Distributed Big Data Assets Sharing & Processing

Big Data Hub infrastructure

Jan Wester, C. de Laat, L. Gommans

TNO
System & Network Engineering, University of Amsterdam
AirFrance KLM
Fading Trust in Internet Dependency

1980

2017

Research Gap!
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?
Big Data Sharing use cases placed in airline context

Global Scale

National Scale

City / regional Scale

Campus / Enterprise Scale

Aircraft Component Health Monitoring (Big) Data
NWO CIMPLO project
4.5 FTE

Cargo Logistics Data
(C1) DaL4LoD
(C2) Secure scalable policy-enforced distributed data Processing
(using blockchain)

Cybersecurity Big Data
NWO COMMIT/SARNET project
3.5 FTE

NLIP iShare project
SAE Use Case envisaged **research** collaboration

**Funding Agency**
- NSF
- NWO

**International Networking**
- iNET
- ESnet
- GÉANT

**Regional / National Networking**
- CENIC
- SoX
- LEARN
- SURFNET

**Local University**
- Stanford
- Georgia Tech
- UT Dallas
- Universiteit van Amsterdam

**Aircraft MRO, OEM & Operators**
- Boeing
- Delta Airlines
- 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**
Example model: Policy Enforced Data Processing

- Bringing data and processing software from competing organizations 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)
Ambition to put capabilities into fieldlab

100 Gb/s

R & E Networks

100 Gb/s

Lightpath

100 Gb/s

Light

ciena

GENI Testbed

SARNET Capable
Cyber-defense

Digital Airport
AMS

CDG

ATL

Big Data sharing
Fast Data Replication

Data Transfer
Node

Private & Secure
Collaboration

Application & Service
chains deployed in
private and secure
Internet slices

SAGE2

SAGE2 Server

Re-enforcing ICT preconditions:
Each envisaged site has similar elements
Science-DMZ

Border Router

Enterprise Border Router/Firewall

WAN

10G Routed

10G Virtual Circuit

Clean, High-bandwidth path to/from WAN

Site / Campus access to Science DMZ resources

Site / Campus LAN

Science DMZ Switch/Router

perfSONAR

High performance Data Transfer Node with high-speed storage

performed security policy control points

High Latency WAN Path

Low Latency LAN Path

High Latency VC Path

CENIC

ESnet
Research goal:
Explore value of academic network research capabilities that enable innovative ways & models to share big data assets.
Networks of ScienceDMZ’s & SDX’s

Internet
Peer ISP’s

Supercomputing centers (NCSA, ANL, LBNL)

ISP

NFV

SDN

ISP

SDX

Func-c1

Func-c3

Func-c4

DTN

DMZ

client 1

client 2

client 3

client 4

client n

Petabyte email service 😊

contains

DTN
Validation Fieldlab and Dissemination

- Experimental facilities from day one!
- Proof of concepts demonstrating secure data sharing
- Blueprint, roadmap and standards where applicable
- Model for FAIR EOSC Infrastructure

UVA - OpenLab
- KLM
- NetherLight
- GENI
- Fed4Fire
- Cloud
- SURFSARA
- ...

TNO - Intrepid
- Smart Data Factory
- Innovations
- Smart Rail
- To-Grip
- ...

C2D – Big DataHubs
- Arena
- KAVE
- AZURE
- Use Cases
- ...

DataHub
DTN

DTN
DataHub

DTN
DataHub

DTN
DataHub

DTN
DataHub

DataHub
DTN

DataHub
DTN

DataHub
DTN
Program at Global Summit I2 in Washington DC April 2017:
15h00 Cees de Laat, University of Amsterdam
   Trusted Data Processing in Untrusted Environments.
15h05 Leon Gommans, Air France KLM
   Trusted Big Data Sharing.
15h25 Rodney Wilson
   Programmable Supernetworks, Science DMZ based Networking.
15h30 Panel of stakeholders Flash talks (~3 min each):
   Inder Monga - ESnet - Data Science Driving Discovery.
   Matt Zekauskas - Internet2 - Thoughts on Internet2 and Trusted Large Data Transfer.
   Jerry Sobieski - NORDUnet - Issues of Big Data Sharing in a Global Science Collaboration.
   Adam Slagell – NCSA - What are we trusting?
15h45 Panel discussion moderated by Cees de Laat
16h00 End of session.