Digital Data Markets:
Trusted Data Processing in Untrusted Environments

Cees de Laat

Systems and Networking Laboratory
University of Amsterdam
Fading Trust in Internet

Trust

Dependency

Research Gap!

1980

2018
Data value creation monopolies

Create an equal playing field

Sound Market principles


Harvard Business Review
Main problem statement

• Organizations that normally compete have to bring data together to achieve a common goal!
• The shared data may be used for that goal but not for any other!
• Data may have to be processed in untrusted data centers.
  – How to enforce that using modern Cyber Infrastructure?
  – How to organize such alliances?
  – How to translate from strategic via tactical to operational level?
  – What are the different fundamental data infrastructure models to consider?
Health use case
Enabling Personal Interventions

RQ4: Adaptive Health Diagnostics

Self Management & Group Management Support providing Advise, Diagnostics, Prognostics.

RQ 1,2 Data Science Algorithm Development

RQ3: Data & Algorithm Distribution

Digital Health Twin

RQ 5: Regulatory Constraints & Data Governance

RQ 6: Infrastructure

Patient Data Delivery System

Patient Data

Medical Use Cases

Health Practitioners Context & Experience

Health product manufacturers

Medical Use Cases

Health Practitioners Context & Experience
Big Data Sharing use cases placed in airline context

### Global Scale
- Aircraft Component Health Monitoring (Big) Data
  - NWO CIMPLO project
  - 4.5 FTE

### National Scale
- Cargo Logistics Data
  - (C1) DL4LD
  - (C2) Secure scalable policy-enforced distributed data Processing (using blockchain)

### City / regional Scale
- NLIP iShare project

### Campus / Enterprise Scale
- Cybersecurity Big Data
  - NWO COMMIT/SARNET project
  - 3.5 FTE
RESEARCH WORKING ALONGSIDE IT INDUSTRY

NETWORK RESEARCH INFRASTRUCTURES

- Data Sharing Infrastructure Model
- Research using Future Internet capabilities

COMMERCIAL DATACENTER INFRASTRUCTURE AS NEUTRAL GROUND

- Goal: How to create a Digital Marketplace Ecosystem

Gobal Lambda
Integrated Facility ➔
Approach

• Strategic:
  – Translate legislation into machine readable policy
  – Define data use policy
  – Trust evaluation models & metrics

• Tactical:
  – Map app given rules & policy & data and resources
  – Bring computing and data to (un)trusted third party
  – Resilience

• Operational:
  – TPM & Encryption schemes to protect & sign
  – Policy evaluation & docker implementations
  – Use VM and SDI/SDN technology to enforce
  – Block chain to record what happened (after the fact!)
Data Processing models

- Bring data to computing
- Bring computing to data
- Bring computing and data to (un)trusted third party
- A mix of all of the above
- Block chain to record what happened
- Block chain for data integrity
- Bring the owner of Data in control!
- Data owner policy + enforcement technology
IETF: Common Open Policy Service (COPS)

- Rfc 2748, 2753, 4261
Secure Policy Enforced Data Processing

- Bringing data and processing software from competing organisations together for common goal
- Docker with encryption, policy engine, certs/keys, blockchain and secure networking
- Data Docker (virtual encrypted hard drive)
- Compute Docker (protected application, signed algorithms)
- Visualization Docker (to visualize output)
Networks of ScienceDMZ’s & SDX’s

- Internet
- Peer ISP’s
- Supercomputing centers (NCSA, ANL, LBNL)

- ISP
- DTN
- SDX
- NFV

Ownership/trust relation

- client 1
- client 2
- client 3
- client 4
- client n

Petabyte email service 😊

Contains a

DMZ

DTN
What have we been doing?

- Studying and defining draft Policy
- Working out model & defining Archetypes
- Implementing a proof of concept using several distributed DTN’s and dockers on kubernetes.
- Working on a demo for SC18 in Dallas TX, 11-16 Nov.
- Tactical operation of Digital Data Markets
- Optimization of degrees of freedom == value
SC16 Demo

DockerMon

Sending docker containers with search algorithms to databases all over the world.

http://sc.delaat.net/sc16/index.html#5
SC17 Posters and proof of concepts & demo’s

http://sc.delaat.net/sc17

Unlocking the Data Economy via Digital Marketplaces
Researching governance and infrastructure patterns in airline context

Use Case: Sharing Aircraft Data to develop a Maintenance Credit System
- A Digital Marketplace estimates data time before maintenance is needed, after data is received from a corresponding aircraft system.
- Algorithm quality increases when data, owned by different aviation operators, can be shared during its development.
- Sharing data assets creates value (e.g. non-compensation).
- Research Question: Can the Digital Marketplace concept organize trust amongst its stakeholders to enable common benefits for a single organization to achieve, whilst observing economic principles?

Digital Marketplace as a means to organize trusted data asset sharing
A Digital Marketplace is a membership organization identified by a common goal: Share data to enable development of a Maintenance Credit System.
Membership organization is institutionalized to create, implement and enforce membership rules.
Member members create digital agreements. Agreements are translated into different software defined infrastructures using infrastructure patterns offered by a Digital Marketplace Ecosystem.

Examples of infrastructure patterns offered by a Digital Marketplace
- Public cloud model
- Container model
- Turntable model

DEMONSTRATION: LIGHT PATHS AND DATA TRANSFER NODES FOR AIRCRAFT MAINTENANCE

Data Transfer Node (DTN) Workflows

Why Data Transfer Nodes (DTNs)
- DTNs can act as an interface to a high-performance link
- Configurable to maximize performance for a given workflow
- Simplifies configuration of client systems
- Multiple clients may share a DTN
- DTNs strategically placed to best benefit clients
- DTNs can be compared to specialized high-speed transport systems of the past

Example: Entry Point for High Speed Transport
A typical use case for DTNs is as a high-speed file transfer service. A computer system’s configuration may allow for the utilization of all available bandwidth in a LAN environment. However, it often the case that in a WAN environment with high latency or packet loss, the same system performs poorly. A DTN could be used to maximize performance on a high-speed link between two locations by breaking up the transfer into smaller segments and routing them through a DTN.

Example: Storage Access Point
Another possible use case for DTNs is to be used to access distributed data from remote locations. In this scenario a system located at a corporate facility requests the data from the local DTN as it is required. This DTN would then transparently route the data from multiple remote sites as received. In contrast to the first example here block level access is provided by the DTN. To the system performing the computations the locally DTN appears to be the actual and the remote datasets to be a local block level access. On the other site DTN may perform same caching, there need not be permanent storage of data at the compute facility.
Training AI/ML models using Digital Data Marketplaces

Creating value and competition by enabling access to additional big data owned by multiple organizations in a trusted, fair and economic way

- AI/ML algorithm-based Decision Support Systems create business value by supporting frontline decision making such as predicting the need for aircraft maintenance.
- Algorithm quality increases with the availability of aircraft data.
- Multiple airlines operate the same type of aircraft.

Research Question: “How can AI/ML algorithm developers be enabled to access additional data from multiple airlines?”

Approach: Applying Digital Data Marketplace concepts to facilitate trusted big data sharing for a particular purpose.

A Digital Data Marketplace is a membership organization supporting a common goal e.g., enable data sharing to increase value and competitiveness of AI/ML algorithms.

Membership organization is institutionalized to create, implement and enforce membership rules organizing trust.

Market members arrange digital agreements to exchange data for a particular purpose under specific conditions.

Agreements subsequently drive data science transactions creating processing infrastructures using infrastructure patterns offered by a Data Exchange as Exchange Patterns.

Dataharbours: computing archetypes for digital marketplaces

Regioud Cordius, Li Zhang, Paolo Grossu, Tim van Zalingen, Joseph Hill, Leon Gommans, Coes de Laut, Vijay Dasawazwan, Purnell Pursbit, Kaladhar Voruganti, Craig Wadstrup, Rodney Wilson, Marc Lynnaes

The problem

How can competing parties share compute and data? The architecture of a digital marketplace is an active research field and has many components to it. Here we investigate a relevant computing platform which is mapped into different archetypes based on trust relationships between organizations.

The components

Concerning: an initial document which brings together organizations that wish to collaborate. It defines static information such as keys to identify parties.

Infrastructure: A single co-located organization infrastructure that secures basic data, compute infrastructure and, optionally, compute infrastructure. The data exchange marketplace implements a set of protocols that allows it to interact with other businesses.

Contracts: Are a set of rules that are shared amongst participating harbours which describe how objects/data/computers can be traded between harbours and who can process data. In its simplest form it is a 3-tuple which binds a user, data object, compute container; contract, connections, and expiry data.

An application is a distributed pipeline which can make use of several harbours. The combination of application and contract defines the archetype of the computation i.e., how data and compute are moved to effect computation.

Auditor: A trusted entity that collects audit trades for use in litigation of policy violations.

In action

Pederized computing on 3 distributed data harbours. Here we illustrate one archetype where KLM and Airfrance do not trust each other and employ a trusted 3rd party to send the data and compute for them.

For the scenario to succeed the different harbours need to effect several transactions which are governed by contractual rules.

The transaction protocol involves first identifying both parties who then negotiate a contract to address challenges and security. The at least one contract is used to drive the transaction. Important steps of the transaction are audit logged i.e., signed and published to an audit log collector.

Researching Exchange Patterns to support Digital Data Marketplaces

Research Infrastructure

The more data - the better: an aircraft maintenance use-case

Digital Data Marketplace enabling data sharing and competition

Research Elements

Data Exchange Model

Research Infrastructure

Research Elements

The more data - the better: an aircraft maintenance use-case

Digital Data Marketplace enabling data sharing and competition

Research Elements

Data Exchange Model

Research Infrastructure

Research Elements
INFRASTRUCTURE PATTERN EXAMPLES

OFFERED BY A DATA EXCHANGE TO MARKETPLACES TO CHOOSE FROM
DMP archetypes and their representation

On the left one of the many collaboration models within a DMP. We call this archetype. One DMP can support multiple archetypes depending on the contracts between partners.

To match application/user requests to the archetype we need to model the archetype on the left in generic ways.
Parties in the DMP collaborate across a number of scopes: data, computing and output. They share data, they share algorithms and they can share results. The matrix on the left represents the level of collaboration between two parties in each of the scope. In the previous slide we had four parties plus the DMP exchange, so we have a 5x5 matrix.
Matching requests

Our work is to match a customer application to the ‘closest archetype’.
Dimensions of DMP

Coverage

How well can we satisfy users request with the available archetypes?

Extensibility

Can a DMP provide more archetypes to user?

Precision

How well the archetype database of a given DMP fits a request by customer?

Flexibility

How easily the requests from potential customers could be satisfied?

We defined four metrics to determine the ‘richness’ of a DMP.
Q&A

• More information:
  – http://delaat.net/dl4ld and http://delaat.net/epi
  – https://towardsamdex.org

• Contributions from:
  – Leon Gommans, Wouter Los, Paola Grosso, Yuri Demchenko, Lydia Meijer, Tom van Engers, Reggie Cushing, Ameneh Deljoo, Sara Shakeri, Lu Zhang, Joseph Hill, Lukasz Makowski, Ralph Koning, Gleb Polevoy, Tim van Zalingen, and many others!
Data Hub System Applicability

**Industry**
- Cross Cutting Field lab
- Innovation with SURF

**Science**
- European Open Science Cloud
- FAIR model
  - Findable – Accessible – Interpretable - Reusable

**Society**
- Smart Cities & Arena
- Streaming Data Decision Support
Experimental Setup

Data Transfer nodes at UvA, KLM and Equinix
Running Kubernetes with a number of dockers (pods) see below.

```
tim@uva-kube-04:~$ kubectl get pods -o wide
NAME                             READY   STATUS    RESTARTS   AGE   IP             NODE          NOMINATED NODE
be2-deployment-c87646848-wt8l5  1/1     Running   0          77m   192.168.5.39   eqx-kube-03   <none>
mq1                             1/1     Running   0          13d   192.168.1.2    uva-kube-02   <none>
oex.airfrance                   1/1     Running   0          10d   192.168.4.16   eqx-kube-02   <none>
oex.klm                          1/1     Running   0          10d   192.168.3.5    eqx-kube-01   <none>
oex.trusted                     1/1     Running   0          13d   192.168.1.3    uva-kube-02   <none>
planner1                       1/1     Running   0          10d   192.168.4.15   eqx-kube-02   <none>
```

- be2 is the backend for the website.
  - It serves the static pages and passes new information and input to the planner.
- mq1 is the message queue that each oex writes logs to.
- oex.* each is one zone or 'object exchange server'.
  - It is responsible for handling requests from others and sending requests to other parties.
  - It should do so in accordance with a preselected contract/archetype.
- planner1 handles requests from be2 when selecting an archetype (and passes it on to the oex's) and when an application or pipeline is started (and passes it on to the oex).
DataHarbours: Computing archetypes for digital marketplaces

Reggie Cushing, Tim van Zalingen, Joseph Hill, Lu Zhang, Paola Grosso, Cees de Laat, Leon Gommans, Vijaay Doraiswamy, Purvish Purohit, Kaladhar Voruganti, Craig Waldrop, Rodney Wilson, Marc Lyonnais
SAE Use Case envisaged research collaboration

**Funding Agency**

- Big Data Hub / Spoke or Industry initiative funding

**International Networking**

- CENIC
- SoX
- LEARN
- SURFNET

**Regional / National Networking**

- Stanford
- Georgia Tech
- UT Dallas
- Universiteit van Amsterdam

**Local University**

**Aircraft MRO, OEM & Operators**

- Boeing
- Delta Air Lines
- Bell Helicopter
- Air France KLM

**Industry Standards Body**

- SAE AeroSpace Group
  - HM-1 working group
  - Use Case on aircraft sensor Big Data

**System and Network Engineering**

**AIR FRANCE KLM**
data Sharing approach: Combine 2&3

MANAGED BY AN INDEPENDENT INDUSTRY MEMBERSHIP ORGANIZATION


Independent Marketplace data platform governed by industry membership organization