Data Usage Control for the Cloud

Florian Kelbert
kelbert@cs.tum.edu
Technische Universität München (TUM), Germany
Supervisor: Prof. Dr. Alexander Pretschner, TUM

***Abstract*—Despite the increasing adoption of cloud-based services, concerns regarding the proper future usage and storage of data given to such services remain: Once sensitive data has been released to a cloud service, users often do not know which other organizations or services get access and may store, use or redistribute their data. The research field of usage control tackles such problems by enforcing requirements on the usage of data after it has been given away and is thus particularly important in the cloud ecosystem. So far, research has mainly focused on enforcing such requirements within single systems. This PhD thesis investigates the distributed aspects of usage control, with the goal to *enforce usage control requirements on data that flows between systems, services and applications that may be distributed logically, physically and organizationally*. To this end, this thesis contributes by tackling four related subproblems: (1) tracking data flows across systems and propagating corresponding data usage policies, (2) taking distributed policy decisions, (3) investigating adaptivity of today's systems and services, and (4) providing appropriate guarantees. The conceptual results of this PhD thesis will be implemented and instantiated to cloud services, thus contributing to their trustworthiness and acceptance by providing security guarantees for the future usage of sensitive data. The results will be evaluated w.r.t. provided security guarantees, practicability, usability, and performance.**

I. Introduction

With the increasing maturity of cloud services, more and more individuals and businesses adopt these services for processing and storing their valuable data. Examples are collaborative creation of documents using SaaS-based text processing, sharing photos using IaaS-based storage, and deploying software on PaaS platforms.

Though, concerns keep many users from using such services: once sensitive data has been released to the cloud, there do not exist reliable means to exercise control over its further usage [1]. To users it remains uncertain whether, how and by whom their data is or may be accessed, stored and used. Worse, backup and replication mechanisms create additional copies that may be located on servers in different geographical locations or on servers under control of different organizational units. Consider a PaaS provider building upon multiple IaaS providers: in such a case the users' data is transparently replicated to different organizations and geographical locations. This leads to the problem that data owners are no longer in full control over their data, as they (intentionally) do not know at which locations and by which organizations their data is stored and processed. Often they also do not know which other entities (e.g., cloud services, their employees, other end users) have accessed, used, stored, or (re-)distributed their data.

The research field of distributed data usage control [2], [3] tackles such problems by providing mechanisms to control the usage of data once access to it *has been* granted. Example policies are "only process my data with application X," "do not redistribute my data to company Y," and "delete my data after thirty days." Usage control is thus particularly important in cloud environments, and, more generally, in distributed system environments. As many users are concerned about processing their sensitive data in the cloud, usage control may act as an enabler for cloud-based processing of sensitive data: Data owners would be able to specify data usage policies and a corresponding infrastructure would guarantee that they are enforced at all data storing and processing sites.

While usage control within single systems has been and is being researched [4]–[6], this thesis investigates the distributed aspects of data usage control. The general research question thus is *"How can usage control requirements be enforced if data flows between systems, services and applications that may be distributed logically, physically and organizationally?"*

As cloud services are one particular instance of such system environments, I plan to instantiate my more general work to real-world cloud computing environments (cf. §III). This thesis is organized in four complementing parts, aiming at comprehensive enforcement of usage control requirements in distributed system environments: (1) generic and fine-grained cross-system data flow tracking and policy propagation in order to know where copies of data reside and to enforce corresponding policies; (2) enforcement of policies for which decisions must be taken in a decentralized manner; (3) addressing the challenge of system and service adaptivity without compromising policy enforcement; (4) providing appropriate guarantees w.r.t. proper usage control enforcement in adaptive distributed system environments.

§II describes related work, its room for improvement and the expected contribution of this thesis. After introducing a use case and the attacker model in §III, §IV motivates more specific research questions, explains the proposed approach, its implementation and evaluation. §V sketches preliminary results and remaining objectives; §VI concludes.

II. Related Work & Expected Contribution

Previous usage control solutions addressing data distribution and sticky policies [7], [8] do not cope with the complexity of today's distributed systems, as they allow for uni-directional data distribution only. Also, these solutions are specific to particular application(-layer protocol)s and thus lack generality.

Martinelli et al. [9] realize usage control for grid computational services by making the grid user deploy her application together with a policy. Different from the approach in this thesis, policies are defined for applications rather than data. As data flows are not considered, cooperating applications may circumvent the usage control mechanisms.

DataSafe [10] aims at preventing "illegitimate secondary dissemination of protected plaintext data by authorized recipients." It allows unmodified applications to transparently use protected data while ensuring compliance with policies. DataSafe builds upon additional hardware and a trusted hypervisor, which is not needed in my approach. Instead, I plan to use Trusted Platform Module (TPM) technology, which is, in contrast to the hardware used in DataSafe, already being deployed on a large scale. Different to DataSafe, this thesis will also allow for the enforcement of policy obligations such as deletion of data after a certain amount of time.

Digital Rights Management (DRM) [11] refers to mechanisms that aim at controlling the usage of copyrighted, read-only, digital information at the data consumer's site. It can therefore be considered a specialization of usage control [3] that focuses on payment-based dissemination [2]. In contrast to this thesis, DRM does not provide means for end users to protect their valuable information. While DRM solutions are limited to read-only content, tailored to specific file types and rely on specific applications to enforce digital licenses, the results of this thesis will not be limited in such a way.

As usage control policies must be enforced at the data processing sites, parts of the infrastructure must be deployed remotely. To provide certain guarantees, remotely deployed components must "behave in a 'good' manner [verifiable] by the policy stakeholder" [12]. Several TPM-based solutions have been proposed [12]–[14]; their basic idea is to measure crucial system components (BIOS, Bootloader, operating system, usage control infrastructure) and verify their integrity using a set of known "good" values. Yet, not much work has been carried out with the goal to use such technologies for securing data-driven usage control infrastructures. I will thus leverage TPM or similar technologies in order to ensure the integrity of remote systems.

Gap Analysis. Many questions regarding generic enforcement of data-driven usage policies within distributed system environments remain unsolved. Existing solutions are limited to particular applications, protocols, or file types; some solutions support uni-directional data flows only or are limited to read-only content; others introduce new hardware components. In sum, comprehensive solutions that generically tackle the problem of usage control enforcement in distributed systems have not yet been investigated, proposed and developed.

The **expected contribution** of this thesis is a conceptual framework, its implementation and its evaluation, allowing for trustworthy, generic, application-independent enforcement of usage policies in adaptive distributed system environments. In particular, the framework will take into account the complexity of today's distributed systems, where data may be distributed by any data possessor and data processing and storing systems keep changing. I expect these results to contribute to data security and privacy within any kind of cloud environment and, thus, to the trustworthiness and acceptance of cloud services.

Scope. This thesis focuses on the distributed aspects of usage control. Policy specification and evolution as well as enforcement within single systems are thus out of the scope. This thesis will leverage respective solutions [6], [15].

III. Use Case Scenario & Attacker Model

As use case, consider an insurance company serving both private and corporate clients. In order to offer attractive contracts, storage, processing and analysis of huge amounts of data obtained from both public sources and customers are at the core of its business. Thus, the insurance set up an in-house Big Data computing cluster which stores and processes large amounts of data in a distributed, potentially virtualized, manner. For business support functions such as email, collaboration and web presence, the insurance demands their employees to use cloud-based services provided by external partners.

As the data of both private and corporate customers is highly sensitive and valuable, the insurance decided to deploy data usage control mechanisms on all in-house computing devices such as the employees' machines and its computing cluster. By using such a solution, the insurance aims at eliminating data misuse and illegitimate data disclosure. As the Big Data software solution being used is quite complex, such incidents may not only happen deliberately or accidentally by employees, but also because of malicious, misconfigured, or corrupted software. Since data could also be leaked or misused by the external cloud-based services, the insurance also demands the corresponding service providers to deploy such mechanisms. §IV motivates research questions on the basis of this use case and outlines the proposed approach.

Attacker Model. End users, cloud services and data processing software are considered threats w.r.t. sensitive data. While data misuse and leakage may happen deliberately, misconfigured or corrupted software may lead to similar effects.

IV. Research Questions and Proposed Approach

A. Cross-System Data Flow Tracking & Policy Propagation

Consider a private client asking the insurance company for health insurance offers using their SaaS-based web forms. In order to get a personalized offer, the client provides her name, address, sex, age, and an overview of her history of diseases. After submission, the information is emailed to an employee who analyzes the data using the in-house computing cluster. She will then compose an offer using the cloud-based text processing service and email it to the client. If the client does not accept the offer within the next weeks, regulations demand deletion of the transmitted health information.

To enforce data deletion across all involved systems (web and email service, computing cluster, and text processing service), the data flow across these systems must have been tracked and policies must have been propagated accordingly.

I plan to answer the corresponding research question *"How can the flow of data between different connected systems be*

tracked and how can data usage policies be propagated to the corresponding decision points?" as follows:

(1) Provide a generic model for cross-system data flow tracking by leveraging an existing model for tracking intra-system data flows [4]. I plan to achieve fine granularity by incorporating application(-layer protocol) specific knowledge into the model and by investigating declassification techniques.

(2) For performance reasons, policy decisions are likely to be taken in a decentralized manner by, or close to, the data processing systems. Policies must thus be propagated whenever the corresponding data flows in-between systems. To this end, I will develop concepts and mechanisms to implement sticky policies for distributed usage controlled systems.

Technically, this will be achieved by monitoring relevant system calls and network messages. In sum, this will allow for knowing in which systems sensitive data resides and for enforcing corresponding policies within these systems.

B. Distributed Policy Decisions

Consider a corporate client requesting an insurance offer via email using a generic email address of the insurance; the email will then be forwarded to responsible employees. Because of a strict Separation of Duty policy, the client does not want the request to be processed by any employees that processed data of a particular competitor in the past. As the email service is not aware of previous request processing and because employees may have accessed the competitor's data using different services or systems, these different systems and services must cooperate in order to take corresponding policy decisions. Thus, the need for taking distributed policy decisions arises. Corresponding mechanisms will be able to take policy decisions if information about data distribution and data usage are distributed across systems.

I plan to answer the corresponding research question *"How can usage control decisions be taken if data and policies are distributed across different systems?"* by providing a model, a system architecture and protocols for a distributed decision process. For this, I will investigate to what extent solutions developed in related research fields (e.g., access control and DRM) can be applied to usage control. After assessing the pros and cons of existing approaches w.r.t. to different usage control requirements and data distribution scenarios, I will provide a corresponding framework for usage control.

Technically, I will investigate synchronization protocols for taking distributed policy decisions in case the sensitive data, as well as information about its distribution and usage are distributed across systems.

C. Adaptivity and Distribution Transparency

A particular property of today's distributed systems is their adaptivity and distribution transparency: Services make use of other services which are in turn implemented in a distributed manner—often transparently w.r.t. location, underlying infrastructures, and distribution. Such adaptivity may stem from technical reasons, e.g. integration of new services or mechanisms such as DHCP, and from organizational reasons, e.g. hiring or reorganization. Thus, future data processing systems, services and users may be unknown at the time of data distribution. Usage control frameworks must be able to cope with such changes without compromising policy enforcement.

Consider the insurance's computing cluster to be built atop of an IaaS cloud for scalability reasons: once a year, the insurance leverages its computing cluster to recalculate insurance premiums for large parts of its contracts. For this, the insurance temporary obtains additional computing power by an external IaaS provider and integrates the machines into its computing cluster. Still, the insurance does not want any of their data to be leaked through these additional machines.

I plan to answer the corresponding research question *"How can usage control requirements be enforced if data processing systems and users are not known upfront and keep changing?"* by providing a model and an architecture for transparent, non-invasive usage control enforcement in adaptive distributed systems. This means that adaptive components (e.g. users, systems, services) must not necessarily be aware of any usage control solutions, thus supporting any kind of legacy software.

Technically, the proposed solution will likely be implemented as a transparent middleware component, thus allowing any new component to integrate into the usage controlled environment without providing additional mechanisms. In sum, the results will allow for the enforcement of usage control policies even if the data processing components keep changing.

D. Providing Appropriate Guarantees

By making use of cloud services for web and email, and by integrating remote IaaS resources into its computing cluster, the insurance deliberately releases valuable data to remote service providers. While contractual agreements may demand these providers to comply with certain data usage and storage restrictions or to deploy usage control mechanisms, compliance with such agreements is hard to verify or enforce. Even if service providers are considered trustworthy, their systems or services may get compromised and thus become a threat to the insurance's data. In particular the fact that formerly unknown systems and services may be entrusted sensitive data yields concerns w.r.t. proper policy enforcement. Thus, it is necessary to provide appropriate guarantees that usage control requirements are enforced at all data processing sites.

I plan to answer the corresponding research question *"Which guarantees for usage control enforcement in adaptive distributed systems can be provided under which preconditions?"* by investigating existing mechanisms for providing guarantees in distributed systems in terms of usefulness and practicability for usage control. I will then integrate appropriate mechanisms into the usage control framework, aiming at (1) mitigation of the downsides of existing mechanisms and (2) providing sufficient guarantees w.r.t. proper usage control enforcement under specified preconditions. If necessary and appropriate, I will extend existing mechanisms.

Technically, technologies such as TPM and smart cards will be integrated into the framework to verify the trustworthiness of remote systems before sensitive data is distributed.

E. Implementation & Evaluation

As a proof of concept I will **implement** a framework that encompasses the proposed models, architectures, and protocols. In order to show its practicability and usefulness in real-world systems, I will instantiate the implementation to the described use case, in particular to cloud-based service instances and an adaptive, IaaS-based computing cluster.

I will **evaluate** both the conceptual work and its implementation w.r.t. provided security guarantees, practicability, usability, and performance. In terms of security guarantees, I plan to investigate under which preconditions (e.g., data usage requirements, data distribution) which guarantees can be given and if/how enforcement mechanisms can be circumvented. This security analysis will also take into account different attacker models, i.e. different malicious entities, such as end users, software developers, service and infrastructure providers as well as administrators (cf. §III). In terms of practicability, I plan to evaluate which technical and organizational changes are required to existing systems and organizations such that security guarantees are in fact provided. I plan to evaluate performance of the prototype; meaningful conclusions will also take into account concrete use cases, potentially considering user interaction. Finally, I plan to evaluate to what extent the framework has an impact on the overall usability of the system by taking into account the precision of cross-system data flow tracking, effectiveness of enforcement mechanisms, and overall performance.

V. Preliminary Results & Remaining Objectives

Preliminary results have been published in [16], [17]: (1) a generic model for cross-system data flow tracking suitable for all Internet-based communication; (2) an instantiation of the model to TCP/IP, therefore allowing for application(-protocol) independent cross-system data flow tracking for many applications and protocols; (3) an architecture implementing the cross-system data flow model for TCP and realizing policy propagation upon data transfer; (4) an evaluation of the work w.r.t. security and performance. [18] shows an instantiation of the above work to a smart meter connected to a web-based social network. In sum, these preliminary results cover large parts of the research question described in IV-A.

Remaining Objectives & Future Plans. These initial results are realized without leveraging application-specific knowledge, thus leading to many false negatives. Currently, I work on enhancing these models and mechanisms to incorporate further knowledge in order to minimize these overapproximations; this work is supposed to be finished in June 2013. The work described in §IV-B and §IV-C will be tackled afterwards and is planned to be finished in November 2013 and March 2014, respectively. I expect that providing appropriate guarantees for remote systems (§IV-D) introduces major challenges, as this requires all participants to deploy Trusted Computing technologies. While this may not be feasible in completely open environments, I aim at giving reasonable guarantees within semi-closed environments as described in §III. I plan to finish the PhD thesis no later than end 2014.

VI. Conclusions

To date, users remain concerned about storing and processing their valuable data in the cloud, as there do not exist reliable means to exercise any further control once data has been released. This PhD thesis contributes to the trustworthiness and acceptance of cloud services by providing means for reliably enforcing requirements on the future usage of data, even if data processing systems and services keep changing and are distributed logically, physically, or organizationally. While the underlying problems, proposed approaches and expected results are more general, they will be applied to cloud services and evaluated using real-world use case scenarios.

References

[1] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling Data in the Cloud: Outsourcing Computation without Outsourcing Control," in *Proc. ACM Workshop on Cloud Computing Security*, 2009.

[2] J. Park and R. Sandhu, "Towards Usage Control Models: Beyond Traditional Access Control," in *Proc. 7th ACM Symp. on Access Control Models and Technologies*, 2002.

[3] A. Pretschner, M. Hilty, F. Schütz, C. Schaefer, and T. Walter, "Usage Control Enforcement: Present and Future," *IEEE Security & Privacy*, vol. 6, no. 4, 2008.

[4] M. Harvan and A. Pretschner, "State-Based Usage Control Enforcement with Data Flow Tracking using System Call Interposition," in *Proc. 3rd Intl. Conf. on Network and System Security*, 2009.

[5] T. Wüchner and A. Pretschner, "Data Loss Prevention Based on Data-Driven Usage Control," in *Proc. 23rd IEEE Intl. Symp. on Software Reliability Engineering*, 2012.

[6] A. Pretschner, E. Lovat, and M. Büchler, "Representation-Independent Data Usage Control," in *Proc. Conf. on Data Privacy Managemnt*, 2011.

[7] B. Katt, X. Zhang, R. Breu, M. Hafner, and J.-P. Seifert, "A General Obligation Model and Continuity-Enhanced Policy Enforcement Engine for Usage Control," in *Proc. 13th ACM Symp. on Access Control Models and Technologies*, 2008.

[8] P. Kumari, A. Pretschner, J. Peschla, and J.-M. Kuhn, "Distributed Data Usage Control for Web Applications: A Social Network Implementation," in *Proc. 1st ACM Conf. on Data and Application Security and Privacy*, 2011.

[9] F. Martinelli and P. Mori, "On Usage Control for GRID Systems," *Future Generation Computer Systems*, vol. 26, no. 7, 2010.

[10] Y.-Y. Chen, P. A. Jamkhedkar, and R. B. Lee, "A Software-Hardware Architecture for Self-Protecting Data," in *Proc. Conf. on Computer and Communications Security*, 2012.

[11] Q. Liu, R. Safavi-Naini, and N. P. Sheppard, "Digital Rights Management for Content Distribution," in *Proc. Australasian Information Security Workshop Conference*, vol. 21, 2003.

[12] X. Zhang, J.-P. Seifert, and R. Sandhu, "Security Enforcement Model for Distributed Usage Control," *Intl. Conf. on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2008.

[13] R. Neisse, D. Holling, and A. Pretschner, "Implementing Trust in Cloud Infrastructures," *Proc. 11th IEEE/ACM Intl. Conf. on Cluster Cloud and Grid Computing*, 2011.

[14] R. Sandhu and X. Zhang, "Peer-to-peer Access Control Architecture Using Trusted Computing Technology," in *Proc. 10th Symp. on Access Control Models and Technologies*, 2005.

[15] P. Kumari and A. Pretschner, "Model-Based Usage Control Policy Derivation," in *Proc. 5th Intl. Symp. on Engineering Secure Software and Systems*, 2013.

[16] F. Kelbert and A. Pretschner, "Towards a Policy Enforcement Infrastructure for Distributed Usage Control," in *Proc. 17th Symp. on Access Control Models and Technologies*, 2012.

[17] ——, "Data Usage Control Enforcement in Distributed Systems," in *Proc. ACM Conf. on Data and Application Security and Privacy*, 2013.

[18] P. Kumari, F. Kelbert, and A. Pretschner, "Data Protection in Heterogeneous Distributed Systems: A Smart Meter Example," in *Dependable Software for Critical Infrastructures*, 2011.