

Digital Data Markets: Trusted Data Processing in Untrusted Environments

Cees de Laat

Systems and Networking Laboratory

University of Amsterdam



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?

Harvard Business Review



Harvard Business Review

Q

Subscribe | Sign In | Register


ECONOMY

Managing Our Hub Economy

by Marco Iansiti and Karim R. Lakhani

FROM THE SEPTEMBER-OCTOBER 2017 ISSUE

WHAT TO READ NEXT



The IT Transformation Health Care Needs

SUMMARY

SAVE


SHARE

COMMENT

TEXT SIZE

PRINT

\$8.95 BUY COPIES



THOMAS M. SCHEER/EYEEM/GETTY IMAGES

I. The Problem

The global economy is coalescing around a few digital superpowers. We see unmistakable evidence that a winner-take-all world is emerging in which a small number of “hub firms”—including Alibaba, Alphabet/Google, Amazon, Apple, Baidu, Facebook, Microsoft, and Tencent—occupy central positions. While creating real value for users, these companies are also capturing a disproportionate and expanding share of the value, and that’s shaping our collective economic future. The very same technologies that promised to democratize business are now threatening to make it more monopolistic.

Data value creation
monopolies



Create an equal
playing field

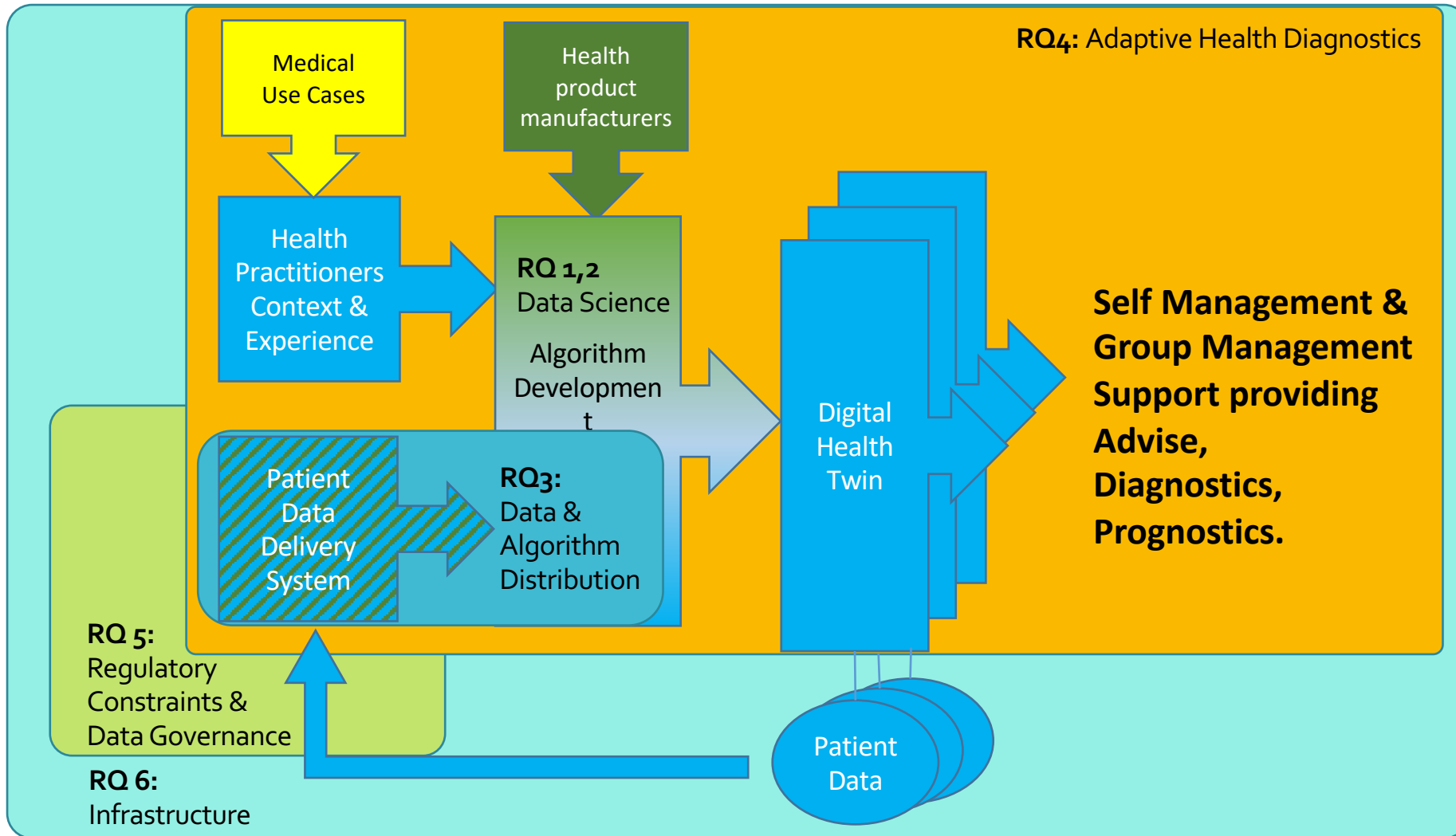


Sound Market
principles

<https://hbr.org/2017/09/managing-our-hub-economy>

Health use case

Enabling Personal Interventions



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)



Cybersecurity Big Data
NWO COMMIT/
SARNET project
3.5 FTE

**City /
regional Scale**

**Campus /
Enterprise Scale**

NLIP iShare project



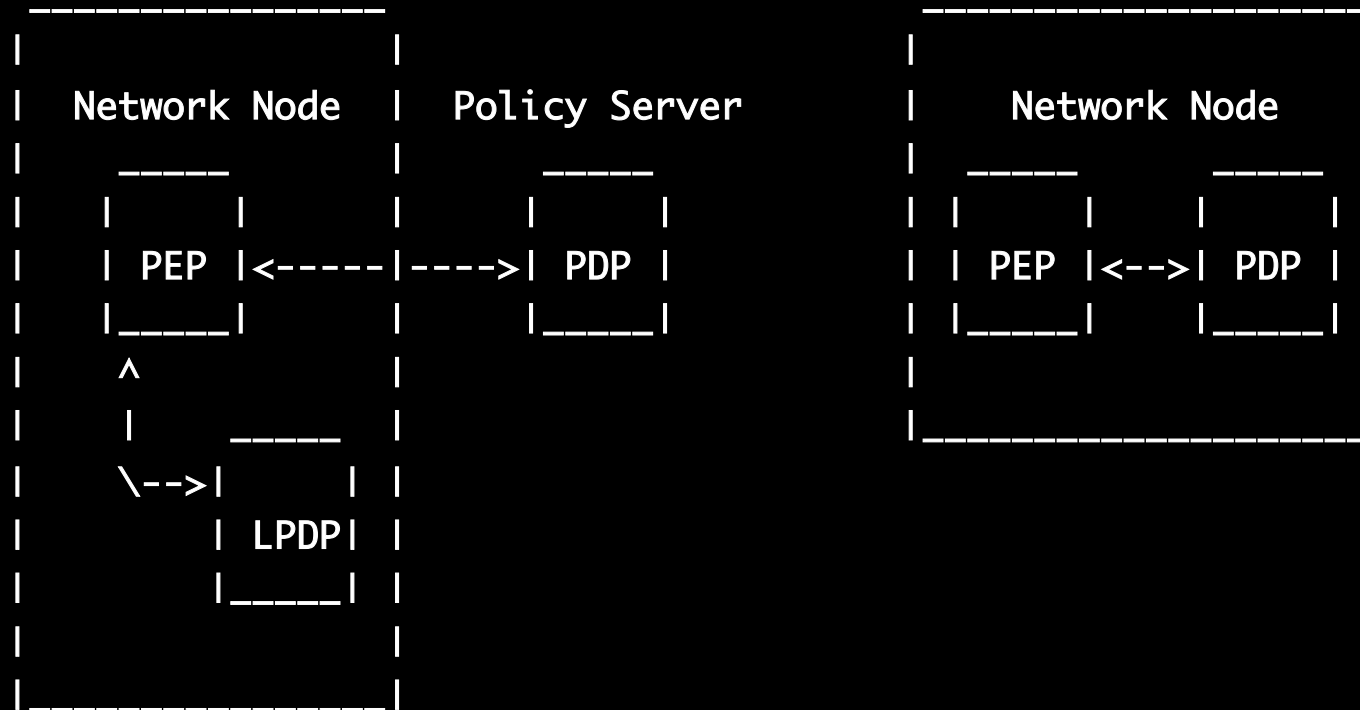
Approach

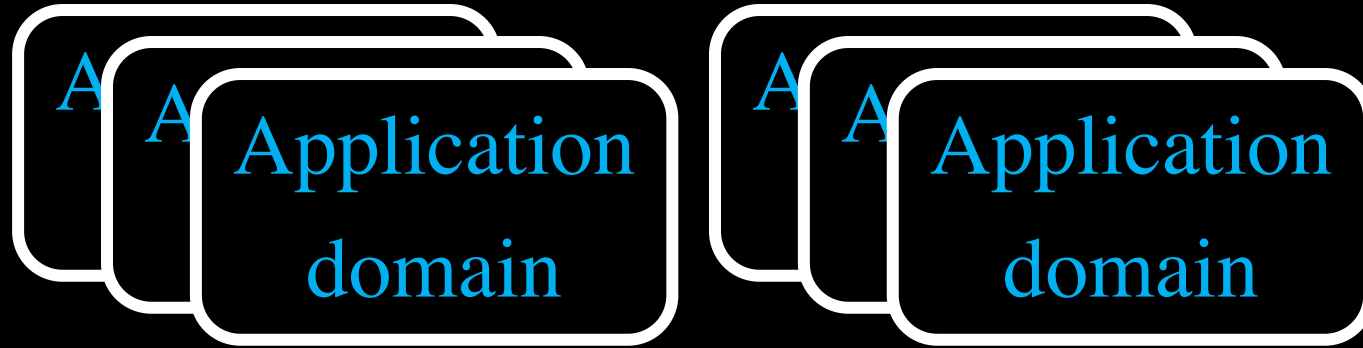
- Strategic:
 - Translate legislation into machine readable policy
 - Define data use policy
 - Trust evaluation models & metrics
- Tactical:
 - Map app given rules & policy & data and resources
 - Bring computing and data to (un)trusted third party
 - Resilience
- Operational:
 - TPM & Encryption schemes to protect & sign
 - Policy evaluation & docker implementations
 - Use VM and SDI/SDN technology to enforce
 - Block chain to record what happened (after the fact!)



IETF: Common Open Policy Service (COPS)

- Rfc 2748, 2753, 4261





AmDex

Data objects & methods
Data & Algorithms service

FAIR / USE

AmsIX

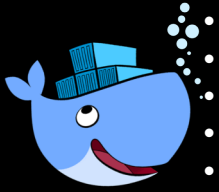
Routers - Internet – ISP's - Cloud
IP packet service

IP / BGP

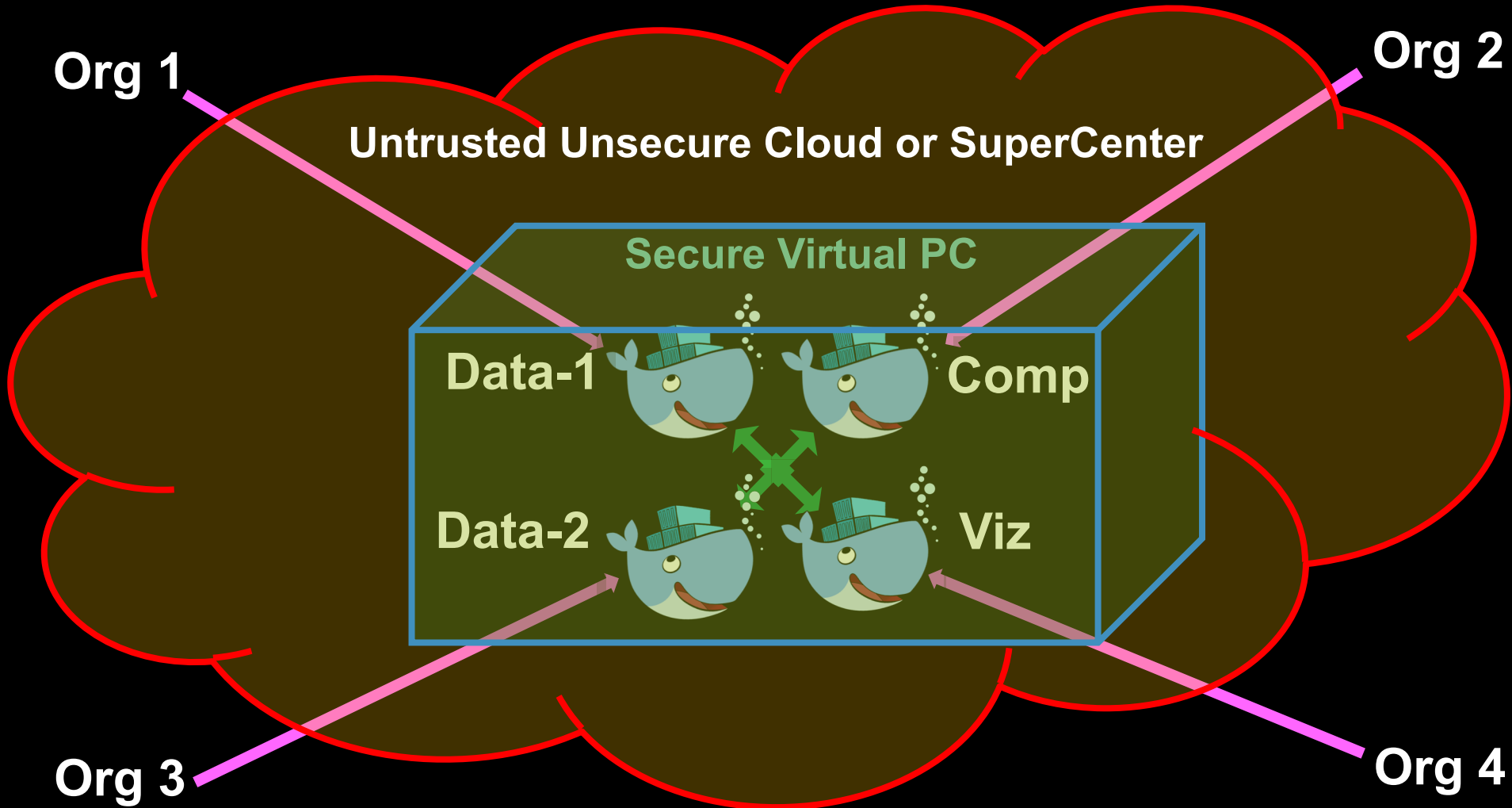
Layer 2 exchange service
Ethernet frames

ETH / ST

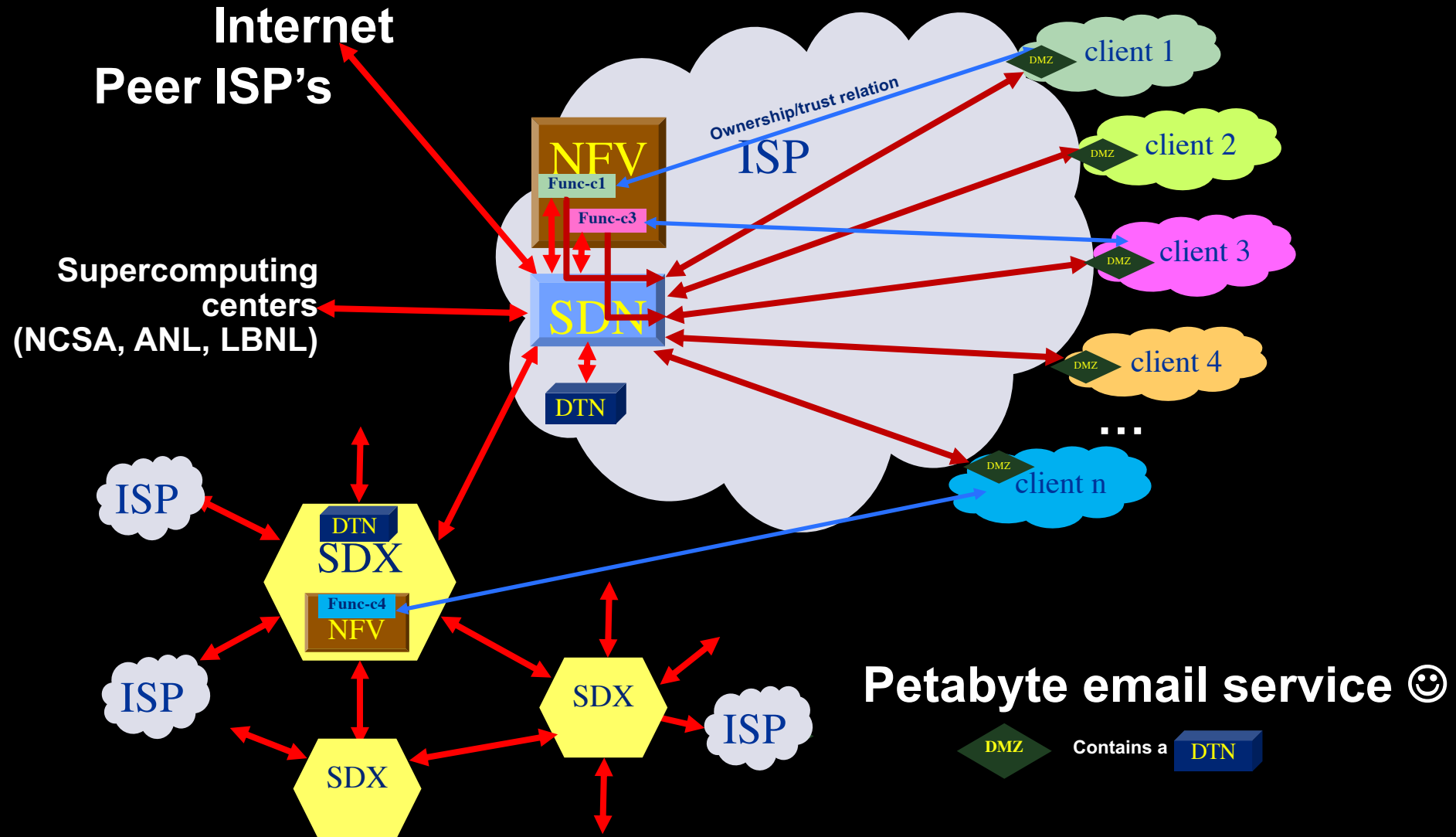
Secure Policy Enforced Data Processing



- 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




SC16 Demo


DockerMon

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

<http://sc.delaat.net/sc16/index.html#5>


UNIVERSITEIT VAN AMSTERDAM

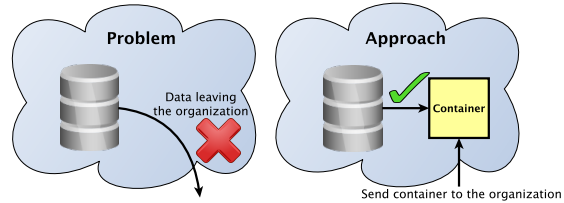
Łukasz Makowski, Daniel Romão, Cees de Laat, Paola Grosso
System and Networking Research Group, University of Amsterdam


System and Network
Engineering

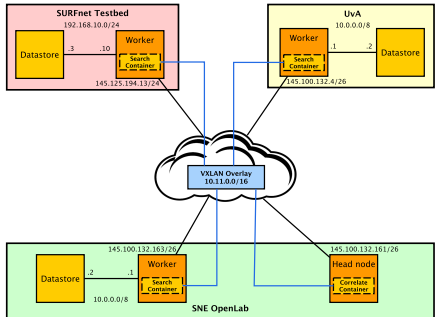
Container-based remote data processing

Problem Description

- Scientific datasets are usually made publicly available
...but data cannot always leave the organization premises
- On-site data processing can be challenging because of incompatibility of systems or lack of manpower
- Can a container-based system perform remote on-site data processing efficiently?
- What are the networking issues to solve?



Underlay and Overlay



Main features:

- Networked containers
- VXLAN overlay
- Containers that perform data retrieval and computation
- Containers built on-demand
- On-site data processing
- Distributed data source
- Multiple sites with datasets

The Game

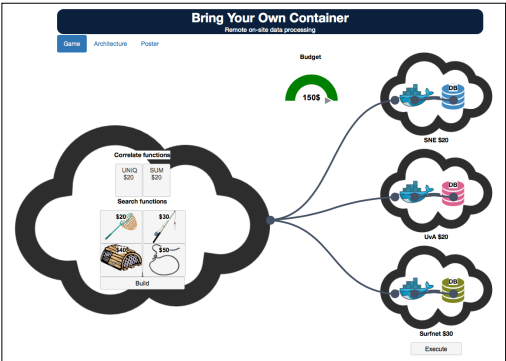
Our SC16 demo is a gamification of the remote dataset processing architecture.

How many different animal species can you find? You have a fixed budget and each function and processing will cost you money!

In our game you will:


- Select a correlate function to combine the results of the different sites.
- Pick different search functions, represented as tools, to find animals in the remote datasets.
- Build containers with the search and correlate functions.
- Execute the containers on the sites of your choice.


Will you have the best score?



More information:

- <http://byoc.lab.uvalight.net/info>
- <http://sne.science.uva.nl/sne/gigaport3>
- <http://delaat.net/sc>

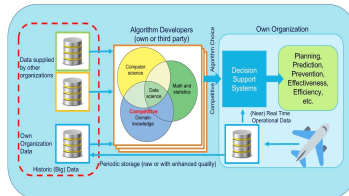




SC18 – Dallas TX

Training AI/ML models using Digital Data Marketplaces

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



- AI/ML algorithm based Decision Support Systems create business value by supporting real-time complex decision taking 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.

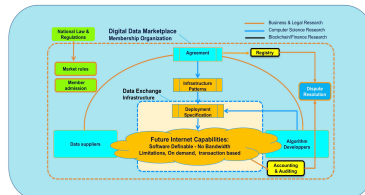
Digital Data Marketplace enabling data sharing and competition

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.*

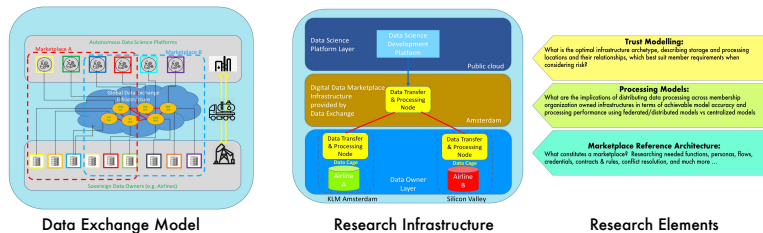
Membership organization is institutionalized to create, implement and enforce membership rules organizing **trust**.

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**.



Researching Exchange Patterns to support Digital Data Marketplaces

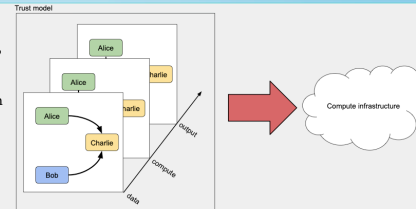


Dataharbours: computing archetypes for digital marketplaces

Reginald Cushing, Lu Zhang, Paola Grosso, Tim van Zalingen, Joseph Hill, Leon Gommans, Cees de Laat, Vijaay Doraiswamy, Purvish Purohit, Kaladhar Voruganti, Craig Waldrop, Rodney Wilson, Marc Lyonnais

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 federated computing platform which is molded into different **archetypes** based on **trust** relationships between organizations.



The components

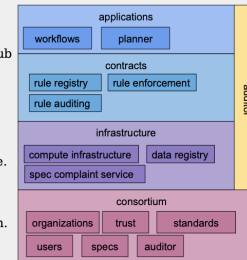
Consortium: is an initial document which brings together organizations that wish to collaborate. It defines static information such as keys to identify parties.

Infrastructure: A single domain organization infrastructure that securely hosts data, compute containers and, optionally, compute infrastructure. We dub this infrastructure a **data harbour**. A harbour implements a set of protocols that allows it to interact with other harbours.

Contracts: Are a set of rules that are shared amongst participating harbours which describe how objects (data, compute) can be traded between harbours and who can process data. In its simplest form is a 7-tuple which binds a user, data object, compute container, contract, consortium, harbour, and expiry date.

An application: Is a distributed pipeline which can make use of several contracts. 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 trails for use in litigation of policy violations.

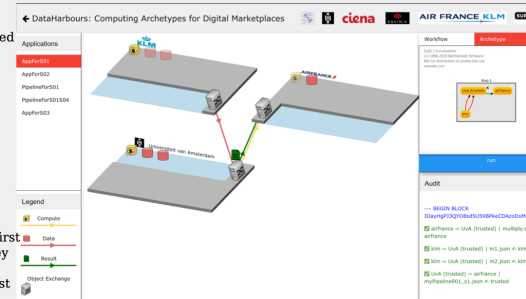


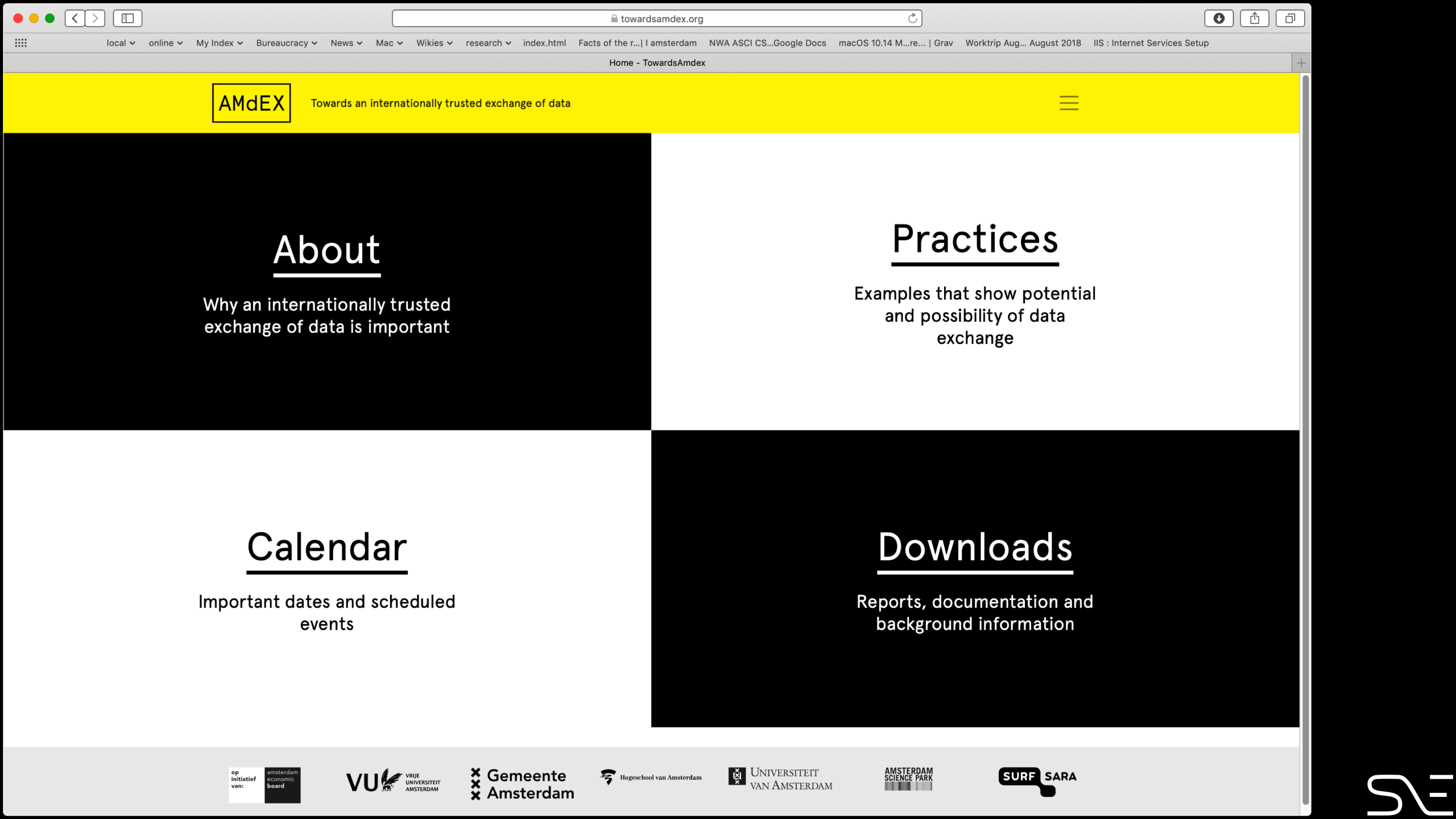
In action

Federated 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 processing.

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 are who they say they are through pub/priv key challenges and secondly, that at least a **contract** rule is matched to allow the transaction. Important steps of the transactions are **audit** logged i.e. signed and published to and audit log collector.





AMdEX

Towards an internationally trusted exchange of data

About

Why an internationally trusted exchange of data is important

Practices

Examples that show potential and possibility of data exchange

Calendar

Important dates and scheduled events

Downloads

Reports, documentation and background information



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!

