extensions to allow each client to cover all the specifics of his job. Each dictionary describes the authorized actors, manipulated objects and the type of actions that can be performed. All of these dictionaries must be registered with a trusted authority that we are going to define as a provenance authority, guaranteeing the meaning of each referenced object . Hence, our global architecture is based on three sub-services:

- Track management: represents the service interface that is visible for clients. It contains the operations of track creation, augmentation, consultation and research. It covers also securing provenance during the collect and all lifecycle phases.
- Dictionary management: defines the different dictionaries, their structure and their interactions with the core module.
- Profile management: profiles allow the customization of the track management policy for every type of client. It contains the used dictionary, hash algorithm, signature algorithm, certification and timestamp authorities.

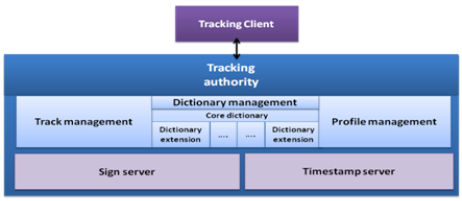

Fig. 5. Global overview of the tracking service architecture

We have implemented a first prototype of this architecture: in the server side, we have developed a java servlet as well as libraries defining documents tracks. For the client side, we have implemented a java and php clients having the same interface. This interface contains methods allowing to:

- Create a track for a document: this action creates a unique identifier for a document (called TID: TrackID) and associates it with different parameters that can be collected from the original document issuer according to the used tracking policy.
- Add a declaration to an existent track: a declaration is defined as an object that contains the action performed on the document, the actor of this action and its identifier as well as the identification of who underwent. It can be added only to an object already known for the provenance service.
- Find track: returns a list of tracks that corresponds to a specific search criteria (hash of the document, the identifier of the document for its owner)

– Get track: returns an XML track containing the whole document lifecycle.

This prototype was tested for the illustrated case study. However, it requires an integration effort in every cloud to communicate with the centralized provenance service. This should be improved to allow seamless integration with different SaaS providers. Regarding provenance integrity, we have developed integrity control mechanisms to check that the whole declarations for a document were not tempered but we need to develop a flexible access control approach regarding that data and provenance can have different confidentiality and privacy requirements. Provenance in the cloud is an open research problem and our future work will be focused on provenance deployment as a service in the cloud, provenance formal analysis, confidentiality and access control. To do that we have first to analyze cloud capabilities for provenance. Regarding security issues, we estimate to introduce semantics to enhance provenance confidentiality. Regarding privacy, we wish to introduce techniques assuring information usefulness in respect with privacy needs without any leakage. This requires the definition of new access-control techniques and data access schemas. In the future, we think that the challenge for provenance systems is to introduce remembrance which is a kind of an RFID for all types of digital objects. This concept was introduced in [13], it consists in augmenting all data objects with persistent memory and to consider that provenance is an intrinsic property of data objects. Remembrance can be very interesting in cloud computing scenarios and can immediately show that a piece of data is originated from an untrusted object or application. It exceeds traditional versioning and snapshots systems and tries to give a complete end-to-end history as an atomic property of any digital object.

7 Conclusion

Provenance is becoming more and more important as information is shared across networks and exceeds internal boundaries. Actually, the most used medium for information sharing, the web does not provide this feature. But the cloud, which is not yet a mature technology, provenance issues can be addressed from now.

In this paper we have presented a pragmatic case study illustrating the need for provenance to ensure electronic documents probative value. It allowed us to illustrate and analyze the challenges and to categorize the inherent constraints motivating our research on information provenance. We aim in our future researches to address provenance in the cloud and how to make it queryable, pervasive, handled automatically and trustworthy. These challenges, if tied together with performance and scalability issues could revolutionize cloud computing.

References

1. Agrawal, P., Benjelloun,O., Sarma,A D., Hayworth, C., Shubha U., Nabar, CU., Sugihara, T., Widom,J.: ULDBs: Databases with Uncertainty and Lineage. Trio: A System for Data, Uncertainty, and Lineage. VLDB 2006: 1151-1154.

2. Bhagwat, D., Chiticariu,L., Tan, W.C., Vijayvargiya, G.: An Annotation Management System for Relational Databases. VLDB, pp. 900-911, 2004.
3. Braun, U., Shinnar, A., Seltzer., M.: Securing provenance. Third USENIX Workshop on Hot Topics in Security (HotSec), July 2008.
4. Buneman, P., Khanna, S., Tan, W.C.:Why and Where: A Characterization of Data Provenance. ICDT, pp. 316-330, 2001.
5. Cameron, G.: Provenance and Pragmatics. Workshop on Data Provenance and Annotation, 2003.
6. Cheney, J., Chong, S., Foster, N., Seltzer, M., Vansummeren, S.: Provenance: A Future History. International Conference on Object Oriented Programming, Systems, Languages and Applications, OOPSLA 2009, pp 957-964.
7. Da Silva, PP., McGuinness, DL., McCool., R.: Knowledge Provenance Infrastructure. IEEE Data Engineering Bulletin, vol. 26, 2003, pp. 26-32.
8. Davidson, S., Cohen-Boulakia, S., Eyal, A., Ludascher, B., McPhillips, T.,Bowers, S., Freire., J.: Provenance in Scientific Workflow Systems. IEEE Data Engineering Bulletin, vol. 32, pp. 44-50, 2007.
9. Goble, C.: Position Statement: Musings on Provenance, Workow and Semantic Web Annotations for Bioinformatics. Workshop on Data Derivation and Provenance 2002.
10. Grand challenges in computing research conference 2008. UK Computing Society: www.ukcrc.org.uk/press/news/challenge08/gccr08final.cfm
11. Hasan, R., Yurcik. W, Myagmar, S.: The Evolution of Storage Service Providers: Techniques and Challenges to Outsourcing Storage. In Proceedings of the 2005 ACM workshop on Storage Security and Survivability.
12. Hassan, R., Sion, R., Winslett, M.: Preventing History Forgery with Secure Provenance. ACM Transactions on Storage (TOS) 2009.
13. Hassan, R., Sion, R., Winslett, M.:. Remembrance: The Unbearable Sentience of Being Digital. Fourth Biennial Conference on Innovative Data Systems Research (CIDR) 2009.
14. INFOSEC Research Council (IRC) Hard problem list. Technical report, november 2005. www.cyber.st.dhs.gov/docs/IRC_Hard_Problem_List.pdf.
15. ISO 14721:2003. Space data and information transfer systems - Open Archival Information System Reference model (OAIS): www.iso.org/iso/$catalogue_detail$.htm?csnumber=24683.
16. Miles, S., Groth, PT., Munroe, S., Jiang, S., Assandri, T., Moreau., L.: Extracting causal graphs from an open provenance data model. Concurrency and Computation: Practice and Experience, 20(5):577-586, 2008.
17. MoReq2 specifications. Model Requirements for the management of electronic records UPDATE AND EXTENSION, 2008: www.moreq2.eu.
18. Muniswamy-Reddy, KK., Holland, DA., Braun, U., Seltzer,M.: Provenance-aware storage systems. In USENIX Annual Technical Conference, General Track, pages 43-56, 2006.
19. NF Z42-013. Electronic archival storage-Specifications relative to the design and operation of information processing systems in view of ensuring the storage and integrity of the recording stored in these systems: http://www.boutique.afnor.org
20. Sar, C., Cao, P.: Lineage file system. Technical Report, Janurary 2005, http://crypto.stanford.edu/ cao/lineage.
21. Simmhan, Y., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Record (Special section on scientific c workflows), 34(3):31-36, 2005.