Edited version published as "*Data Flow Management and Compliance in Cloud Computing*" in Cloud Computing, Special Issue on Legal Clouds. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7270240

Seeing through the clouds: Managing data flow and compliance in cloud computing

Jatinder Singh[1], Julia Powles[2], Thomas Pasquier[1] and Jean Bacon[1]

[1]Computer Laboratory, University of Cambridge
[2]Faculty of Law, University of Cambridge

Abstract

As cloud computing becomes an increasingly dominant means of providing computing resources worldwide, legal and regulatory issues associated with the cloud also become more pronounced. In particular, there is a heightened focus on ensuring the privacy and integrity of end-users' personal data. At present, the cloud is opaque, a black-box. The technical means for enforcing and demonstrating compliance with data management practices lag behind legal and regulatory aspirations.

After reviewing existing methods for containing, accessing and encrypting data, we introduce Information Flow Control (IFC) as a technology enabling auditable, fine-grained management as data moves throughout systems. We describe how IFC offers potential in improving the visibility and control over data flows within and between cloud services and cloud-hosted applications. This is demonstrated through real-world legal/regulatory examples, which show how IFC can help satisfy data management obligations, and improve the accountability of responsible parties.

I. Responsibility in the cloud

Cloud computing is an industry with rapid and continued growth, reflecting the efficiencies and cost reductions that can be obtained through economies of scale, improved global accessibility, and simplified, 'outsourced' management and configuration.

Legal issues concerning data in the cloud derive primarily from four areas: contract; data protection; law enforcement; and regulatory and common law protections for particularly sensitive domains such as health, finance, fiduciary relations, and intellectual property assets.

From a technical perspective, these legal requirements all impose information management obligations on the transmission and sharing of data within cloud-hosted applications and services. They may restrict how, when, where, and by whom, data may flow and be accessed. These issues must be managed not only between applications, but through the entire, potentially global, cloud supply chain.

Currently, cloud providers employ access controls, to prevent unauthorised access to data and services; and containment mechanisms, to prevent data leaking between tenants (those consuming cloud services) using shared infrastructure. But these tend to be security rather than compliance focused, often applying at specific application, system or user boundaries. Further, cloud services tend to be opaque and black-box in

Contact author: jatinder.singh@cl.cam.ac.uk

nature. Despite some management tools (that depend on the service model/application), there is typically little scope for tenants to visualise, let alone specify, how data should be managed once within the cloud, nor the precise circumstances in which data can be transferred.

Tenants and providers must ensure and demonstrate that they meet their legal and regulatory obligations. However, current technical mechanisms offer limited means for controlling data from afar, and insufficient tools for determining compliance and/or apportioning responsibility. This means that providers—and potentially their whole supply chain—must be *trusted* to act appropriately. This not only hinders accountability, but represents a barrier to cloud adoption, particularly for personal data use and for industries such as healthcare and finance where additional regulatory requirements pertain.

It is clear that more is required. We argue that one way forward is for flexible data-centric technical mechanisms that enable the visibility and *control of data flows inside, within and between cloud services.* As an exemplar, we introduce our ongoing work on *Information Flow Control* (IFC), to explore in this article how greater technical controls over data flows can allow parties to better manage their legal obligations, improve accountability, and offer verifiable data trails for audit and compliance.

Note that our focus here is on management and compliance with respect to civil, administrative and criminal law obligations and responsibilities. Surreptitious actions, such as those by government agencies or malicious parties, are beyond the scope of this particular discussion, and must be addressed with robust international policy efforts and domestic legal reforms.

A. Parties and relationships

Commercial cloud services are offered by *providers*, which may use third-party (*sub-provider*) services as part of their supply chain. A cloud service is used by *tenants* who contract with the provider in order to leverage the service. Tenants use cloud services to host applications and services that are, in turn, used by *end-users.*

Cloud services therefore involve a series of direct (typically contractual) relationships between: a) end-users and tenants; b) tenants and cloud providers; and c) cloud providers and sub-providers. Data flows tend to correspond to these relationships.

B. Responsibility and compliance

Legal and regulatory considerations for data flows in the cloud revolve around four primary dimensions:

1) **Contractual obligations.** Cloud services involve a number of contracts such as service level agreements (SLAs) and privacy policies. These documents impose obligations for which it would be valuable to audit data flows and therefore a) verify compliance; b) detect breaches; and c) apportion responsibility. Although many cloud contracts appear non-negotiable, in practice there may be some room for negotiation—particularly for larger organisations [1]. A benefit of mechanisms that enable managed, auditable data flows is an increased likelihood of negotiable terms between different parties, in addition to giving tenants and end-users a greater capacity for control.
2) **Data protection.** Data protection laws, adopted in many countries, place obligations and responsibilities on tenants and providers for the management of personal data. The fundamental premise of data protection is that all uses of information identifiable to an individual should be strictly regulated and controlled, with various audit mechanisms, flow and purpose restrictions, and penalties (in theory) for non-compliance. Building on the foundation of the OECD's 1981 Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, the most influential set of laws are those concerning the European Union's Directive 95/46/EC and the new EU Regulation currently under negotiation. In some jurisdictions, such as the United States, data protection regulates only some industries and types of processing, rather than providing a general schema for all data processing. In this article, we draw particularly on the EU data protection authorities' 2012 guidelines for cloud operators [2], which set out rigorous requirements for technical and organisational measures that ensure transparency, purpose

specification and limitation, and erasure of data. Note, however, that the technical concepts presented are general, and therefore can operate in other data protection regimes, be they the US or elsewhere.

3) **Law enforcement** access for crime/national security. For global businesses with international clientele, there is increasing pressure to report on government demands for data. In this article, we draw particularly on the recent high-profile Microsoft-Ireland case, and on the potential for IFC to track unauthorised (direct requests to cloud providers) and authorised (via warrant and mutual legal assistance treaties, possibly only to an accredited public institution) data flows out of a specified territory or agreed location.
4) **Regulatory and common law protections** for particularly sensitive domains such as health, finance, doctor-patient and lawyer-client relations, as well as, in a commercial context, protection of trade secrets and other intellectual property.

Within each of these areas, the physical (geo)location of data, storage, processing and equipment are particularly pertinent considerations [3], [4], as these are all relevant factors in determining legal jurisdiction to legislate, adjudicate and enforce obligations on otherwise delocalised cloud operations.

Different types/instances of data can come with different obligations and responsibilities, meaning that tenants may have a number of data management constraints, and therefore require flexible technical controls.

The critical dimension that seems to be missing from existing legal and regulatory obligations, and the way that providers have responded to date, are guarantees that responsibilities are being met: *demonstrable compliance through technical means.* This article furthers the proposition that proper steps/policies should be in place to show explicitly how, where, when and by whom data is accessed. This not only provides assurances to cloud consumers and to authorities, but offers a verifiable audit trail so that there is evidence available if something goes wrong, or responsibility needs to be apportioned. This can consolidate and be reinforced by external, general audits, to move beyond higher-level reactive checks and into proactive, data-centric and context-aware compliance.

II. Data management mechanisms: the status quo

These legal and regulatory issues concern data. It follows that the technical mechanisms for data management directly affect the ability of a party to meet, demonstrably, their obligations. Currently, the well-established and commercially deployed control mechanisms tend to focus more on security; compliance and accountability, despite being crucial cloud considerations [5], are given considerably less attention.

Cloud service provisioning

The type of cloud service offering determines the capacity for management. Cloud offerings tend to be described in terms of a *service model*, which reflects the parts of the cloud stack that are managed by the provider (Fig. 1). That is, the flavour of the service model relates to the degree of control a tenant has over the service. Generally, tenants have limited (if any) means to influence or view those aspects managed by the provider.

For *Infrastructure as a Service* (IaaS), the cloud provider manages the low-level aspects (hardware, software-hardware interfaces and isolation mechanisms), giving tenants the freedom to determine the operating system (OS), software stack, and of course, deploy and manage their own applications. *Software as a Service* (SaaS) is at the opposite end of the spectrum, where the entire application is offered and managed by the provider. This might be a university email service run by a large webmail provider, for example. SaaS tenants have far less freedom, as any management is determined by the configuration functionality offered by the application. *Platform as a Service* (PaaS) offerings run tenant-provided applications on top of the cloud provider's software and service stack. That is, the provider offers the runtime environment, and the tenant the application code.

For all service models, the tenant has a lack of control and visibility, to some degree, over various aspects of service. (Note IaaS tenants may use pre-configured virtual machine images, and often leverage other provider-managed services such as storage.) For end-users, the situation is similar to the SaaS provider-tenant scenario, in that any controls to manage their data are determined by the functionality of the (tenant's) application.

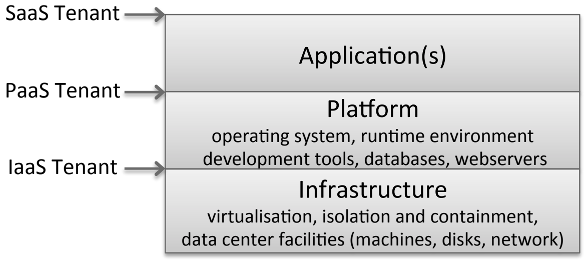
SaaS Tenant
PaaS Tenant
IaaS Tenant
Application(s)
Platform
operating system, runtime environment
development tools, databases, webservers
Infrastructure
virtualisation, isolation and containment,
data center facilities (machines, disks, network)

Figure 1: Cloud service provision models. The cloud provider manages all that is below the arrow, the tenant manages that which is above.

This can be analogised to technical control over one's Facebook user data, or Amazon customer data, which are determined by the privacy settings' functionality of those services. Cloud services may be composed. For instance, Dropbox (a SaaS offering for end-users) runs over Amazon (an IaaS offering), and SalesForce Heroku PaaS runs over Amazon IaaS. Thus it may not be clear to end-users, or even tenants, which offerings comprise the service's supply chain and therefore where responsibility lies.

A. Containment

Given the shared nature of the cloud, a key focus has been on isolating tenants (data and processing) in order to prevent information leakage. There are few mechanisms for tenants and providers to determine if/when data has been leaked, e.g. as a result of some misconfiguration, software bug, or as a security issue.

A common approach involves containing tenants by allocating them their own *virtual machines* (VMs), meaning that only the provider-managed hardware and hypervisor are shared between tenants. (Though there may be other shared infrastructure such as storage services.) Each VM maintains its own OS and the software layers above. This type of strong isolation may be appropriate for unrelated tenants, especially those using an IaaS offering. More recently, *containers* have enabled strong isolation of tenants over a shared OS. The goal of isolation is to segregate tenants, protecting their data and computation, and to limit a tenant's (direct) knowledge of others.

Though strong isolation is of clear importance, many applications and services will require data sharing across and outside of isolation boundaries. This may be in order to span a range of applications or to directly access other service components and resources, e.g. storage or billing services. Such interactions are managed through access controls.

B. Access controls

Access controls regulate the actions that a *principal*—human user, application, software, process, and so on—may perform, e.g. to read or write data, reconfigure a system, use a particular service, etc. These actions typically relate to data.

Applications/services use access controls to manage data they hold, through authentication (identification) of the principal ("you are who you say you are") and authorising the actions the principal attempts to take.

Authorisation involves applying a policy at a particular *policy enforcement point* within the application/service. This determines whether the action is allowed. As a simple illustration, one may log-in to a social media platform (authentication), where authorisation rules ensure that one may only view a profile's detail if they are 'friends'. This would be evaluated on an attempt to view a profile.

Access controls tend to operate within the scope of the particular application/service. Policy is enforced as the action is attempted, considering the principal(s) directly involved.

In a cloud context, this means that access controls generally govern the end-user–tenant, tenant–cloud provider, and provider–sub-provider interactions at the interface between them. These mechanisms typically do not, of themselves, offer control beyond that point; e.g. they would not regulate the indirect interactions between an end-user–sub-provider. Further, the application-centric nature of many access controls renders it difficult to have a consistent management policy that can apply across the range of different applications and services.

C. Encryption

The main roles encryption plays within the cloud are: a) to protect a communication channel, preventing eavesdropping during transmission; and b) to protect data items when transferred outside their boundary of control. Regarding the former, *Transport Layer Security* (TLS/SSL) transmission is commonplace, particularly in the post-Snowden era where large providers use encrypted communication channels, even within their data centres [3].

As a tool of data management, encryption provides an orthogonal form of protection to the access controls just described. Encryption does not restrict physical access to data, but rather, affects its usability by making it unintelligible. Access to the (intelligble) data is regulated through the distribution of keys that enable decryption. In a cloud context, this means that if a user places encrypted data in the cloud, this data will not be accessible by the provider, or anyone else, unless they hold the requisite keys.

This brings a number of considerations (see [4] for discussion). First, key management is hard: keys must be distributed to the relevant parties, and revoked (and reallocated) when conditions change. This comes with overhead, and quickly becomes unmanageable in dynamic and distributed environments such as the cloud. Second, being mathematically-based, there are no means for determining when, where and by whom data was decrypted. This makes detecting leaks difficult and hinders accountability. We argue that even the distribution of encrypted data should still be carefully managed, as a broken encryption scheme or compromised key at any time in the future places the data at risk.

The practice of encrypting data before upload may be appropriate for storage services, and more generally to protect against surreptitious access. However, many cloud service offerings entail data processing. This generally requires the provider to have access to the customer's intelligible data (and/or their keys if data is encrypted) to provide the service. There is ongoing work on homomorphic encryption, which allows operations to be performed on encrypted data without revealing the *plaintext* [6], but the current state of the art is not yet practical for use at scale.

D. The need for more

These controls clearly have their place: containment prevents data leaking between tenants on shared infrastructure, and access controls regulate the circumstances in which particular applications/services (data) may be accessed. However, these mechanisms were not designed to demonstrate compliance with respect to contractual, legal and regulatory obligations, nor to account for the fact that applications and the provision of cloud services necessarily entail data sharing.

To improve accountability, there is a clear role for technical mechanisms that:

- Enable data to be managed beyond application and system boundaries, i.e. within and between applications, cloud services, and throughout the cloud supply chain. This is particularly relevant as cloud becomes part of wider architectures, such as for the Internet of Things.
- Facilitate visibility to determine when/where data flows, to help identify the occurrence of any leakage and/or other data obligation failures. This provides evidence indicating who may (or may not) be responsible.
- Flexibly deal with the subtleties and nuances of data management requirements, which may be contextual, or apply only to certain data items, etc.

III. Managing the flow

Current cloud mechanisms depend heavily on *trust*—trust between parties that data will be properly managed once it is transferred beyond their direct control. To improve on this, we are developing mechanisms for *managing information flows within and between cloud services and applications.* The aim is to **complement** other control and management mechanisms, by providing visibility and control across isolation, application and system boundaries, and providing evidence for compliance purposes.

Controlling Information Flow

Conceptually, controlling data flows involves coupling data to a known management policy that can be enforced wherever the data flows. This is implemented by tagging data, where lightweight tags—representing particular management aspects/concerns—are tightly linked with data. The policy is enforced when tagged data then flows between components of the system, within or between machines. Enforcement of policy may involve allowing or preventing information flows, or transforming (and re-tagging) data so that it will be allowed to flow. Importantly, this approach means that data management policy need not be encoded within the software-logic (i.e. application-centric), but rather enables separation of policy that can therefore apply across different infrastructures and be managed over time.

A. Information Flow Control (IFC)

Information Flow Control (IFC) is a data flow control model that enforces policy against every flow in the system. To achieve IFC, tags are linked with data to represent various properties and policies concerning the data. Entities that can touch data are similarly tagged in order to control the flows between them. The tags are collected into two labels: a) a secrecy label representing the data's privacy/confidentiality/sensitivity; and b) an integrity label representing the data quality/provenance/authority.

The state of the labels represents a *security context*, and a flow is only permitted if the security context of the data and entity accord. This is through tag-matching rules (evaluated by subset relationships) [7] that ensure that outgoing flows that would violate privacy/confidentiality are prevented (secrecy) and that incoming data has certain (quality/sanitisation etc.) properties (integrity). Or in simple terms, secrecy constraints concern the decision to release data, integrity constraints the decision to receive. For example, data tagged as private-bob in its secrecy label can only flow to a processing component that contains a secrecy tag private-bob in its secrecy label, e.g. ensuring that data produced by Bob's mobile phone is only processed by a cloud application running on his behalf. Similarly, a storage component tagged as validated-data in its integrity label can only input data that has gone through some system validation or sanitisation process and has been assigned a tag validated-data in its integrity label.

Importantly, IFC provides the flexibility for fine-grained, data-centric control, and is inherently dynamic, meaning that flows can be managed in accordance with changes in context. *Decentralised Information Flow Control* (DIFC) [8] allows IFC tags to be managed in a decentralised fashion meaning that policy can be expressed and tailored to infrastructure, application or even individual, end-user specific needs.

Further, policy is enforced against every flow in the system, which is made possible by the fact that each entity runs in a security context, defined by its labels. This facilitates a more complete audit, as the policy decision for every flow can be logged, as well as the security contexts for the entities involved.

B. IFC in the cloud

There has been a good deal of research into IFC, as well as some practical realisations. Only recently has IFC been considered for cloud computing [9]. We have developed a full DIFC prototype for PaaS offerings, that provides a lightweight IFC policy enforcement regime to protect every OS-mediated I/O operation (i.e. within a virtual machine), with an integrated middleware responsible for policy-compliant inter-machine data exchange. Full technical details of this work are provided in [7].

IFC can help ensure that data is being properly protected by acting as a safety net around the isolation mechanisms for all service models. Further, it operates beyond specific enforcement points, thus providing the mechanisms to secure information flows across various parties and aspects of the infrastructure. Other approaches to wide-scale policy enforcement, such as those leveraging sticky policies [10], are complementary as they tend to be at higher levels of abstraction, where policy definition and interpretation is comparatively heavyweight.

As a general approach, IFC offers flexibility over who manages policy. For instance, end-users are able to express IFC policy for their data, which is respected by the tenant applications and cloud infrastructure. Alternatively, IFC policy can be managed by the cloud provider without tenant involvement to guarantee, for example, data-location requirements as in [11]. A mixture of these is also possible.

IFC can only provide guarantees above the level at which it is enforced. For instance, our implementation enforces IFC at the OS kernel level, and thus protects user-space flows but not lower-level aspects, e.g. concerning the hardware or the hypervisor. So we assume that a cloud provider that implements IFC does not actively try to circumvent its enforcement. This is not unreasonable, since a degree of trust in the cloud provider is implied by the use of a cloud service. This is reinforced by contractual relationships between tenants and cloud providers, and the role of regulators that operate in many jurisdictions. There is work on using trusted hardware components to enforce geo-location guarantees for cloud instances [12]. A similar approach could be taken for IFC policy enforcement.

IFC in managing personal data

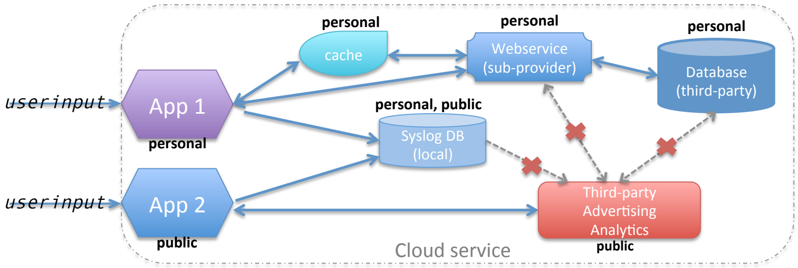

Figure 2: Separating personal and public data.

This figure shows two tenant applications, where 'App 1' is labelled as dealing with personal information. The data from 'App 1' can only flow to entities marked as properly dealing with personal information, and thus cannot flow to the 'Advertising' service which is not certified to process personal data. IFC ensures this constraint is adhered to as it flows throughout the cloud, including to sub-providers and other third parties. This is visible through audit, as each flow is recorded.

IV. Flow controls and legal case studies

IFC offers an additional security mechanism for cloud services and also provides the technical means for: a) greater control over data flows, helping parties to meet their data management responsibilities; and b) improved accountability, through providing detailed logs to demonstrate compliance and/or failures. This is illustrated in the following examples.

A. Compliance with contractual and data protection obligations

IFC is an effective mechanism for enforcing policy requirements set out in contracts and data protection regulations. We take as an example the stringent standards set out in the 2012 opinion by European data protection authorities on cloud computing [2], which requires cloud providers and tenants to implement technical and organisational measures guaranteeing to a cloud client (tenant or end-user):

- **Transparency.** There must be transparency concerning all sub-providers contributing to the provision of the cloud service, as well as the physical locations of all data centres in which personal data may be processed (including storage, caching and computation). These must be auditable by the tenant or end-user or by a certified third-party. Fig. 2 illustrates that IFC can help improve audit by providing fine-grained, data-centric constraints over data flows within the cloud, increasing visibility to those involved, rather than relying on external physical/process audits and certifications that may go out of date, e.g. by a simple configuration change. Auditable data flows provide the means for identifying (and bounding the fall-out from) situations of misconfiguration or compliance failures. Though the Figure focuses on sub-provider relations, location constraints can also be taken into account—we have explored this [11], and revisit it in the Microsoft-Ireland example below.
- **Purpose specification.** It must be guaranteed that personal data is not used in ways incompatible with the purposes for which it was originally provided. IFC provides information on the paths data has taken, which can be used to indicate proper usage at a fine-grained, data-centric level. Further, if the purpose can be encapsulated within labels, it could proactively ensure that purpose constraints are met.
- **Data erasure.** Guarantees must be in place to ensure that personal data no longer required is erased or truly anonymised. If this data cannot be erased due to legal retention rules (e.g. tax regulations), access to this personal data should be blocked. Since personal data may be kept in a number of locations, each instance (including fragments) must be erased irretrievably, including from backup, caches and potentially log files. IFC assists with erasure concerns. First, as data flows are audited, it is possible to determine where data has gone, and thus can ensure (verifiably) that the deletion requests are directed to all relevant entities. Further, if the erasure operation is a defined, encapsulated process, IFC could provide evidence that data was moved through a deletion operation (i.e. similar to recording what was sent to 'Trash').

B. Microsoft-Ireland: Location and law enforcement

Our second example concerns the ongoing Microsoft-Ireland legal dispute, in which a search warrant issued in a US drug trial has been claimed to extend to accessing a Microsoft customer's email data held exclusively in Ireland. Microsoft and a number of interested parties are appealing against the decision on the basis that it involves evading the EU-US Mutual Legal Assistance Treaty (MLAT), which ordinarily governs cross-border law enforcement requests to access personal data.

In this example, IFC offers the potential to regulate and audit data flow across jurisdictions, ensuring that any transfer of Irish user data to US authorities aligns to a visible, auditable and traceable process. This is illustrated in Fig. 3. The data residing in Ireland is tagged [Ireland] to represent this fact. For data to move across jurisdictions, the *integrity* tags require successful passage through a privileged Government Cross Border Transfer (GCBT) process, which if appropriate, changes the tags of the data to make it suitable to flow to [US]. The GCBT could reflect the application of an MLAT-approved request, but also offers more, in that any transfer abroad is forced through the process.

This example illustrates several benefits of an IFC approach. First, it shows the containment of data by location according to the affixed *integrity tags*. Data cannot simply flow to the US (or indeed other jurisdictions); it must travel a certain path. Second, all flows are recorded, including the flows in and out of the GCBT process. This means the data transferred can be audited, for example by Irish public authorities, to ensure that *only* data processed according to an MLAT-approved request is transferred. There is also the potential for individuals to verify that their information has not been transferred abroad. However, as this

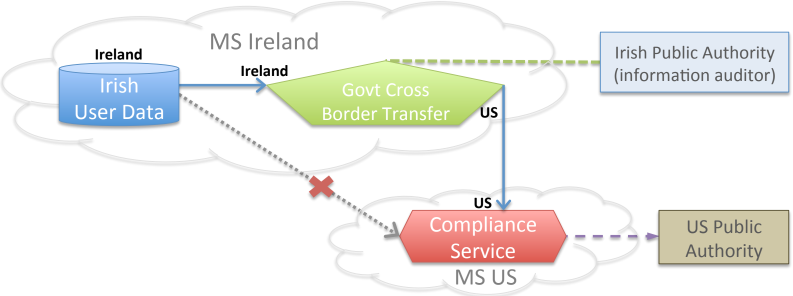

Figure 3: Illustration of the Microsoft-Ireland situation, where IFC can regulate the flow of information to the US for law enforcement, forcing it through a cross-border transfer process.

is not always appropriate for cases of law enforcement, access to the audit log would likely be protected through access controls.

The IFC approach we propose provides a stronger safety net than the current situation, when both MLAT and non-MLAT-processed requests are invisible and trust-based, both to individuals and to authorities, with few (technical) guarantees regarding data transfers.

C. Strict processing constraints, e.g. health data

Personal health data is intrinsically sensitive, but it can also be useful for medical research, and is essential for population health studies. For personal data to be used for medical research purposes, there are generally strict requirements about informed consent, anonymisation processes and appropriate ethics and governance frameworks. IFC can help ensure and provide evidence that these constraints and requirements are respected. For example, Fig 4 illustrates a simplified scenario concerning consent and anonymisation (in practice, more checks may be necessary) where:

- An anonymisation process (Anonymiser) takes personal medical data as input, producing an anonymised (according to some approved algorithm) version as an output.
- *Integrity* constraints ensure that the Anonymiser can only receive data where patients have consented to its use in medical research (i.e. with the tag consent). This allows a veto for those concerned about possible re-identification, or use of their data for research more generally.
- The Anonymiser outputs data with an *integrity* tag (anon) marking that the data has gone through the designated anonymisation process. The (personal) *secrecy* tag is removed, given the changed level of sensitivity.
- The Medical Research Database is tagged such that it will only receive—or be willing to accept, e.g. for reasons of liability and responsibility—data that has been anonymised, and where consent has been given.
- Individual researchers/projects are bound by same IFC constraints as the Medical Research Database. In transacting with the Medical Research Database, additional controls—such as *differential privacy* [13] techniques—could also apply.

We present this example to show how IFC can assist in the managing of particularly sensitive data and where additional regulatory requirements may pertain. Here the only means for personal data to flow to researchers is through a particular, designated anonymisation process, where consent is a prerequisite. As all flows are recorded, this can be audited. It allows patients who gave consent to see whether their data has actually been staged for research use. More complex constraints are also possible, such as consenting for use only in particular research projects.

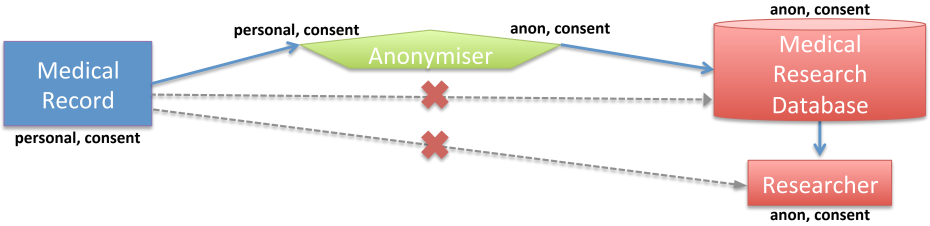

Figure 4: Illustration of the health-data example, involving the transfer of a personal medical record to a medical research database through an anonymisation process. Only the pertinent labels are shown.

Data processing regimes

The examples demonstrate the effect of the flow controls to essentially isolate (or more accurately, ensure non-interference of) data transmission and processing. This essentially sets up a *data processing regime*, potentially down to the level of a data-item, where flows of that data are transparent, and importantly, the flow of data into and out of that regime can be tightly managed. If the high-level management concerns can be specified in tags, then they can be technically enforced.

One can envisage cloud providers leveraging such mechanisms to offer services based on the processing regime, rather than on infrastructure aspects such as the service model. Sector-specific clouds, such as a *Financial Services Authority* compliant cloud are feasible, where data is guaranteed only to flow to processes certified as compliant with particular requirements. There is also the potential for tenants to define their own regimes. The extra control and transparency should improve levels of trust in cloud services [5], and therefore encourage cloud uptake; while effecting such controls imposes comparatively little effort on the cloud provider when compared to segregated infrastructure offerings.

There is ongoing discussion concerning the current conflation of legal jurisdiction and physical location. IFC is relevant to this, as it can ensure particular data management aspects based on and/or irrespective of the physical location of technical infrastructure. See [3] for discussion.

V. Moving forward

Though we argue that flow control offers a way forward in assisting with legal and regulatory dimensions of cloud computing, it is a novel area and research is ongoing. To date, there are no commercial IFC deployments by cloud services. Further work is needed to bring IFC to mainstream providers, including the resolution of issues of trusted enforcement, global naming schemes, policy authoring mechanisms, tag sensitivity and management, and so forth (see [7] for discussion). All of the examples discussed in this article have nevertheless been implemented in our prototype, demonstrating the promise of such an approach, and indicating that novel technical mechanisms can assist in enforcing higher level concerns. Again, our vision for IFC is to complement other management technologies, be they well-established or the subject of ongoing research.

It is also worth noting that the directions in cloud computing are towards smaller clouds. Ongoing research, including *cloudlets* [14] and *droplets* [15], aims at reducing the size and scope of cloud deployments. Such research is particularly important in the context of the emerging Internet of Things. With smaller clouds, cloud instances may be dynamically moved between hosts, e.g. from a portable device (collecting data while on the move) to a larger provider (to perform more complex processing on that data), facilitating mobility and location-based services, flexible processing, assisting in resource management, and so forth. Clouds can also be dedicated to particular purposes, e.g. one for an individual's fitness data, another concerning the usage of appliances in the home. As clouds become smaller, their function becomes more obvious and explicit. This could ease management by allowing more focused and informed policy definitions. Flow controls will become even more relevant, as more interactions (data flows) will occur within and between these clouds. Mechanisms enabling data management policy to apply continuously and consistently across applications and infrastructure will be crucial.

VI. Concluding remarks

As cloud computing becomes the dominant form of application and service provisioning, the need for more proactive and integrated responses to legal and regulatory requirements grows stronger. We have considered the primary contours of cloud providers' data management and compliance obligations. Tenant isolation plays an important role, but as cloud services mature, finer-grained, auditable, data-centric control mechanisms are required to help manage the information flows inherent in cloud provisioning.

We have argued that new technical mechanisms are required to enable control and transparency within and across cloud services and cloud-hosted applications—to "see through the clouds". Towards this, we have presented IFC as an exemplar, targeting issues of privacy, confidentiality and data quality system-wide, providing a means for demonstrable compliance with contractual and regulatory obligations. Such an approach has the potential to raise the general level of trust and confidence in cloud services.

Acknowledgement

The authors from the Computer Laboratory are supported by the UK Engineering and Physical Sciences Research Council under grant EP/K011510, and Microsoft through the Microsoft Cloud Computing Research Centre. We thank our colleagues at the Centre for Commercial Law Studies, Queen Mary University of London for their input and assistance with the legal aspects.

References

[1] C. J. Millard, Ed., *Cloud Computing Law.* Oxford University Press, 2013.

[2] European Commission, "Opinion 05/2012 on Cloud Computing," 2012. [Online]. Available: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp196_en.pdf

[3] K. Hon, C. Millard, C. Reed, J. Singh, I. Walden, and J. Crowcroft, "Policy, Legal and Regulatory Implications of a Europe-Only Cloud," Queen Mary University of London, School of Law, Tech. Rep., 2014, accessed: 14th October 2015. [Online]. Available: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2527951_code1577160.pdf

[4] J. Singh, J. Bacon, J. Crowcroft, A. Madhavapeddy, T. Pasquier, W. K. Hon, and C. Millard, "Regional Clouds: Technical Considerations," University of Cambridge, Tech. Rep. UCAM-CL-TR-863, 2014, accessed: 14th October 2015. [Online]. Available: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-863.pdf

[5] Cloud Security Alliance, "Security guidance for critical areas of focus in cloud computing v3.0," 2011, accessed: 14th October 2015. [Online]. Available: https://cloudsecurityalliance.org/research/security-guidance/

[6] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proc. 3rd ACM Workshop on Cloud Computing Security (CCSW).* ACM, 2011, pp. 113–124.

[7] T. Pasquier, J. Singh, D. Eyers, and J. Bacon, "CamFlow: Managed Data-Sharing for Cloud Services," 2015, arXiv:1506.04391. [Online]. Available: http://arxiv.org/abs/1506.04391

[8] A. C. Myers and B. Liskov, "A Decentralized Model for Information Flow Control," in *17th Symposium on Operating Systems Principles (SOSP).* ACM, 1997, pp. 129–142.

[9] J. Bacon, D. Eyers, T. Pasquier, J. Singh, I. Papagiannis, and P. Pietzuch, "Information Flow Control for Secure Cloud Computing," *IEEE TNSM SI Cloud Service Management*, vol. 11, no. 1, pp. 76–89, 2014.

[10] S. Pearson and M. C. Mont, "Sticky Policies: An Approach for Managing Privacy across Multiple Parties," *Computer*, vol. 44, July 2011.

[11] T. Pasquier and J. Powles, "Expressing and Enforcing Location Requirements in the Cloud using Information Flow Control," in *IC2E International Workshop on Legal and Technical Issues in Cloud Computing (Claw'15).* IEEE, 2015.

[12] K. R. Jayaram, D. Safford, U. Sharma, V. Naik, D. Pendarakis, and S. Tao, "Trustworthy Geographically Fenced Hybrid Clouds," in *ACM/IFIP/USENIX Middleware.* ACM, 2014.

[13] C. Dwork, "Differential privacy," in *Automata, languages and programming.* Springer, 2006, pp. 1–12.

[14] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The Case for VM-based Cloudlets in Mobile Computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, 2009.

[15] J. Crowcroft, A. Madhavapeddy, M. Schwarzkopf, T. Hong, and R. Mortier, "Unclouded Vision," in *Distributed Computing and Networking.* Springer, 2011, pp. 29–40.

Jatinder Singh is a Senior Research Associate at the Computer Laboratory, University of Cambridge. His research interests concern management control in distributed systems, particularly regarding cloud and the Internet of Things.

Julia Powles is a lawyer and researcher at the Centre for Intellectual Property and Information Law, University of Cambridge, where she works on the law and politics of information-based assets, from data privacy to patent law.

Thomas Pasquier is a PhD student and a Research Assistant at the University of Cambridge. His MPhil from Cambridge included a project on “Prevention of identity inference in de-identified medical records”.

Jean Bacon is a Professor of Distributed Systems at the University of Cambridge, and leads the Opera research group, focussing on open, large-scale, secure, widely-distributed systems.