Services for Sensitive Data (TSD) - data management challenges

Leon du Toit

2019-02-27

Outline

- TSD as a generic eInfrastructure
- unique data management challenges
- use cases
- generalisations

TSD generically

- logically
  - identities
  - policies: security, access control, usage
  - administrative data
- basic intrastructure
  - physical and virtual compute, networking
  - administrative systems
- services
  - compute
  - storage
  - data transport
  - data management (combining compute, transport, storage, policy)

Important dimensions of eInfrastructures

- eInfrastructures differ in:
  - scale: users, data volume, organisaitonal complexity
  - complexity: security, requirements, user competence
  - domain: potential consequences of service delivery failures

TSD overview

- 2500+ users, 562 active projects
- 3800 years of research
- 600+ VMs, 2 PB data
- 3000+ support cases per year,
- 12 people 2nd-line support, development, operations
- HPC cluster, FPGA gene sequencer, VM rig
- Sigma2, Tryggve2, EOSC, PRACE

TSD complexity

- security
  - machine room - physical separation
  - projects - layer 2, and later 3 network separation
  - Two Factor Authentication
  - own IdP, support for external IdPs with high Levels of Assurance
- requirements
  - HPC, DBs
  - mobile apps
  - open source and licensed software
  - Windows and linux
- end user competence
  - some users we could easily hire
  - others have basic computer literacy

TSD domain

- sensitive data
- bioinformatics
- medicine
- social science

Data management challenge, in general

1. Propagating and enforcing central security policies,
2. over different protocols,
3. to diverse systems,
4. at different levels of the technology stack,
5. while providing research services with good user experience

Use cases

1. Secure video recording
2. Survey data collection
3. Dynamic digital consent
4. Publishing patient data

1. Secure video streaming

- videos of patient therapy from Psychology department
- challenges
  - record video securely
  - efficient transfer to TSD
  - restrict access control to student and study leader
  - post process video files before playback
- solutions
  - custom desktop app for recording
  - streaming HTTPS upload, with personal login, no intermediate storage
  - use TSD identities and groups for file system access control
  - access control and processing peformed locally in the project

2. Survey data collection

- challenges
  - real-time delivery
  - centralised processing for automatic dataset creation
- real-time delivery
  - transitioning away from a batch pipeline
  - batch: ssh, rsync, copy to nfs mount
  - HTTPS: client POSTs data, ready in project
- dataset creation
  - decrypt individual responses
  - compile dataset incrementally as new responses arrive
  - scalability issues: security vs. operational load

3. Dynamic digital consent

- challenge
  - GDPR compliance
- solution
  - collect consent using survey tools
  - digital signatures of consent
  - revocable
  - APIs to check consent status and history
  - portals for consentors, and researchers
- vision
  - full feature implementation of the GDPR
  - enforce consent on collection and usage
  - right to insight
  - right to access
  - right to be forgotten

4. Publishing patient data

- challenges
  - enforcing ownership using an external IdP
  - project separation at the application layer
- ownership
  - allow researchers to share data, specifying ownership
  - use external identities with high LoA
- application layer project separation
  - maintain lower-level network configuration internally
  - look up project affiliation of HTTP request at request-time
  - combine with user credentials
  - remain open for new implementations of projects

Data flow policies

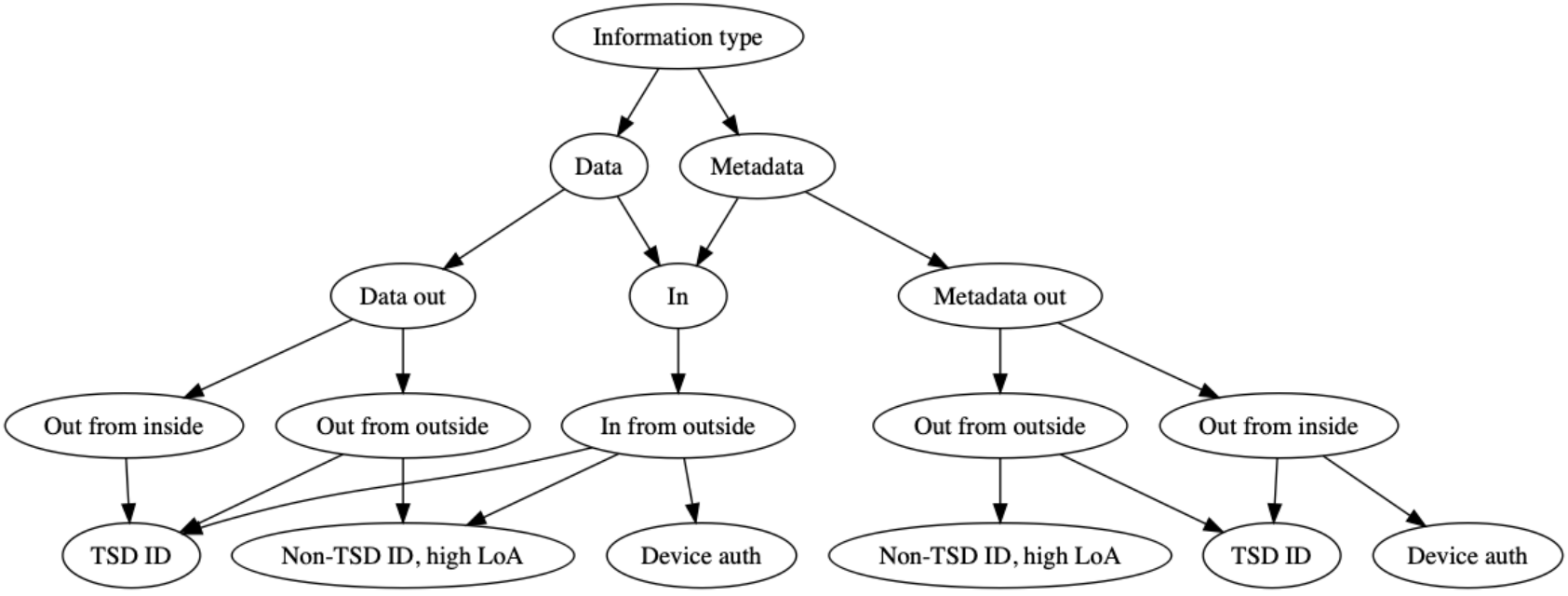

Figure 1: Required authentication

Recommendations (1/2)

- avoid intermediate data storage
- user and group based access control is always relevant
- create APIs which give client developers more capabilities
- end-to-end solutions are simpler to reason about and to maintain
- careful consideration about which details to expose to researchers
- create extensible services so projects can cater to very specific needs themselves
- centralise where possible to reduce operational complexity

Recommendations (2/2)

- separate policy from mechanism: enforce consent requirements on behalf of users
- security features best considered as inherent in system design
- embrace the web - mature, flexible, large talent pool and knowledge base
- difficult to enforce access control at many levels of the technology stack
- data ownership should be a first class concept in all systems