

# Data as a Service, Data Marketplace and Data Lake – Models, Data Concerns and Engineering

Hong-Linh Truong  
Faculty of Informatics, TU Wien

[truong@dsg.tuwien.ac.at](mailto:truong@dsg.tuwien.ac.at)  
[@linhsolar](http://dsg.tuwien.ac.at/staff/truong)

# Outline

- Data-as-a-Service concepts
- Data governance & Data concerns for DaaS
- Evaluating data concerns
- Data marketplace
- Datalake

# From previous student projects

„Use of several health, food and recipe services, in order to collect general food information“

“Latest data on air quality is fetched from London Air API”

„collect location-data from multiple Sources .... combine location- with social-data“

„real time production information from photovoltaic panels“

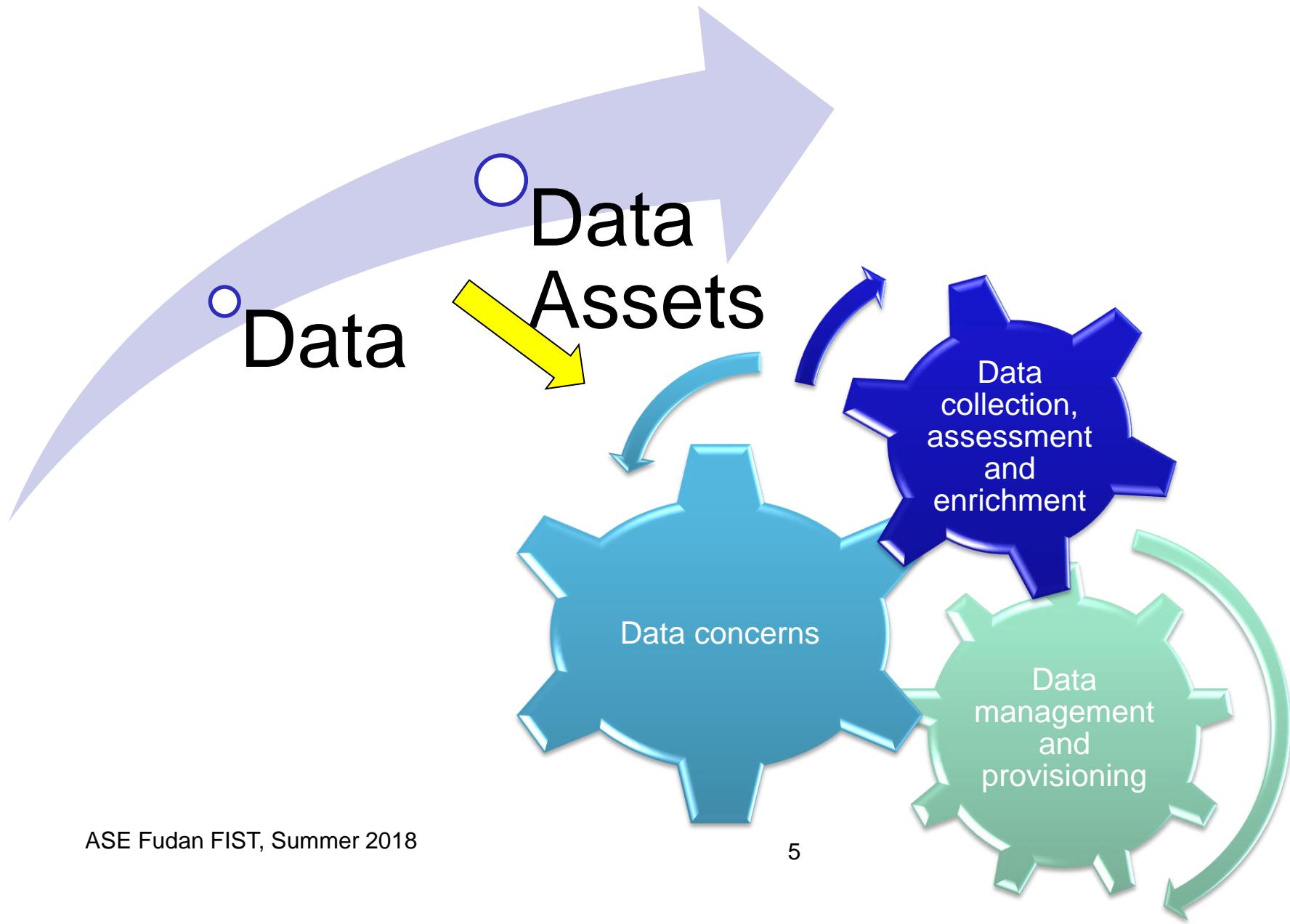
“running face recognition on customers entering the store, captured by security cameras, and assigning them anonymous identities which persist across store visits.”

“Measure and report water quality metrics”

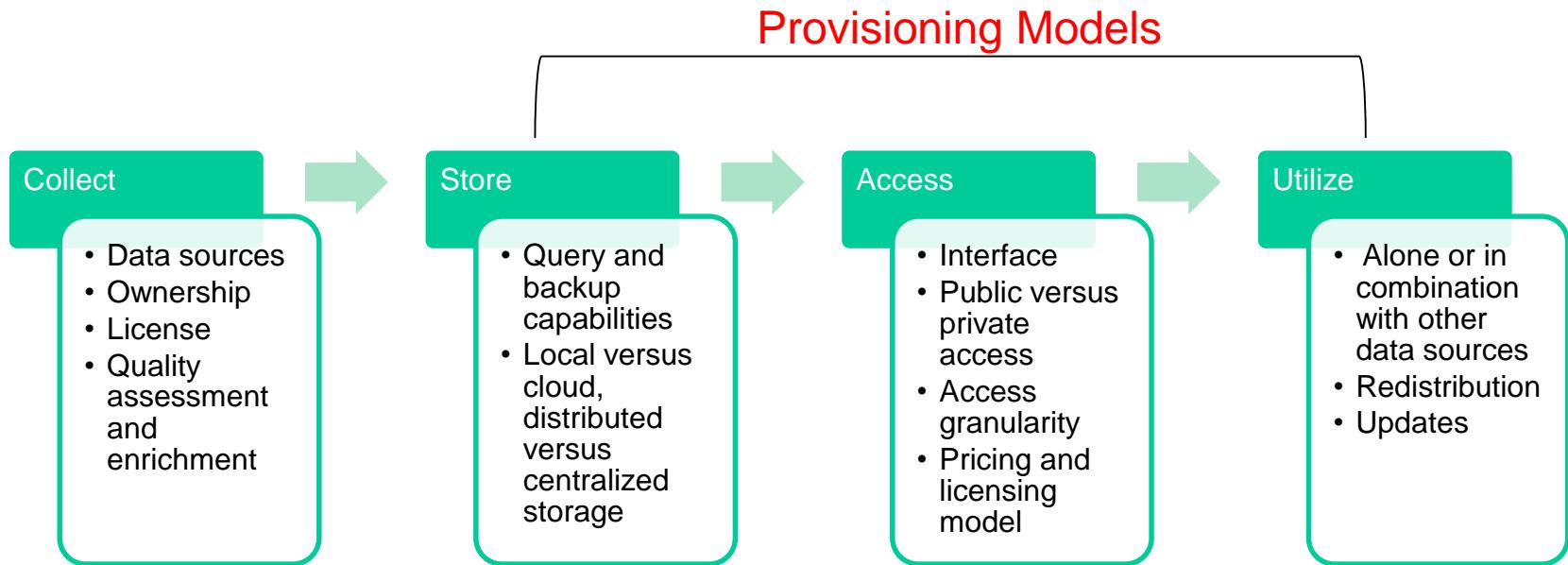
“give data about crimes in an area .... ranking of data quality ”

# DATA AS A SERVICE

# Data versus data assets

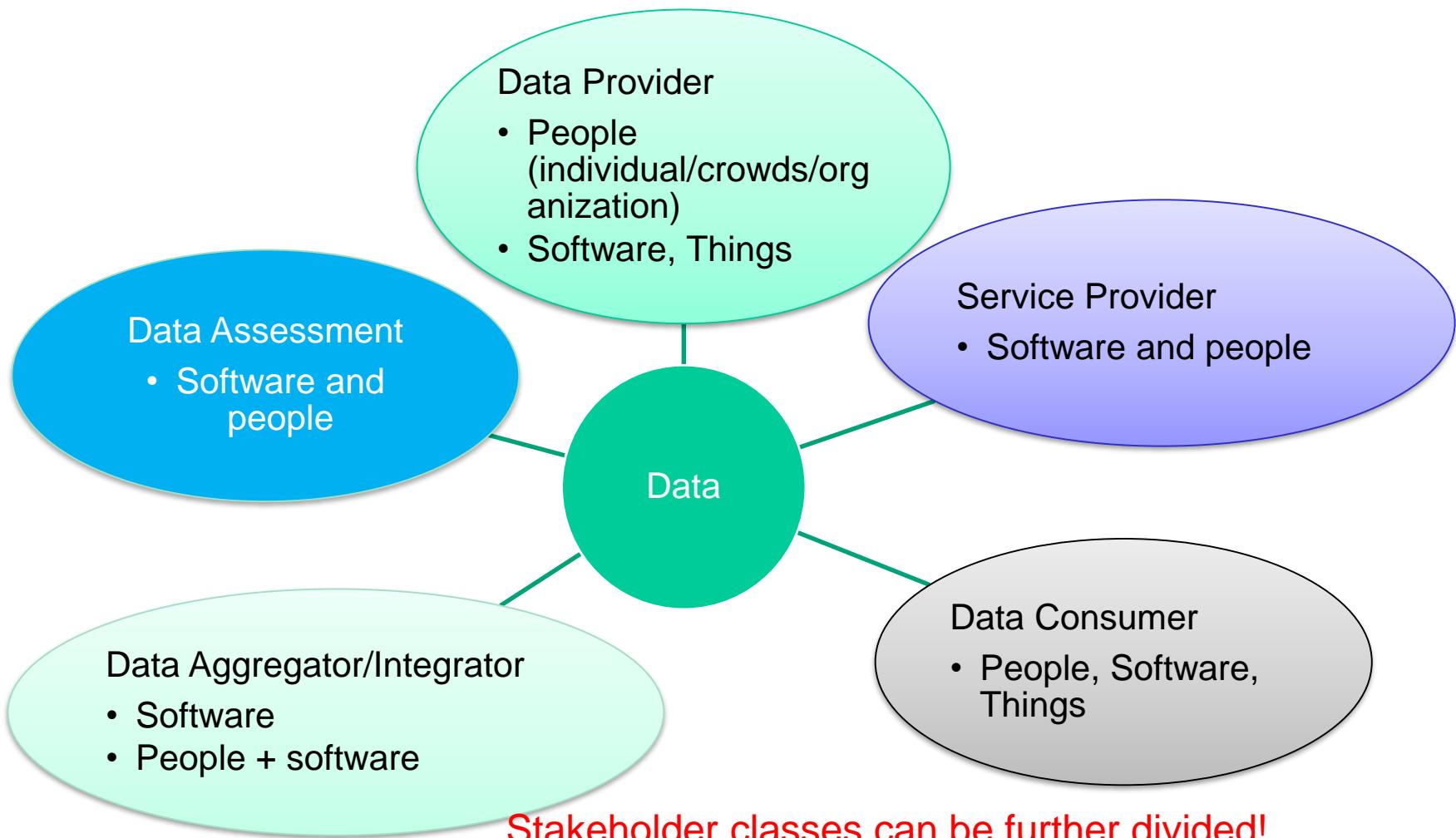


# Data provisioning activities and issues



Non-exhaustive list! Add your own issues!

# Stakeholders in data provisioning

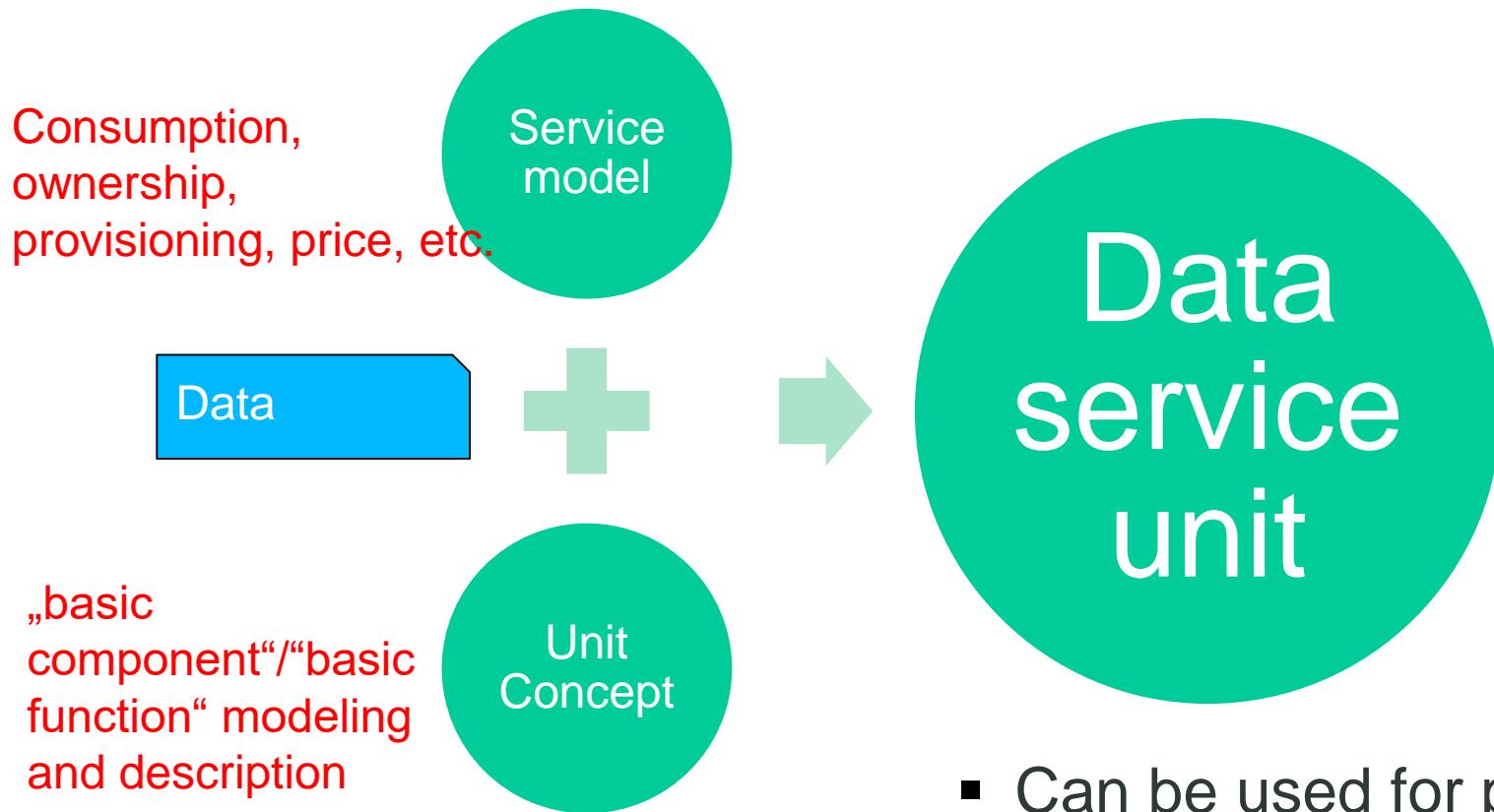


Stakeholder classes can be further divided!  
Domain-specific versus domain-independent functions

How do you see the data sharing and data provisioning situations in China?

# Conceptual data service unit

Microservices mindset!

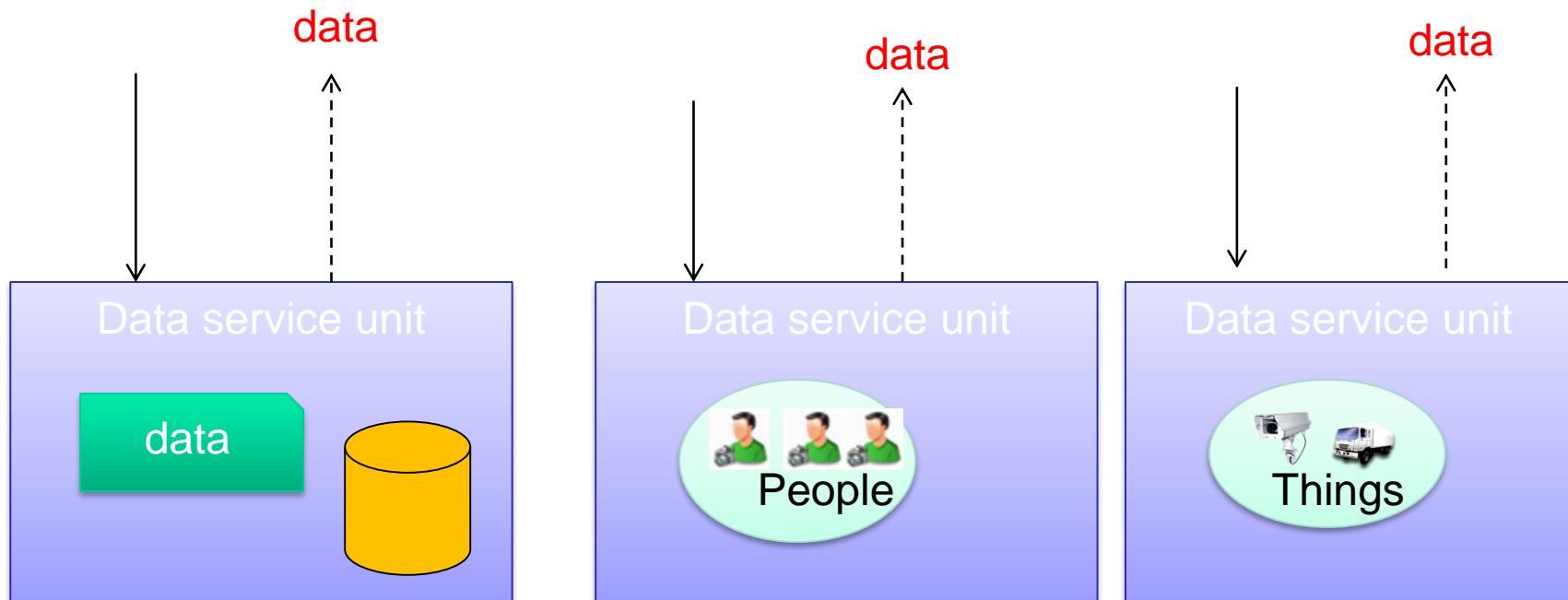


- Can be used for private or public
- Can be elastic or not

# Data service units in clouds

- *Provide data capabilities* rather than provide computation or software capabilities
- Providing data in edges/clouds is an increasing trend
  - In both business and e-science environments
- Now often in a combination of **data + analytics of the data + AI → to provide data assets**

# Data service units in distributed edge and cloud systems



# Data as a Service -- characteristics

Let us use NIST's definition

- *On-demand self-service*
  - Capabilities to provision data at different granularities
- *Resource pooling*
  - Multiple types of data, big, static or near-realtime, raw data and high-level information
- *Broad network access*
  - Can be accessed from anywhere
- *Rapid elasticity*
  - Easy to add/remove data sources
- *Measured service*
  - Measuring, monitoring and publishing data concerns and usage

# Data as a Service – service models and deployment models

## Data-as-a-Service – service models

Data publish/subscription  
middleware as a service

Database-as-a-Service  
(Structured/non-structured  
querying systems)

Sensor-as-a-Service

Storage-as-a-Service  
(Basic storage functions)



IoT, Edge & Cloud Systems

# Examples of DaaS

**Windows Azure Marketplace**

Region: United States ▾ | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA

category

- BANKING (2)
- CAPITAL MARKETS (6)
- INSURANCE (1)

Sort By: Date Added Name Publisher

1 2 3 4 5 ►

**D&B** Decide with Confidence

**Bustling Manufacturers & Business Services List** data  
published by: DNB

Bustling Manufacturers & Business Services list is a market segmentation that covers over 30,000 large and medium-sized businesses with an average annual sales volume of \$40 million. The companies in this list also have high trade activity, maintained steady size in last 4 years and have been in business for an average of 20 years.

**Crime Statistics for England & Wales** data  
published by: Custom Web Apps, Ltd

The crime data is released by the National Policing Improvement Agency (NPfIA) at the end of every month and contains all recorded crime and anti-social behaviour for England & Wales. Data is available from Dec 2010 to present to a level of UK postcode as well as postcode sector, postcode district, and postcode area.

**DIRECTORY SERVICES** data  
Searchable directory of objects and permissions

**DATA SERVICES** data  
Time-Series Archiving

**BUSINESS SERVICES** data  
Device provisioning, activation and management

**MESSAGE BUS** data  
Real-time message management and routing

**Xively™ API** REST, Sockets, MQTT

**Xively™ Applications**  
Developer Workbench  
Device Management Console

**Customer Backend Services**  
Customer CRM Service

**Applications**

**Connected Objects**

**ASE Fudan FIST, Summer 2018**

**GNIP** The Social Media API™ Product

Gnip is the Largest Provider of Social Media Data to the Enterprise - Never Miss a Tweet, Post, Comment or Like

Try Gnip! CONTACT US TODAY

Twitter Feeds GET STARTED!

**DATA.GOV.UK** Beta Opening up Government

Home Data Participate Apps Location Linked Data Library Lab About

Search | Map Search | Publishers | Tags | Public Roles & Salaries | Spend Browser | Spend Reports

**Search Datasets**  
8729 Datasets

Search... Search

**Tags**  
View all tags »  
national-indicators: Health health Spending Data care  
spend-transactions communities school NERC\_DDC  
local-government transparency nhs children  
health-well-being-and-care population finance child  
health-and-social-care education disclosure

**Publishers**  
View all publishers »

- Office for National Statistics (847)
- Department for Communities and Local Government (739)
- NHS Information Centre for Health and Social Care (514)
- British Geological Survey (364)
- Centre for Ecology & Hydrology (326)
- Department for Environment, Food and Rural Affairs (322)
- Welsh Government (241)
- Department of Health (239)
- Department for Children, Schools and Families (227)
- Home Office (221)

**UK Location**  
Conduct Map Based Search »

The UK Location Programme has introduced over 1000 location data records into data.gov.uk and tools to support their use. To find which of these datasets cover a particular location, you can use Map Based Search.

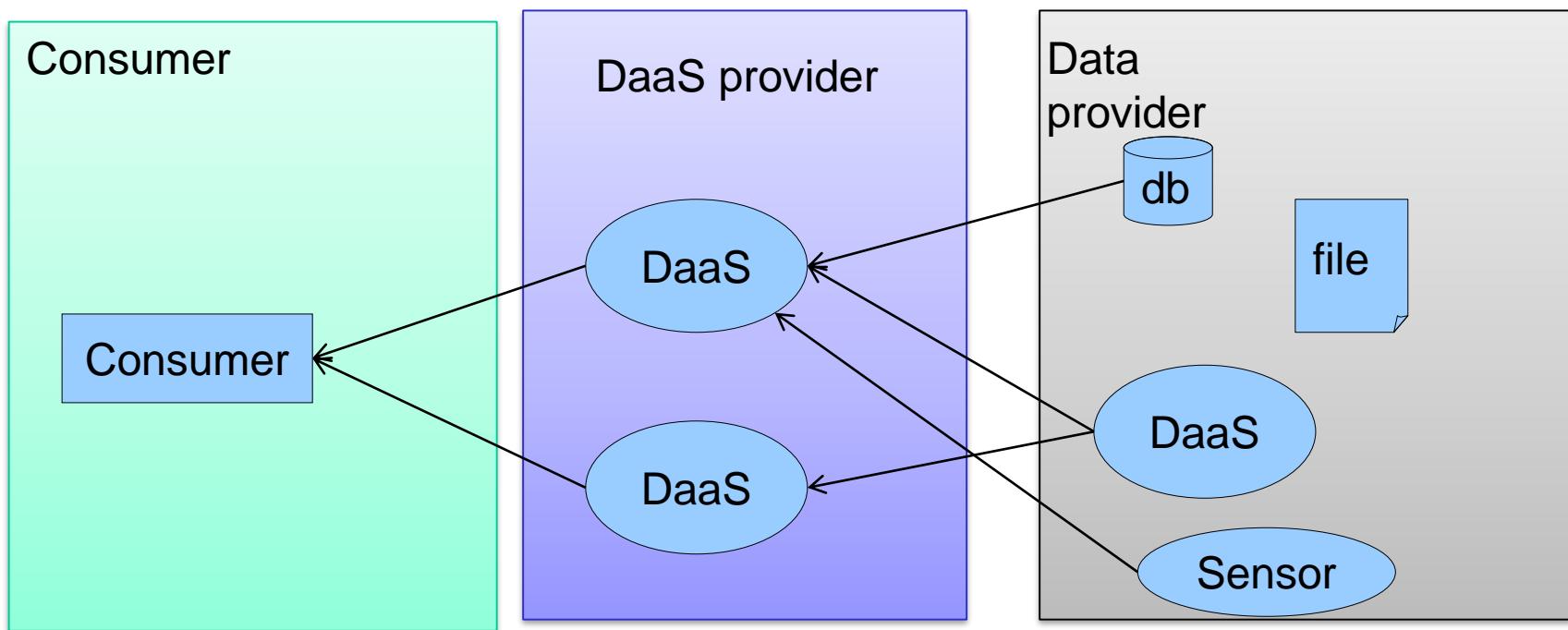
Many of these datasets provide a Web Map Service too, and for some preview of the data is available. Click to find out more about Map Based Search and about Preview on Map.

# DaaS design & implementation – APIs

- Read-only DaaS versus CRUD DaaS APIs
- Service APIs versus Data APIs
  - They are not the same w.r.t. data/service concerns
- REST & Streaming data APIs

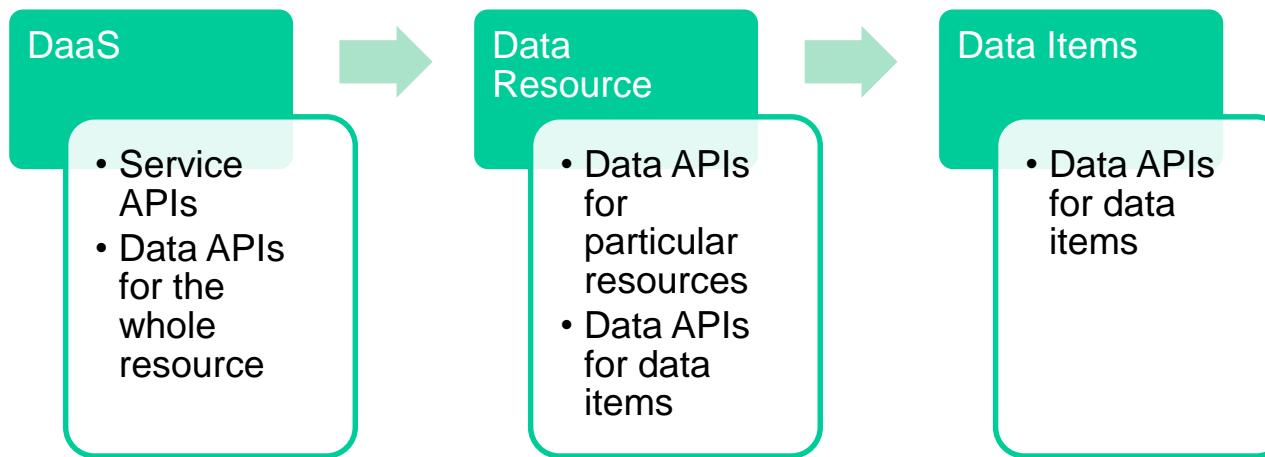
# DaaS design & implementation – service provider vs data provider

- The DaaS provider is separated from the data provider



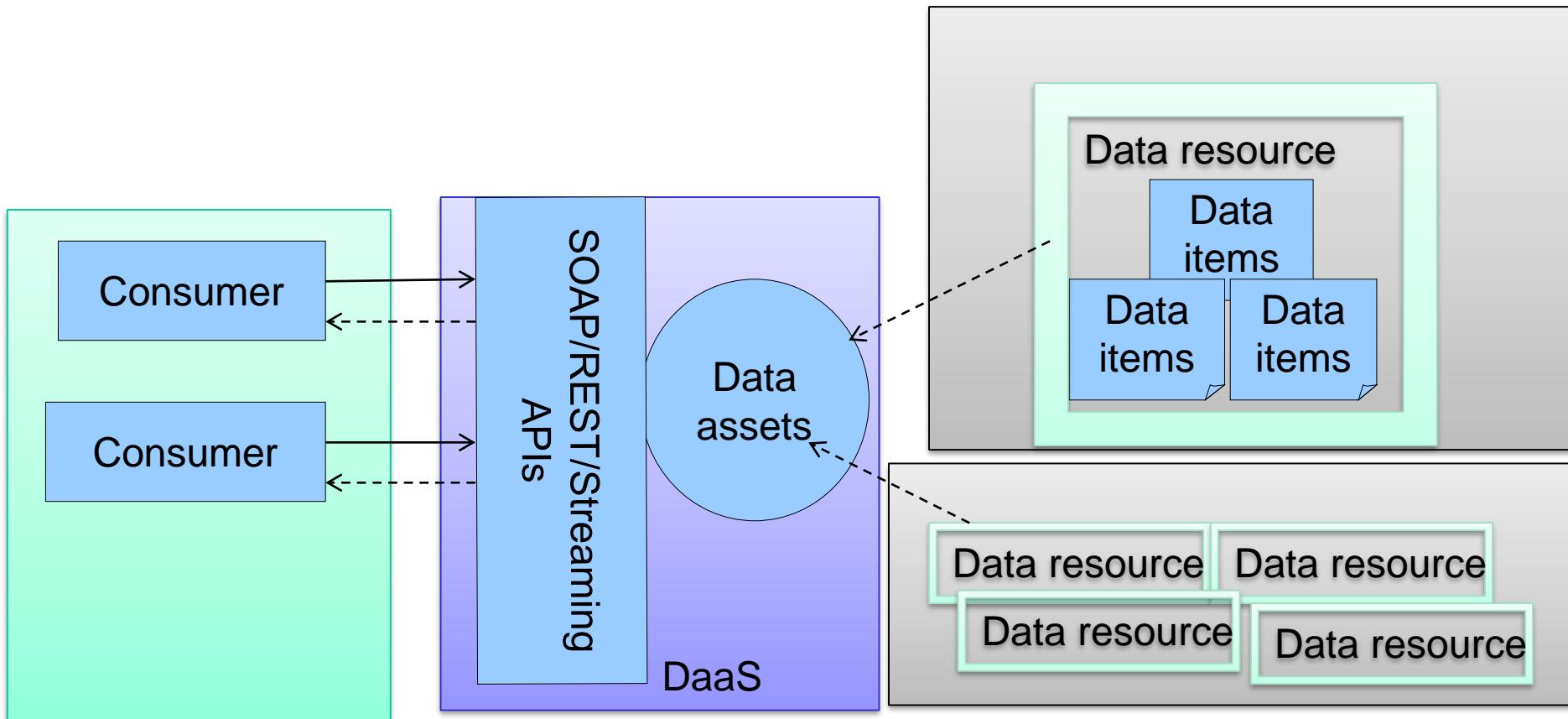
# DaaS design & implementation – structures

Three levels

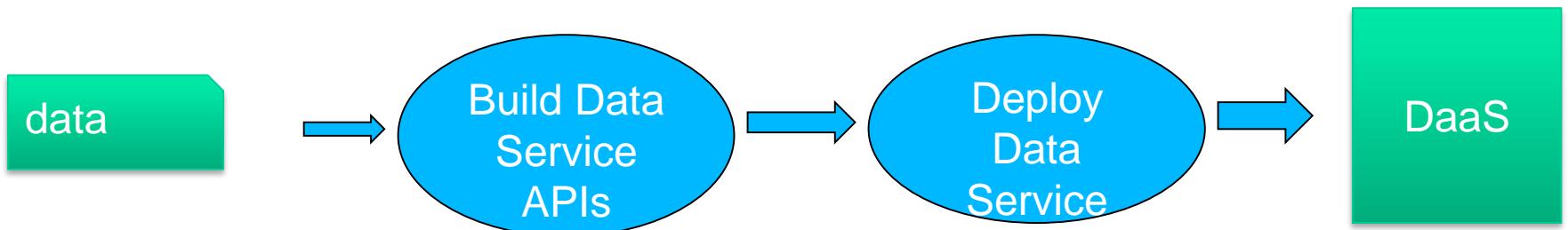


- DaaS and data providers *have the right to publish* the data

# DaaS design & implementation – structures (2)



# DaaS design & implementation – patterns for „turning data to DaaS“ (1)



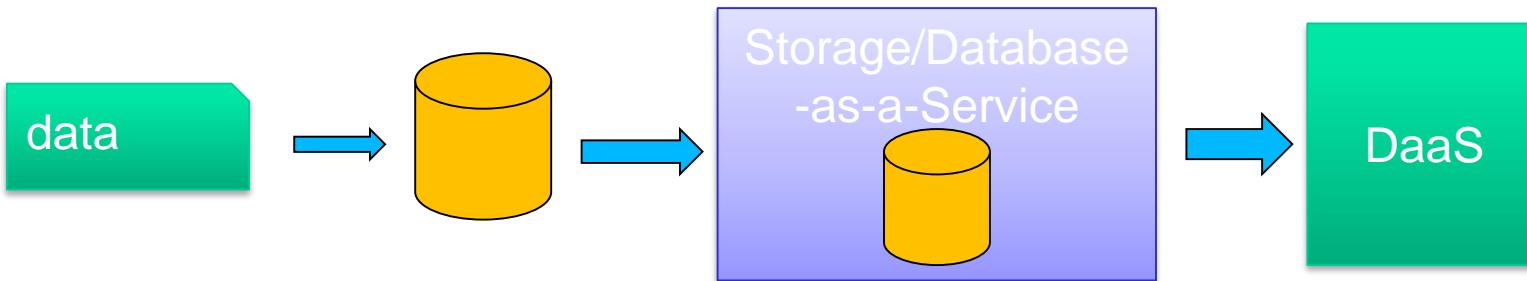
Examples: using WSO2 data service

The screenshot shows the WSO2 Data Services Manager interface. On the left, the 'Edit Query' panel is displayed, containing fields for 'Query ID' (getAvailabilityAll), 'Data Source' (QWSDataSet), 'Result (Output Mapping)' (Grouped by element: serviceAvailability, Row name: service, Row namespace: http://www.infosys.tuwien.ac.at/SOD1/d), and a table for mapping elements to columns. The table has three rows:

Element Name	SQL Column Name	Mapping Type	Allowed User Roles	Schema Type	Actions
availability	availability	element	everyone	xs:double	Edit  Delete
serviceName	ServiceName	element	everyone	xs:string	Edit  Delete

At the bottom of the 'Edit Query' panel are 'Save' and 'Cancel' buttons. On the right, the 'Edit Operation(getAllServiceAvailability)' panel is shown, with fields for 'Operation Name' (getAllServiceAvailability) and 'Query ID' (getAvailabilityAll). It also has 'Save' and 'Cancel' buttons.

# DaaS design & implementation – patterns for „turning data to DaaS“ (2)

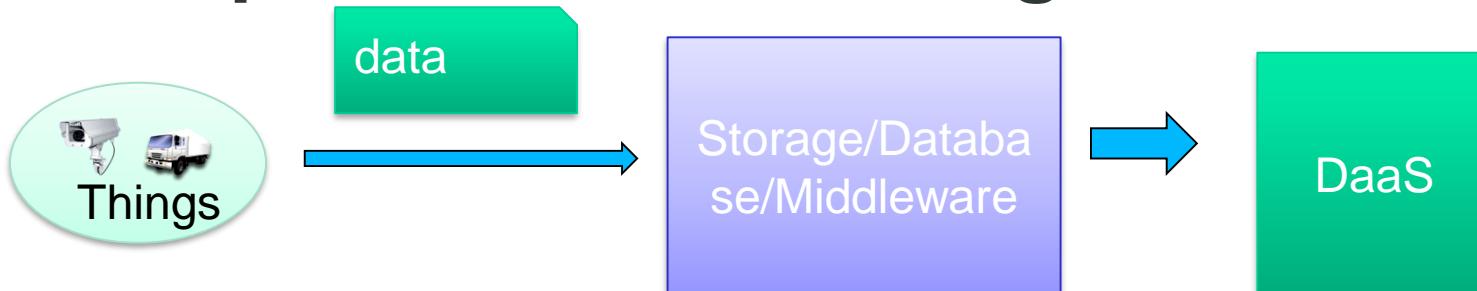


Examples: using Amazon S3

The screenshot shows the AWS S3 console interface. At the top, there's a navigation bar with links like Elastic Beanstalk, S3, EC2, VPC, CloudWatch, etc. Below the navigation bar, there's a sidebar with "Buckets" and a "Create Bucket" button. The main area shows a list of objects under the "smad" bucket. The list includes numerous files with names starting with "ASA\_GM1\_1PNPDE" followed by various identifiers and file extensions (.N1). The columns in the table are "Name", "Size", and "Last Modified". The "Name" column lists the full file paths, the "Size" column shows file sizes ranging from 4.6 MB to 27 MB, and the "Last Modified" column shows dates from June 24, 2011, to June 25, 2011.

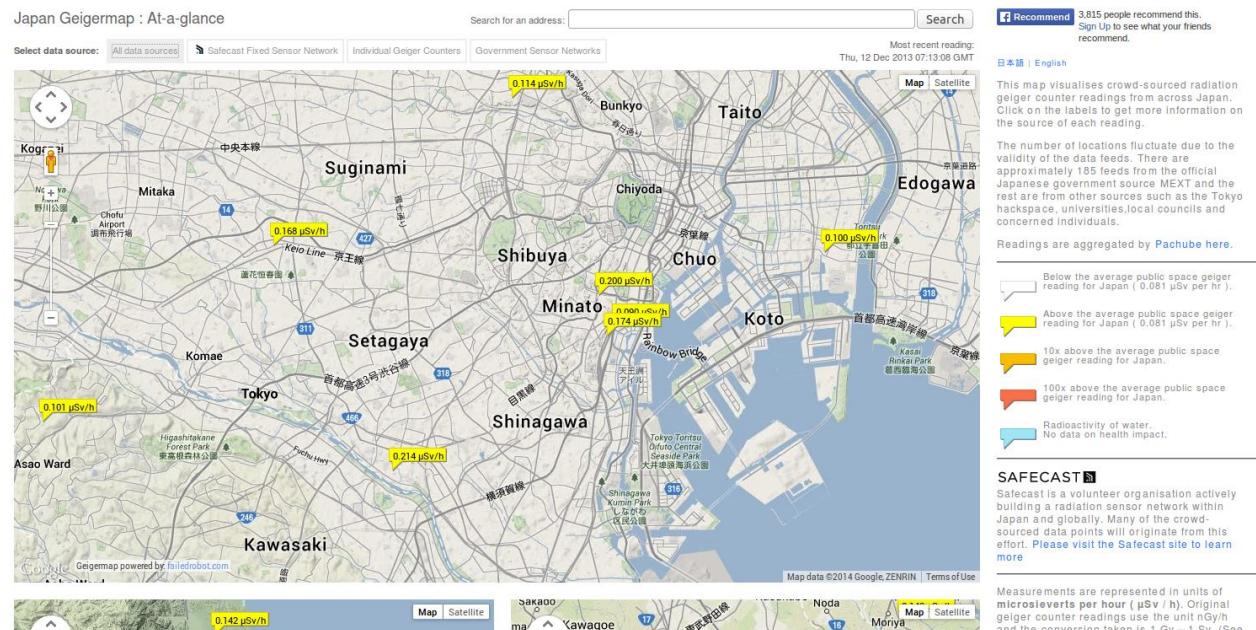
Name	Size	Last Modified
ASA_GM1_1PNPDE20050225_213424_000005862035_00058_15643_0353.N1	12.9 MB	Fri Jun 24 13:32:17 GMT+200 2011
ASA_GM1_1PNPDE20050228_213954_000005862035_00101_15686_0705.N1	12.9 MB	Fri Jun 24 13:32:59 GMT+200 2011
ASA_GM1_1PNPDE20050316_213820_000005012035_00330_15915_2268.N1	11 MB	Fri Jun 24 13:32:59 GMT+200 2011
ASA_GM1_1PNPDE20050401_213800_000001502036_00058_16144_4008.N1	3.2 MB	Fri Jun 24 13:33:20 GMT+200 2011
ASA_GM1_1PNPDE20050405_210811_000005192036_00115_16201_4433.N1	11.4 MB	Fri Jun 24 13:34:01 GMT+200 2011
ASA_GM1_1PNPDE20050408_211356_000005012036_00158_16244_4730.N1	11 MB	Fri Jun 24 13:34:18 GMT+200 2011
ASA_GM1_1PNPDE20050411_211942_000003922036_00201_16287_5118.N1	8.5 MB	Fri Jun 24 13:34:43 GMT+200 2011
ASA_GM1_1PNPDE20050417_213511_000002112036_00287_16373_5662.N1	4.6 MB	Fri Jun 24 13:34:50 GMT+200 2011
ASA_GM1_1PNPDE20050420_214112_000003202036_00330_16416_5947.N1	7 MB	Fri Jun 24 13:34:59 GMT+200 2011
ASA_GM1_1PNPDE20050427_210649_000011052036_00429_16515_6487.N1	27 MB	Fri Jun 24 13:35:11 GMT+200 2011
ASA_GM1_1PNPDE20050430_211232_000010932036_00472_16558_6730.N1	26.7 MB	Fri Jun 24 13:35:49 GMT+200 2011
ASA_GM1_1PNPDE20050503_213221_000002472037_00015_16602_6950.N1	5.3 MB	Fri Jun 24 13:36:32 GMT+200 2011
ASA_GM1_1PNPDE20050522_213525_000002112037_00287_16874_7819.N1	4.6 MB	Fri Jun 24 13:36:41 GMT+200 2011
ASA_GM1_1PNPDE20050525_213129_000009002037_00329_16916_R126.N1	20.6 MB	Fri Jun 24 13:36:50 GMT+200 2011

# DaaS design & implementation – patterns for „turning data to DaaS“ (3)

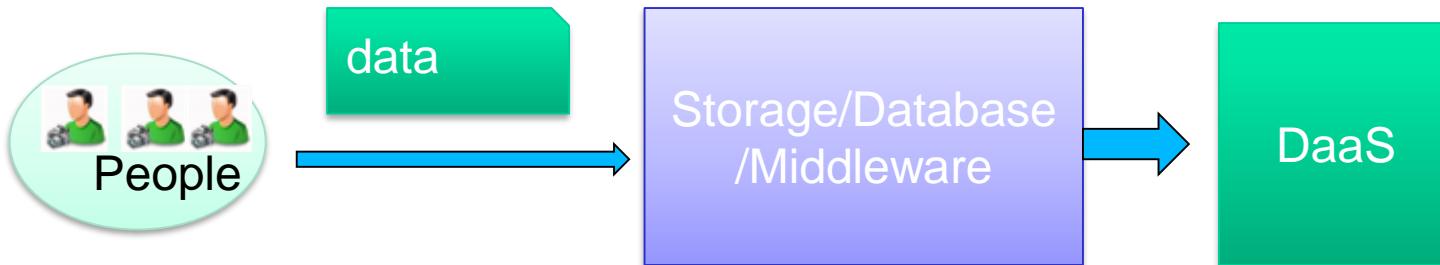


One Thing → 10000... Things

Examples: using Crowd-sourcing with Pachube in 2013 (Note: *the information was not up-to-date*)



# DaaS design & implementation – patterns for „turning data to DaaS“ (4)

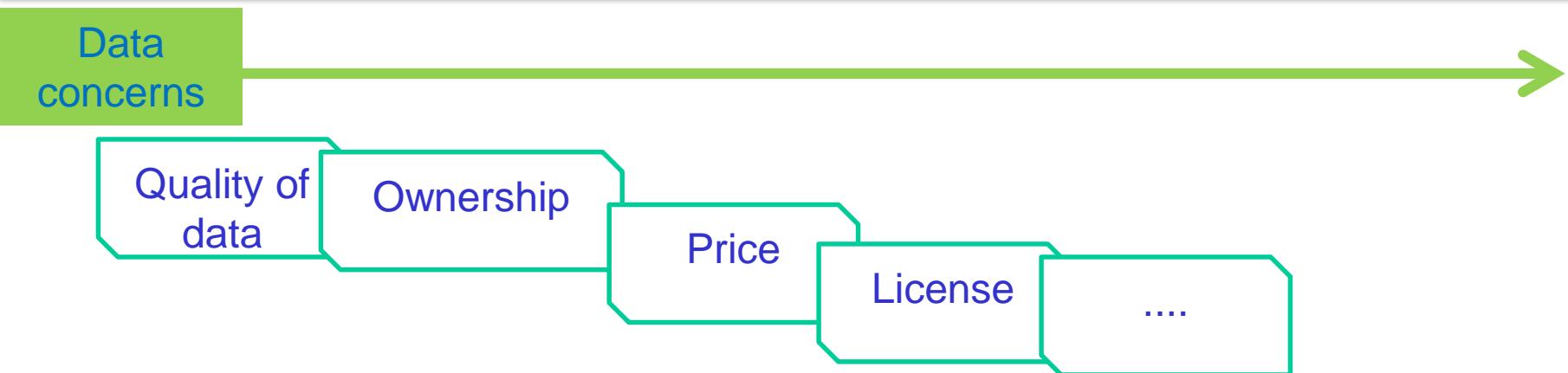
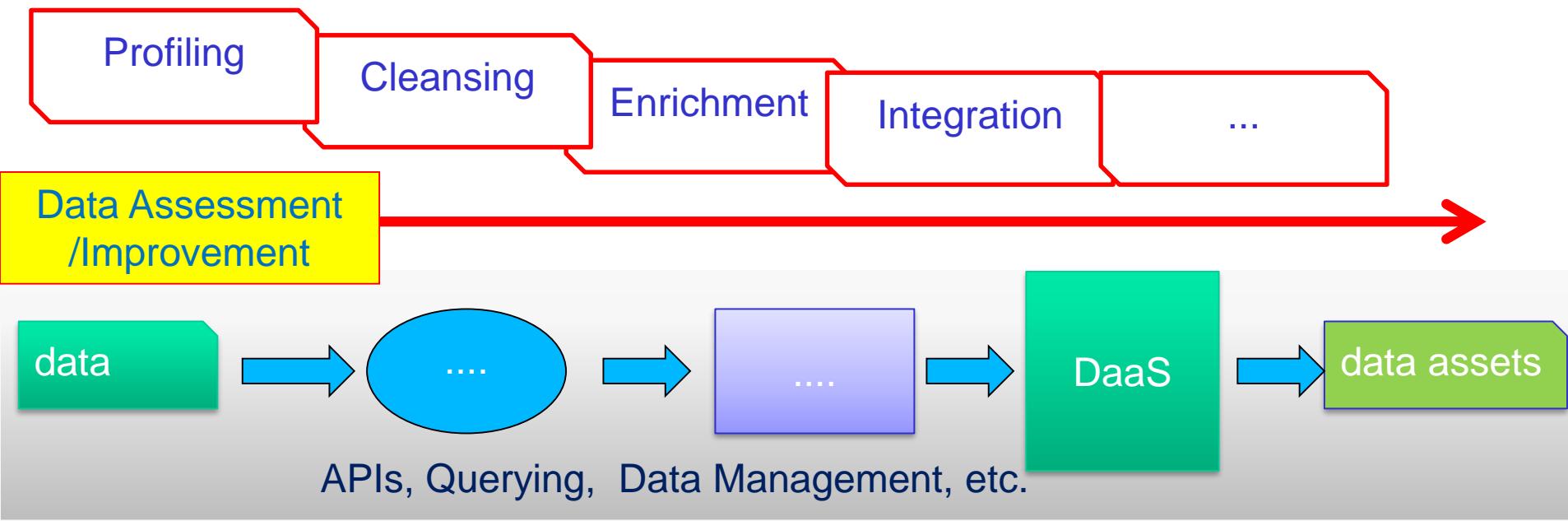


Examples:  
using Twitter

The screenshot shows a portion of the Twitter API documentation. At the top, there's a navigation bar with links for Developer, Use cases, Products, Docs, More, and Apply. Below the navigation, there's a search bar with placeholder text "Search all documentation...". The main content area has a sidebar on the left with sections for Basics, Accounts and users, and Tweets. Under the Tweets section, there's a list of endpoints: Get Tweet timelines, Curate a collection of Tweets, Optimize Tweets with Cards, Search Tweets, Filter realtime Tweets, Sample realtime Tweets, Get batch historical Tweets, Rules and filtering, and Premium enrichments. To the right of the sidebar, the title "Get Tweet timelines" is displayed in large bold letters. Below the title, there are tabs for Overview, Guides, and API Reference, with "Overview" being the active tab. A descriptive paragraph explains what a timeline is, mentioning it's a list or aggregated stream of Tweets. It also notes that the Twitter API has several endpoints for timelines. A table below lists the API endpoints and their descriptions:

API endpoint	Description
GET statuses / home_timeline	Returns a collection of the most recent Tweets posted by the authenticating user and the users they follow.
GET statuses / user_timeline	Returns a collection of the most recent Tweets posted by the indicated by the <code>screen_name</code> or <code>user_id</code> parameters.
GET statuses/mentions_timeline	Returns the 20 most recent mentions (Tweets containing a user's @handle) for the authenticating user.

# DaaS design & implementation – not just „functional“ aspects (1)



# DaaS design & implementation – not just „functional“ aspects (2)

Understand the DaaS ecosystem

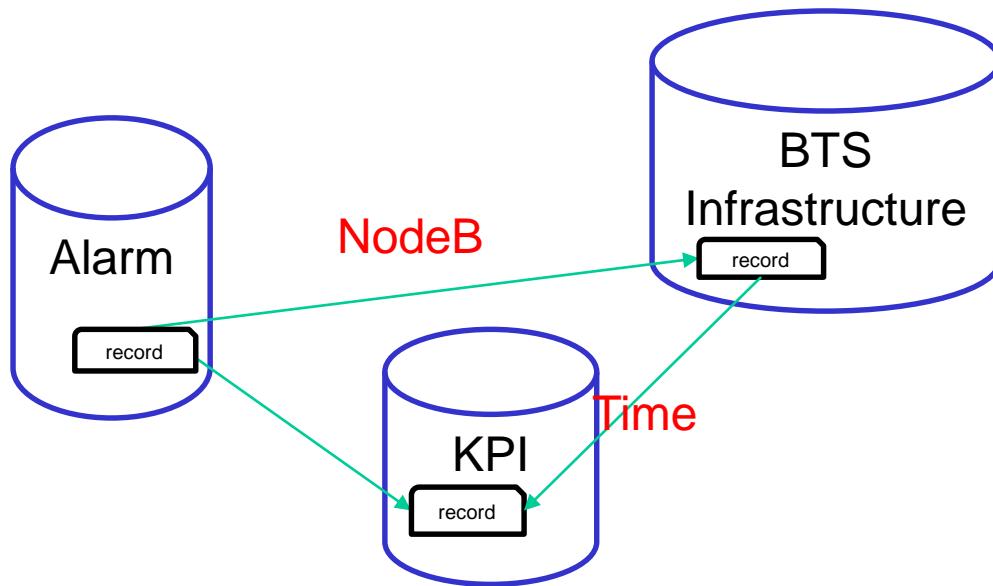
Specifying, Evaluating and Provisioning *Data concerns and Data Contract*

# Example

<https://www.informatica.com/products/data-quality/data-as-a-service.html>

# DATALAKE

# Example: Linking data in telco management



You can continue to have different data sources like that but you need to make sure they are linked

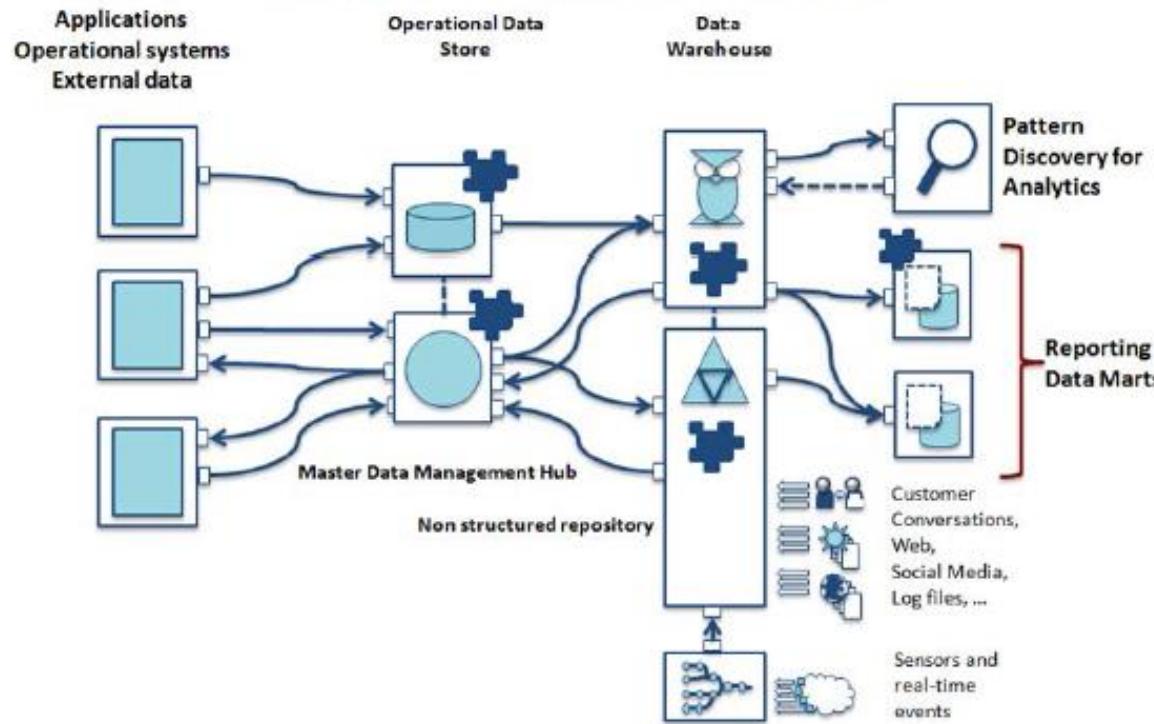
# Data lakes

- A lake of data
    - Ingest and integrate as many as possible types of data
    - To archive a lot of data so that potentially many analytics and applications can access
- Data lake is a concept so you can implement it based on your requirements and needs

# Example

## Existing Decision Support System

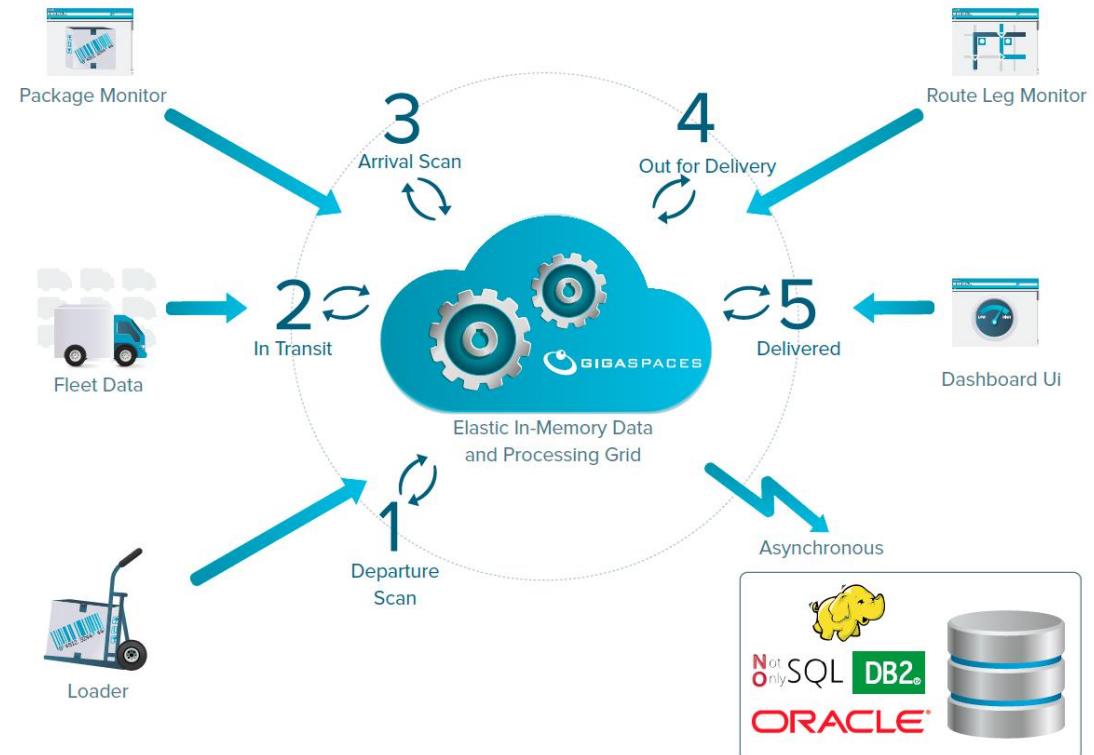
*Through unstructured and new IoT data sources*



Cedrine Madera and Anne Laurent. 2016. The next information architecture evolution: the data lake wave. In Proceedings of the 8th International Conference on Management of Digital EcoSystems (MEDES). ACM, New York, NY, USA, 174-180.  
 DOI: <https://doi.org/10.1145/3012071.3012077>

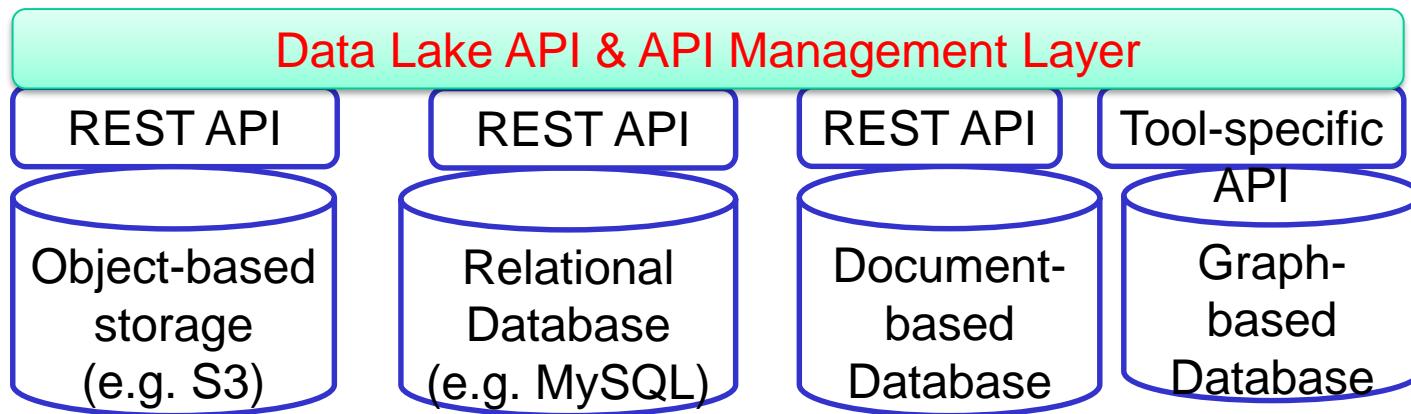
# Implementation

Can we build a data lake using the concept of “data space”



Source: <http://www.gigaspaces.com/logistics-and-shipping-management>

# Data Lake through Data Access API & API Management



Data access APIs can be built based on well-defined interfaces

Help to bring the data object close to the programming language objects

# DATA GOVERNANCE

# Data Governance

**“Data governance is a control that ensures that the data entry by an operations team member or by automated processes meets precise standards, such as a business rule, a data definition and data integrity constraints in the data model.”**

From [https://en.wikipedia.org/wiki/Data\\_governance](https://en.wikipedia.org/wiki/Data_governance)

# Why data governance is very important for ASE?

# Example: General Data Protection Regulation (GDPR)

- <https://www.eugdpr.org/>, effective from May 2018
  - Related to EU citizen data used and managed by services, regardless of where a service is located (e.g., in China)
- Heavy fine if violated
- Consequences:
  - Minimum usage and storage of EU citizen data
  - Clear consent/agreement
  - Major changes in software design and operations
  - Etc.

# Data governance Process

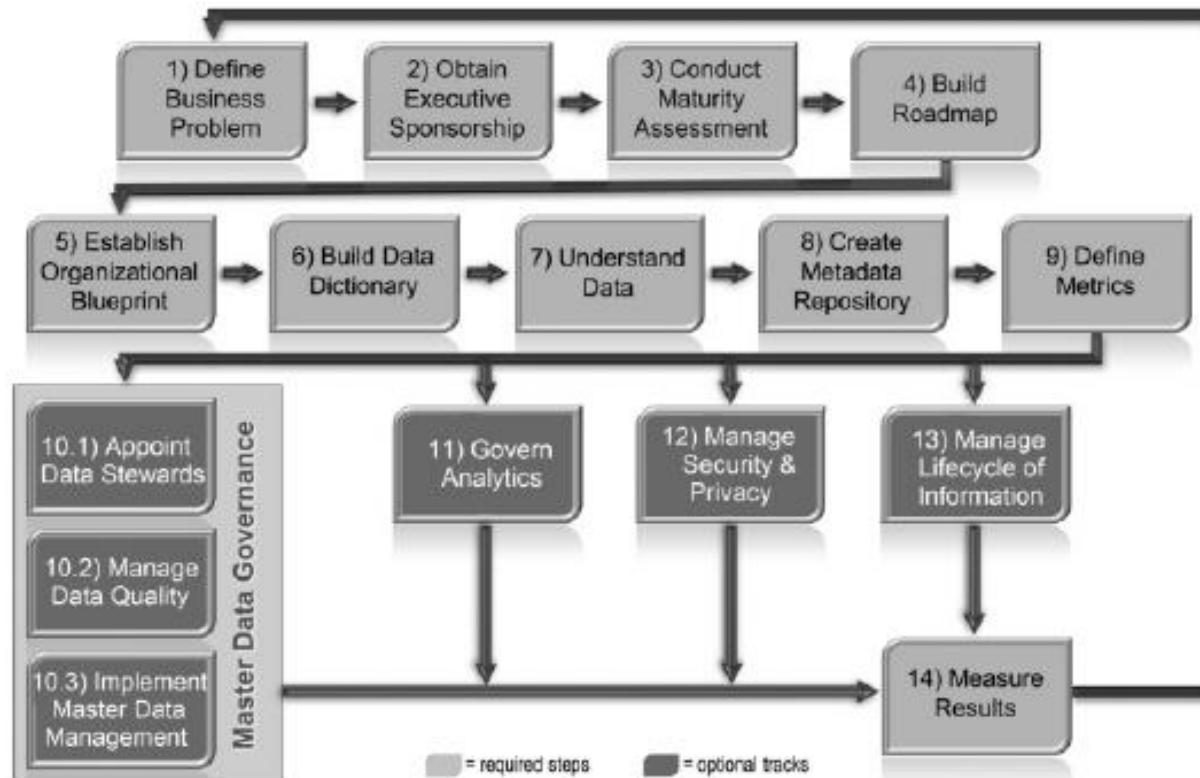


Figure 2.1: An overview of the IBM Data Governance Unified Process.

Sunil Soares. 2010. The IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices. MC Press, LLC.

# Decision domains for data governance

Figure 1: Key organizational assets to be governed; adapted from Weill and Ross.<sup>10</sup>

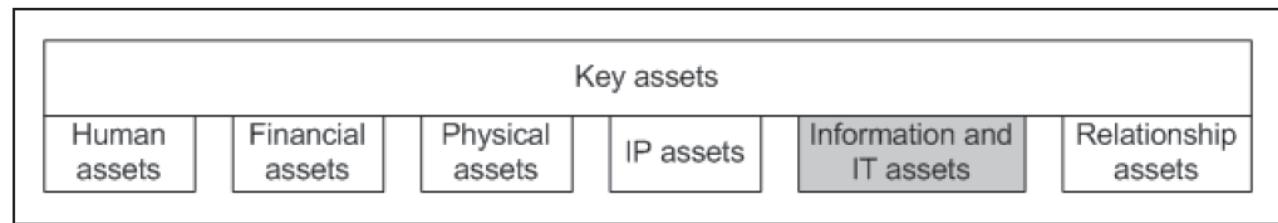
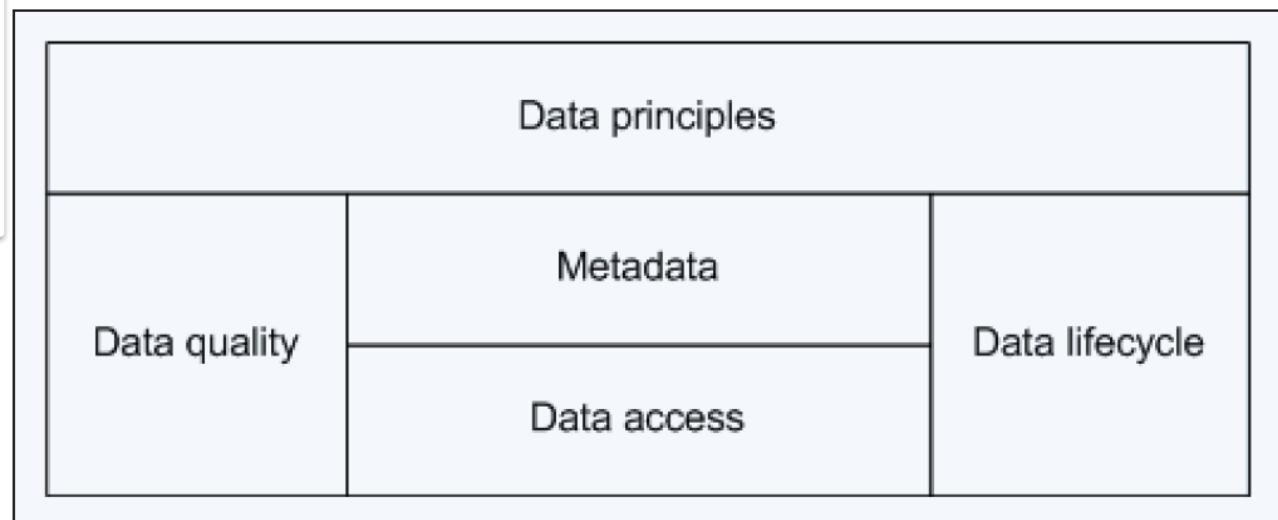


Figure 2: Decision domains for data governance.



Vijay Khatri and Carol V. Brown.  
2010. Designing data governance.  
Commun. ACM 53, 1 (January  
2010), 148-152.  
DOI=<http://dx.doi.org/10.1145/1629175.1629210>

# Framework for domain decisions

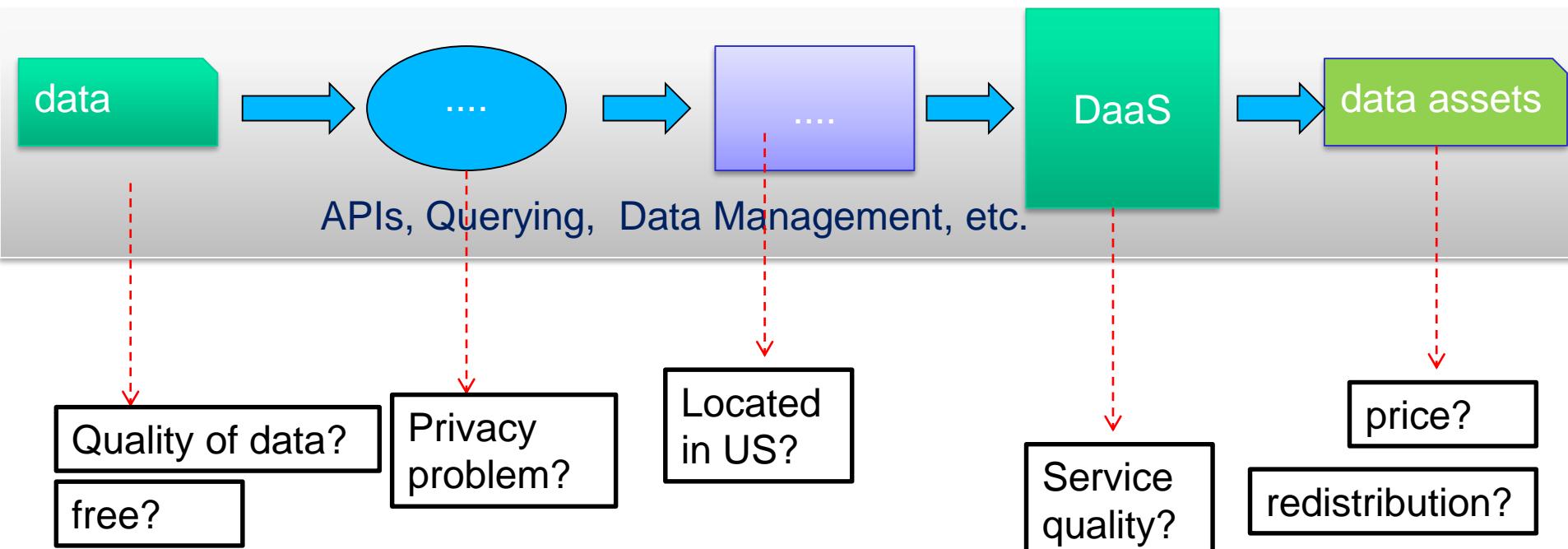
Vijay Khatri and Carol V. Brown.  
 2010. Designing data governance.  
 Commun. ACM 53, 1 (January 2010), 148-152.  
 DOI=<http://dx.doi.org/10.1145/1629175.1629210>

**Table 1: Framework for data decision domains.**

Data Governance Domains	Domain Decisions	Potential Roles or Locus of Accountability
<b>Data Principles</b> • Clarifying the role of data as an asset	<ul style="list-style-type: none"> <li>• What are the uses of data for the business?</li> <li>• What are the mechanisms for communicating business uses of data on an ongoing basis?</li> <li>• What are the desirable behaviors for employing data as assets?</li> <li>• How are opportunities for sharing and reuse of data identified?</li> <li>• How does the regulatory environment influence the business uses of data?</li> </ul>	<ul style="list-style-type: none"> <li>• Data owner/trustee</li> <li>• Data custodian</li> <li>• Data steward</li> <li>• Data producer/supplier</li> <li>• Data consumer</li> <li>• Enterprise Data Committee/Council</li> </ul>
<b>Data Quality</b> • Establishing the requirements of intended use of data	<ul style="list-style-type: none"> <li>• What are the standards for data quality with respect to accuracy, timeliness, completeness and credibility?</li> <li>• What is the program for establishing and communicating data quality?</li> <li>• How will data quality as well as the associated program be evaluated?</li> </ul>	<ul style="list-style-type: none"> <li>• Data owner</li> <li>• Subject matter expert</li> <li>• Data quality manager</li> <li>• Data quality analyst</li> </ul>
<b>Metadata</b> • Establishing the semantics or "content" of data so that it is interpretable by the users	<ul style="list-style-type: none"> <li>• What is the program for documenting the semantics of data?</li> <li>• How will data be consistently defined and modeled so that it is interpretable?</li> <li>• What is the plan to keep different types of metadata up-to-date?</li> </ul>	<ul style="list-style-type: none"> <li>• Enterprise data architect</li> <li>• Enterprise data modeler</li> <li>• Data modeling engineer</li> <li>• Data architect</li> <li>• Enterprise Architecture Committee</li> </ul>
<b>Data Access</b> • Specifying access requirements of data	<ul style="list-style-type: none"> <li>• What is the business value of data?</li> <li>• How will risk assessment be conducted on an ongoing basis?</li> <li>• How will assessment results be integrated with the overall compliance monitoring efforts?</li> <li>• What are data access standards and procedures?</li> <li>• What is the program for periodic monitoring and audit for compliance?</li> <li>• How is security awareness and education disseminated?</li> <li>• What is the program for backup and recovery?</li> </ul>	<ul style="list-style-type: none"> <li>• Data owner</li> <li>• Data beneficiary</li> <li>• Chief information security officer</li> <li>• Data security officer</li> <li>• Technical security analyst</li> <li>• Enterprise Architecture Development Committee</li> </ul>
<b>Data Lifecycle</b> • Determining the definition, production, retention and retirement of data	<ul style="list-style-type: none"> <li>• How is data inventoried?</li> <li>• What is the program for data definition, production, retention, and retirement for different types of data?</li> <li>• How do the compliance issues related to legislation affect data retention and archiving?</li> </ul>	<ul style="list-style-type: none"> <li>• Enterprise data architect</li> <li>• Information chain manager</li> </ul>

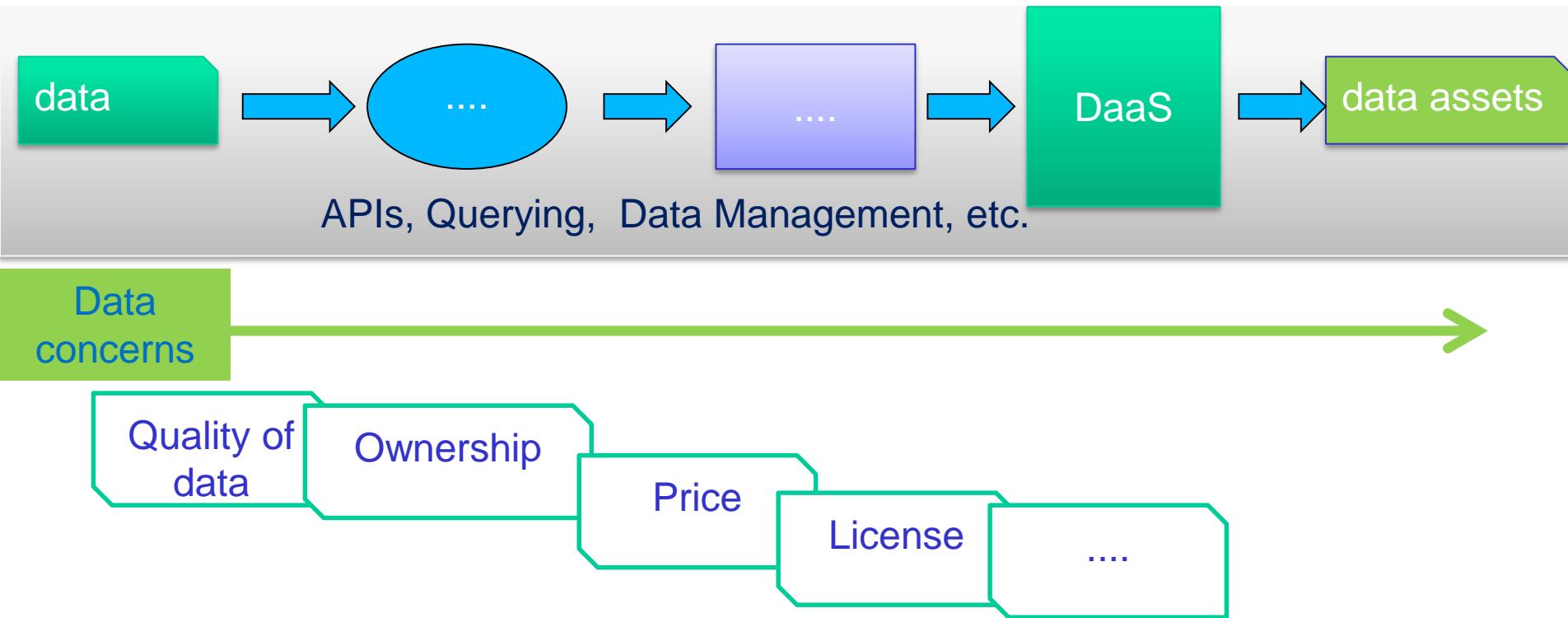
# DATA CONCERNS

# What are data concerns?



Read: Carlo Batini, Monica Scannapieco: Data and Information Quality - Dimensions, Principles and Techniques. Data-Centric Systems and Applications, Springer 2016, ISBN 978-3-319-24104-3, pp. 1-449

# DaaS concerns



DaaS concerns include QoS, quality of data (QoD), service licensing, data licensing, data governance, etc.

# Why DaaS/data concerns are important?

- Overloading data returned to the consumer/integrator are not good
- Results are returned without a clear usage and ownership causing data compliance problems
- Consumers want to deal with dynamic changes

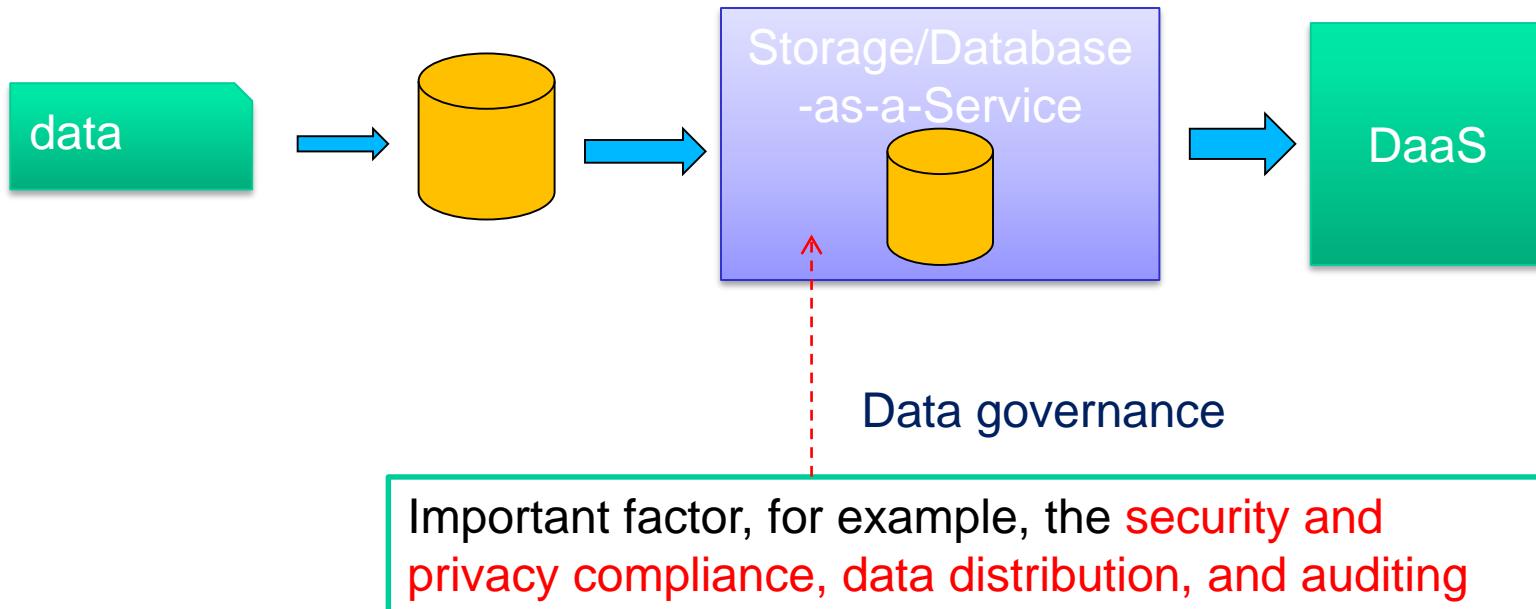
Ultimate goal: to provide *relevant* data with *acceptable constraints on data concerns in different provisioning models*

# DaaS concerns analysis and specification

- Which concerns are important in which situations?
- How to specify concerns?

Hong Linh Truong, Schahram Dustdar On analyzing and specifying concerns for data as a service. APSCC 2009: 87-94

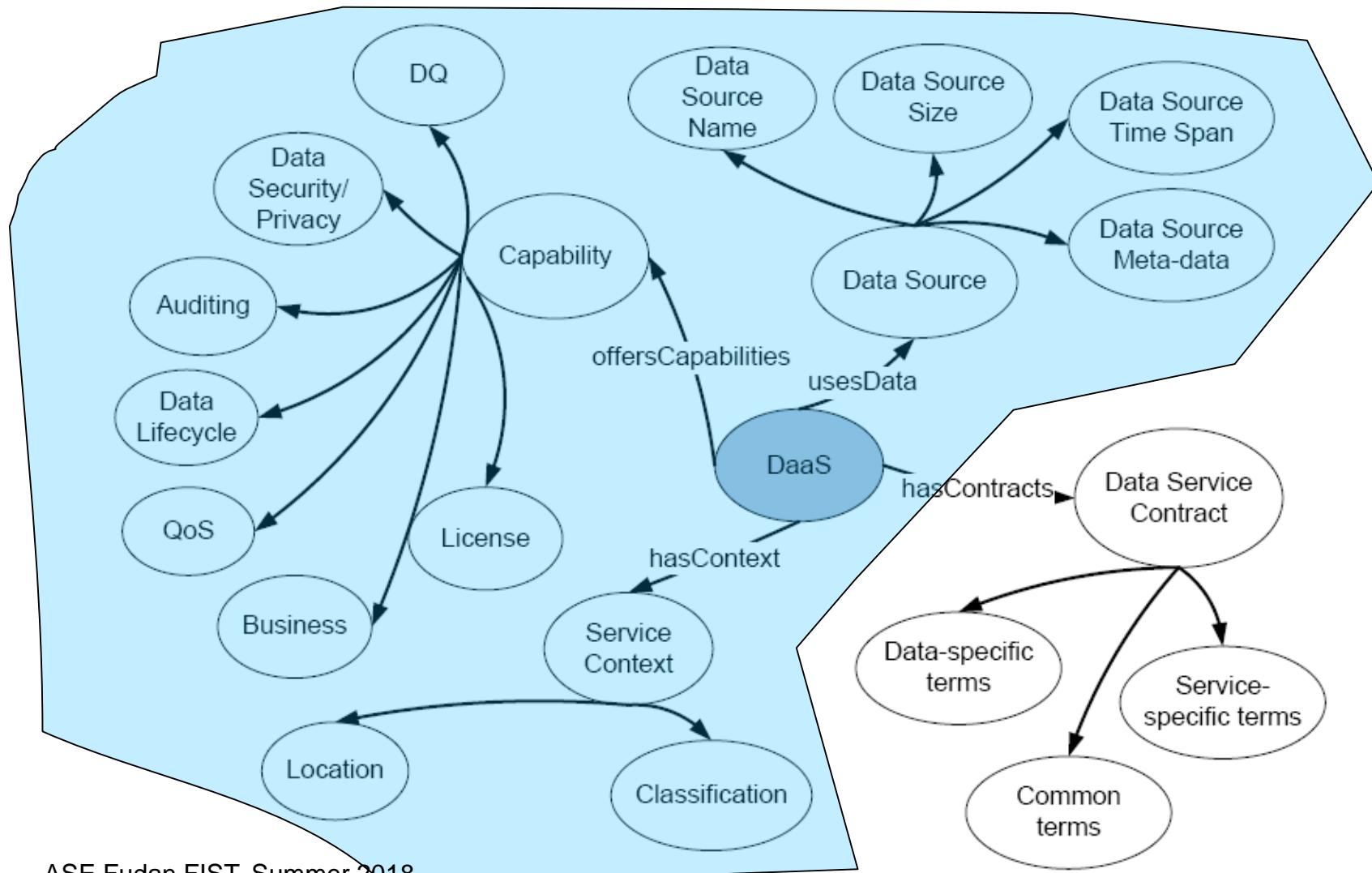
# Data governance



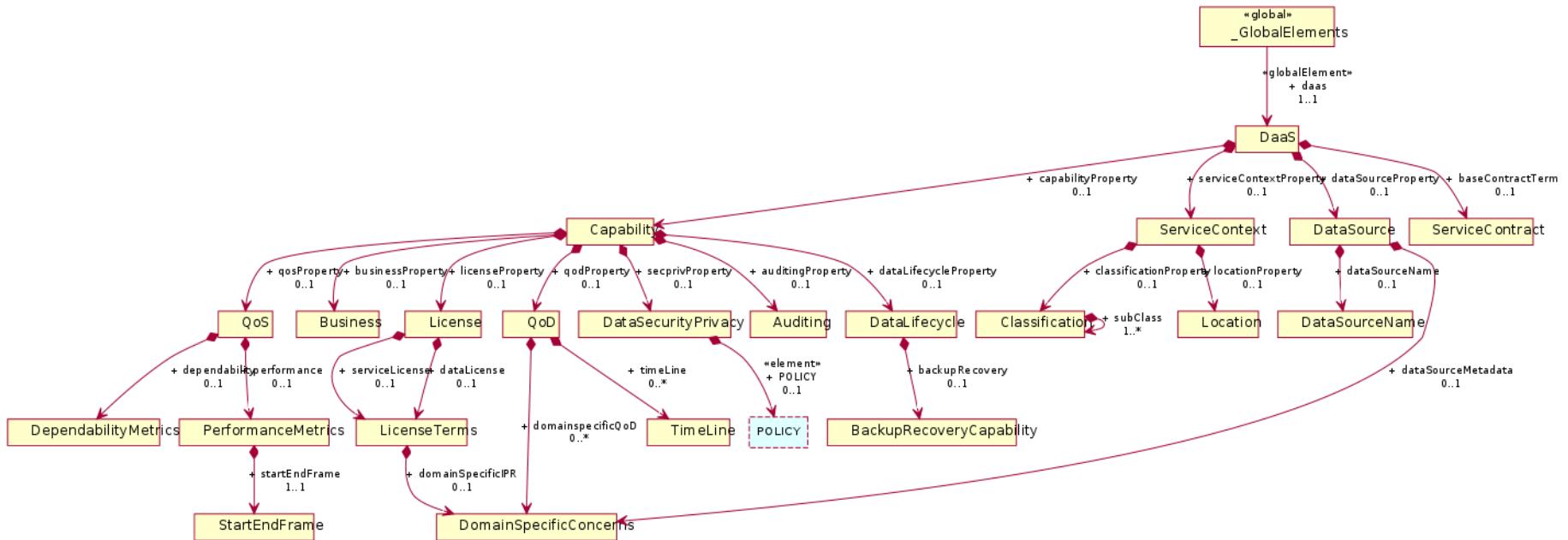
# Types of concerns

- Quality of data:
  - For example, the **accuracy** and **completeness** of the data, whether the data is **up-to-date**
- Data and service usage:
  - In particular, **price**, data and service **APIs** **licensing**, **law enforcement**, and **Intellectual Property** rights
- Quality of service:
  - in particular, **availability**, **response time**, **dependability**, and **security**
- Context:
  - useful factor, such as **classification** and **service type** (REST), **location**

# Conceptual model for DaaS concerns and contracts



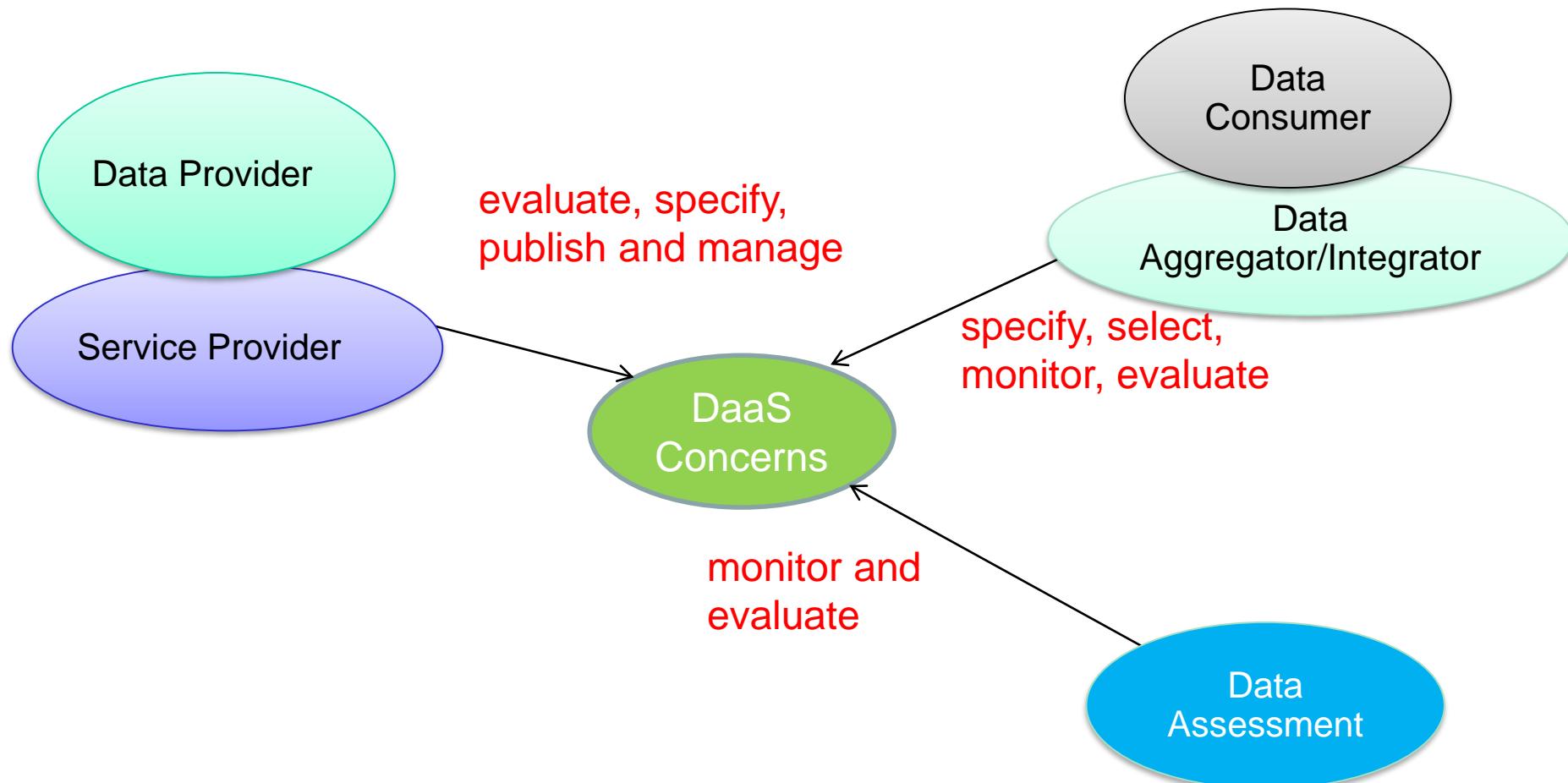
# Example of implementations



Check <http://www.infosys.tuwien.ac.at/prototyp/SOD1/dataconcerns>

# Populating DaaS concerns

The role of stakeholders in the most trivial view

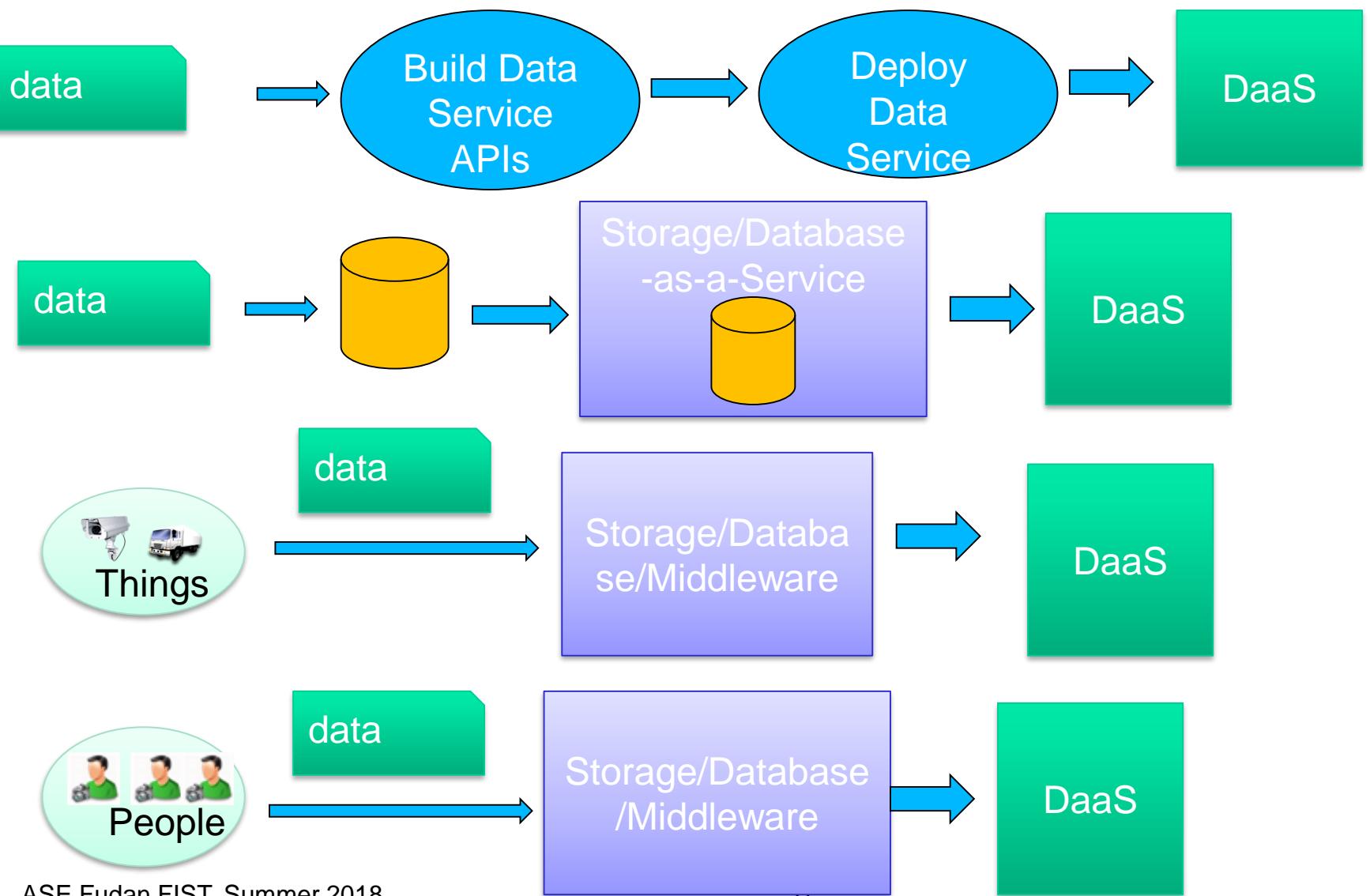


# **SOFTWARE DESIGN FOR EVALUATING DATA CONCENRS FOR DATA ASSETS**

# Data concern measure

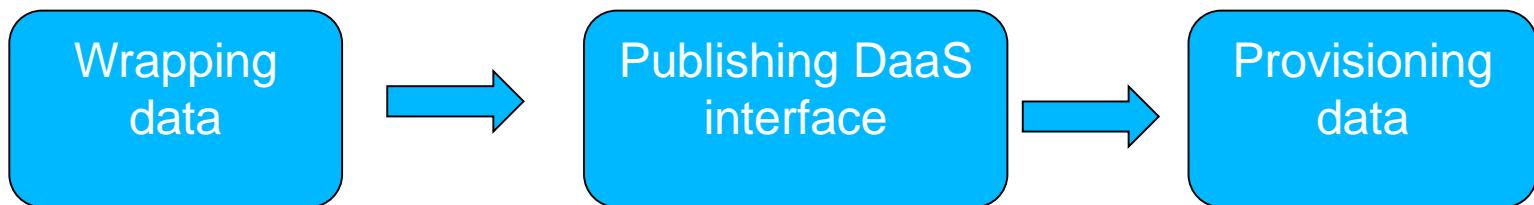
- They are domain-specific and data-specific
  - You need to look at specific definitions in order to calculate/determine values of metrics
- But key principles of software design are quite similar
- We are focusing on software design

# Patterns for „turning data to DaaS“

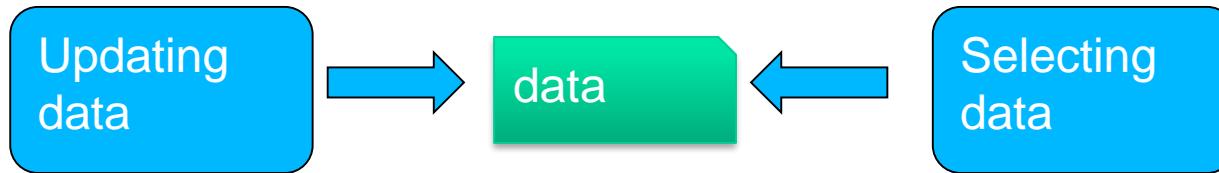


# Data-related activities

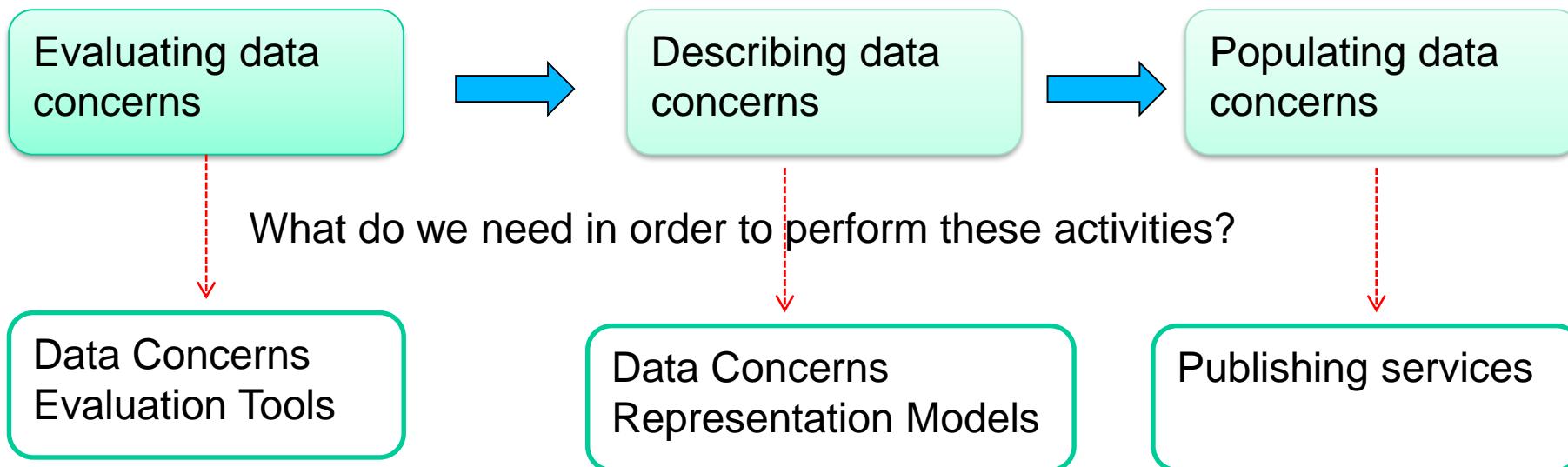
Typical activities for data wrapping and publishing



Typical activities for data updating & retrieval

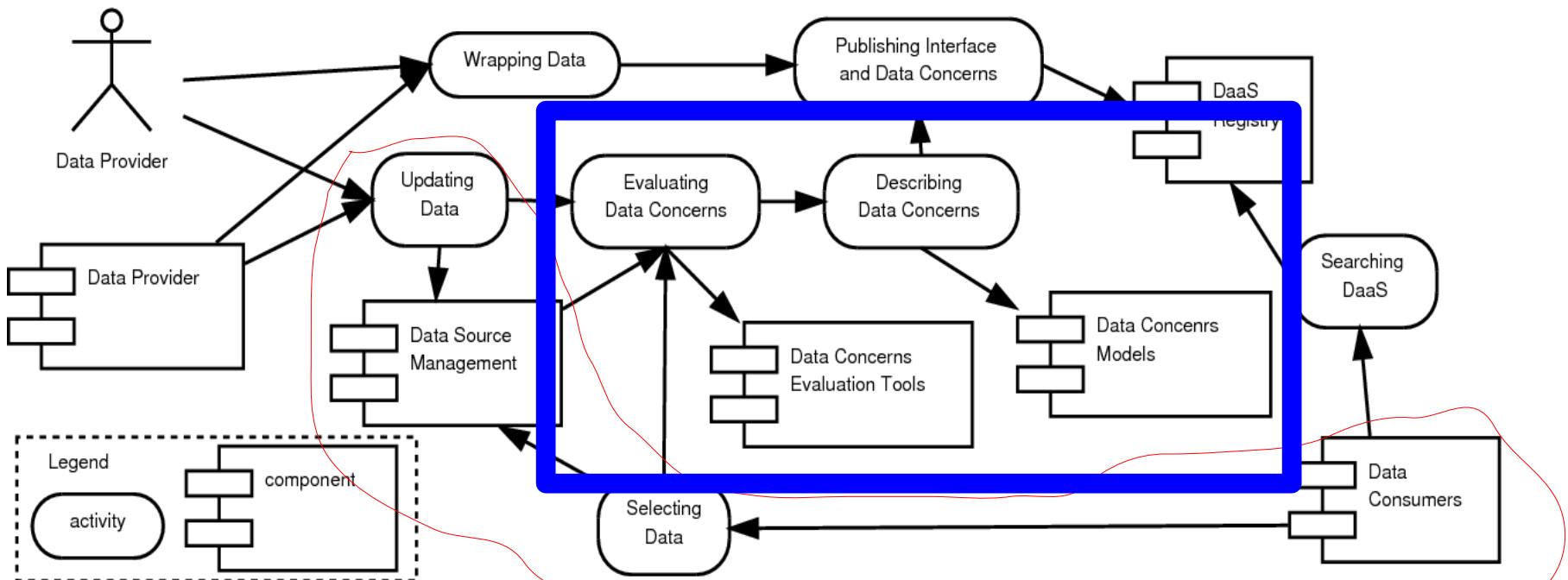


# Typical data concern evaluation



# Data concern-aware DaaS engineering process

## Typical activities for data wrapping and publishing



## Typical activities for data updating & retrieval

Hong Linh Truong, Schahram Dustdar: On Evaluating and Publishing Data Concerns for Data as a Service. APSCC 2010: 363-370

# Evaluating data concerns – the three important points

evaluation scope

- At which level the evaluation is performed?

evaluation modes

- When the evaluation is done?

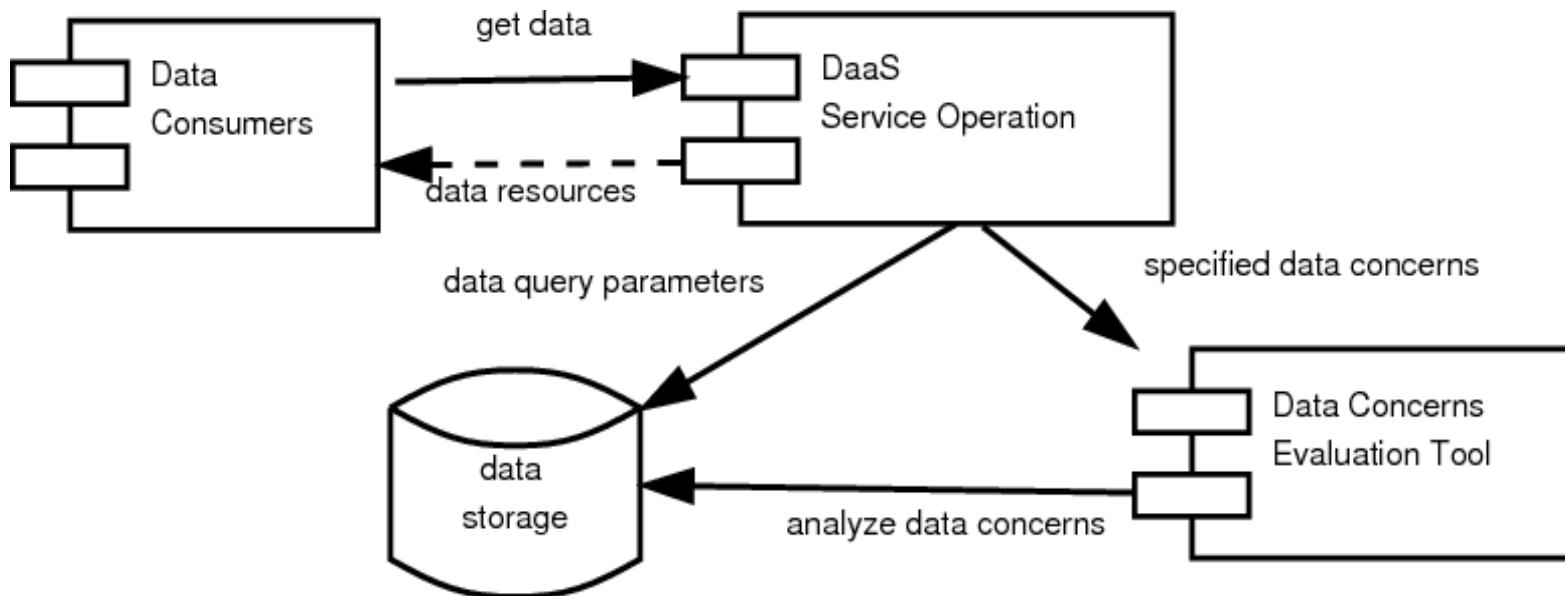
integration model

- How the evaluation tool is invoked?

Hong Linh Truong, Schahram Dustdar: On Evaluating and Publishing Data Concerns for Data as a Service. APSCC 2010: 363-370

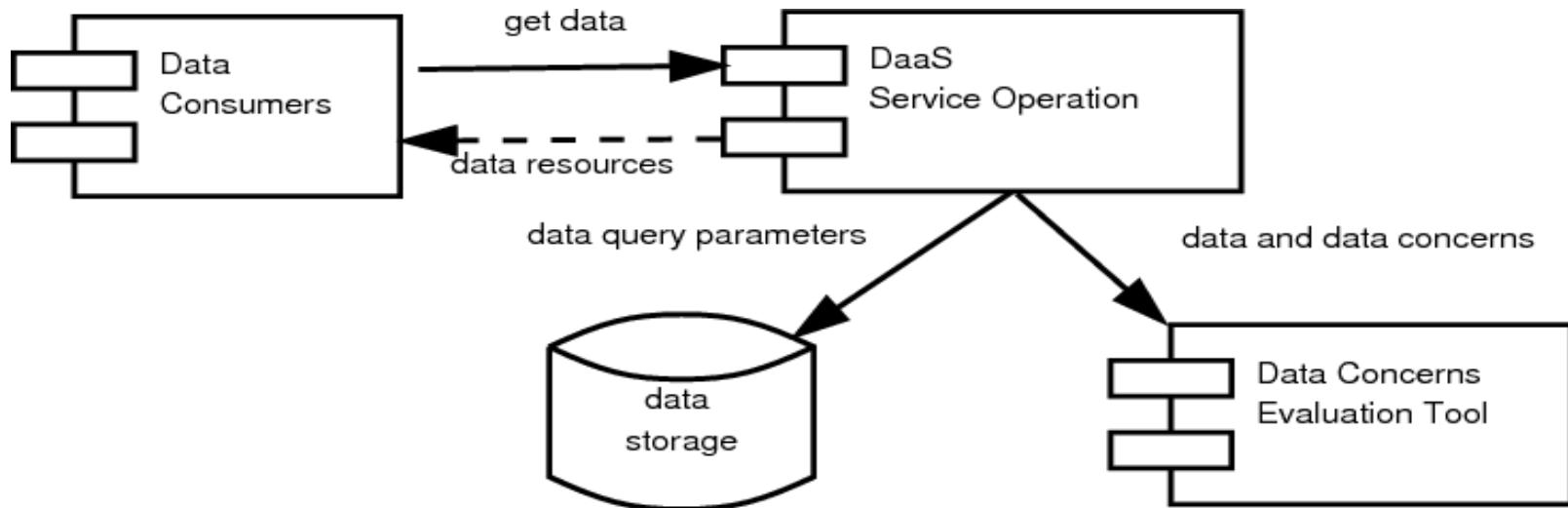
# Evaluating data concerns – some patterns (1)

## Pull, pass-by-references



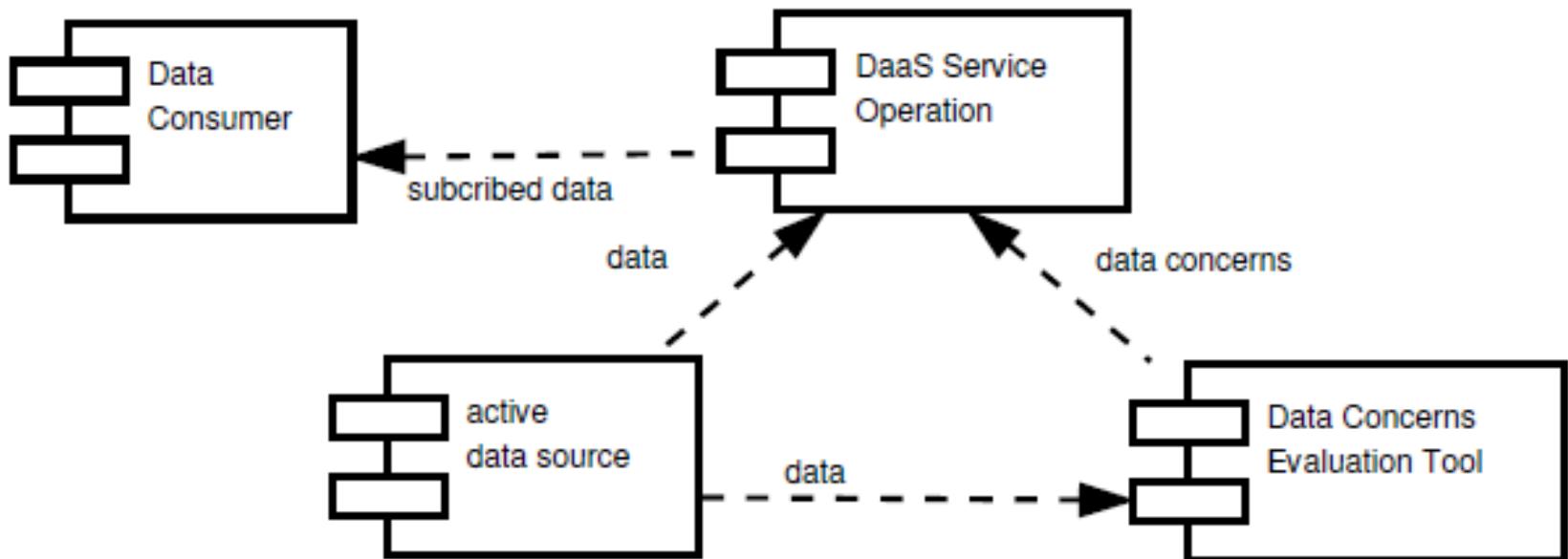
# Evaluating data concerns – some patterns (2)

## Pull, pass-by-values



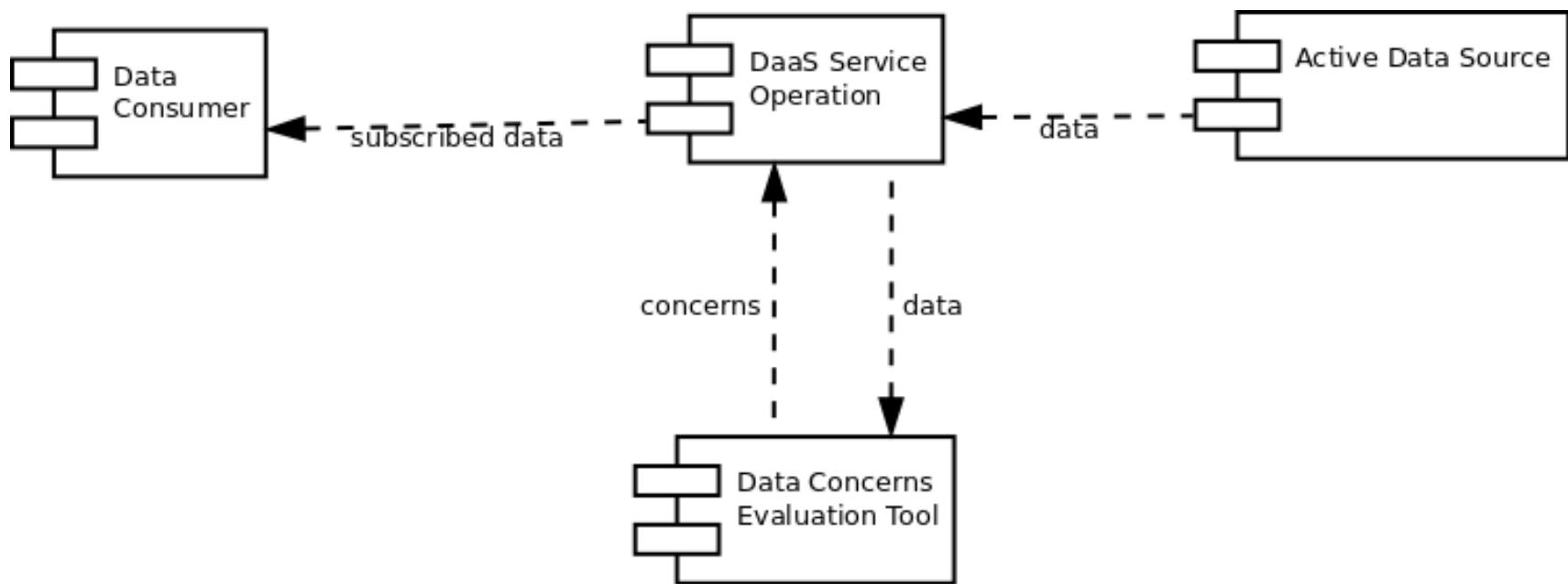
# Evaluating data concerns – some patterns (3)

## Push, pass-by-values (1)



# Evaluating data concerns – some patterns (4)

## Push, pass-by-values (2)



# Evaluation Tool – internal Software components

- Self-developed or third-party software components for evaluation tool
- Advantages
  - Tightly couple integration → performance, security, data compliance
  - Customization
- Disadvantages
  - Usually cannot be integrated with other features (e.g., data enrichment)
  - Costly (e.g., what if we do not need them)

# Evaluation tool – using cloud services

- Evaluation features are provided by cloud services
- Several implementations
  - Informatica Cloud Data Quality Web Services, StrikeIron,
- Advantages
  - Pay-per-use, combined features
- Disadvantages
  - Features are limited (with certain types of data)
  - Performance issues with large-scale data
  - Data compliance and security assurance

# Evaluation Tool -- using human computation capabilities

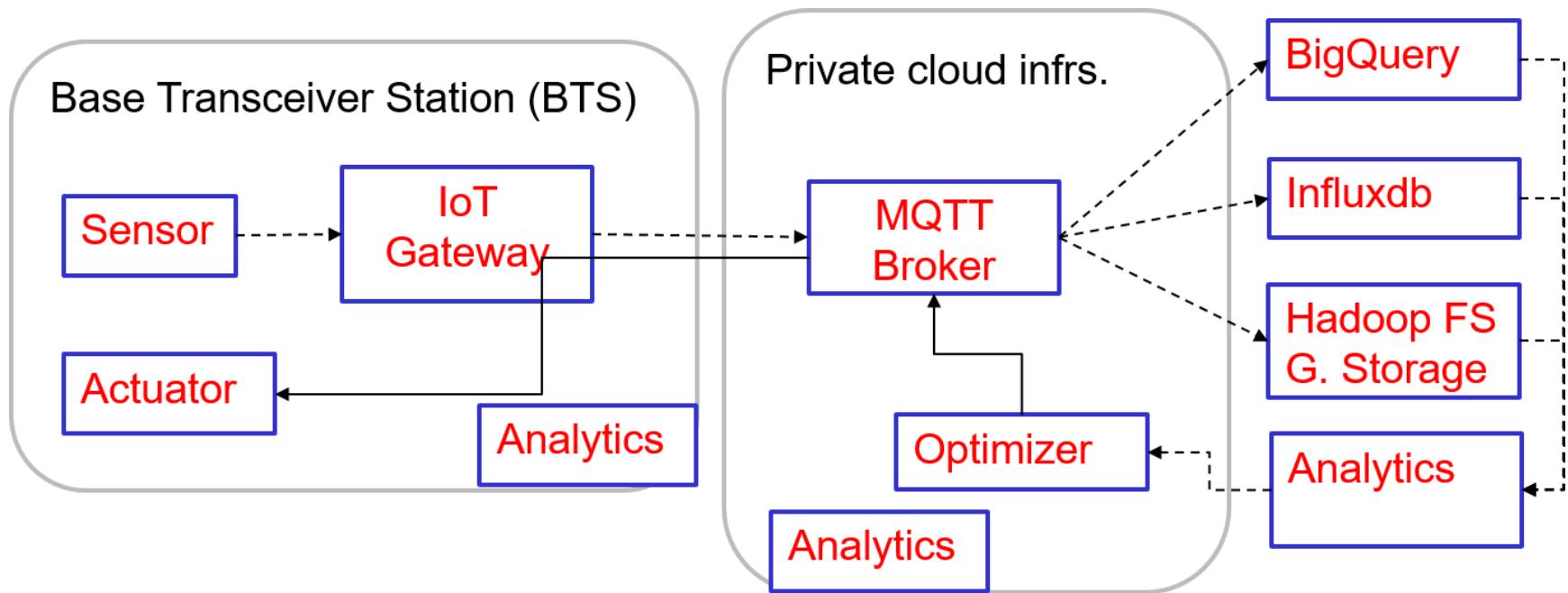
- Professionals and Crowds can act as data concerns evaluators
  - For complex quality assessment that cannot be done by software
- Issues
  - Subjective evaluation
  - Performance
  - Limited type of data (e.g., images, documents, etc.)

Michael Reiter, Uwe Breitenbücher, Schahram Dustdar, Dimka Karastoyanova, Frank Leymann, Hong Linh Truong: A Novel Framework for Monitoring and Analyzing Quality of Data in Simulation Workflows. eScience 2011: 105-112

Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, Jens Lehmann: Crowdsourcing Linked Data Quality Assessment. International Semantic Web Conference (2) 2013: 260-276

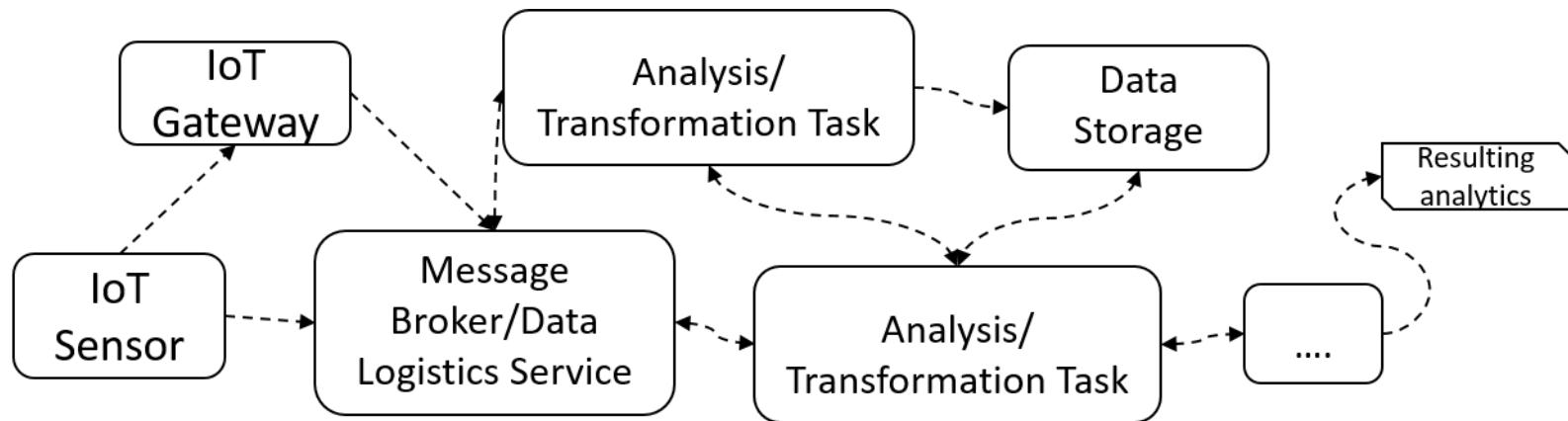
Óscar Figuerola Salas, Velibor Adzic, Akash Shah, and Hari Kalva. 2013. Assessing internet video quality using crowdsourcing. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (CrowdMM '13). ACM, New York, NY, USA, 23-28. DOI=10.1145/2506364.2506366 <http://doi.acm.org/10.1145/2506364.2506366>

# Evaluating concerns in in the big data pipeline



data concerns: why, how and what?

# Abstract software components for identifying “How” and “What”



Large number of data sources (e.g., IoT devices)

Large-scale brokers & data transfer/logistics services

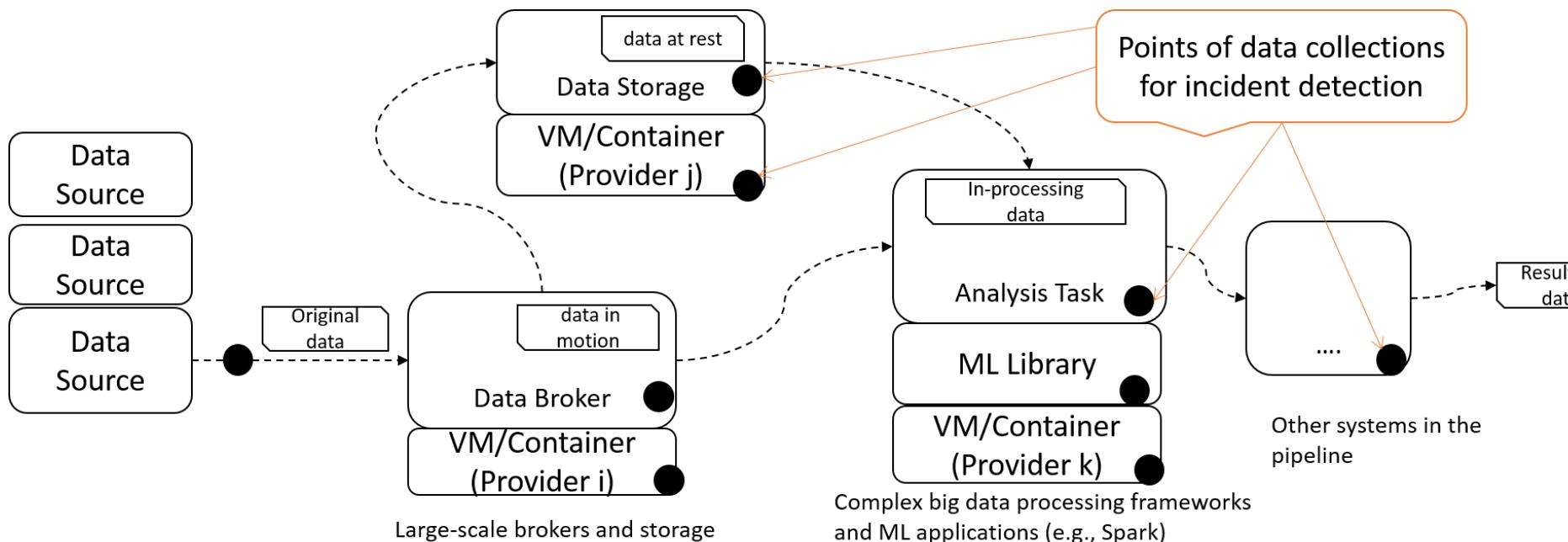
Complex big data processing frameworks

Other systems in the pipeline

Hong-Linh Truong, Manfred Halper, „**Characterizing Incidents in Cloud-based IoT Data Analytics**”, Proceedings of The 42nd IEEE International Conference on Computers, Software and Applications (COMPSAC 2018)

# Monitor and evaluating concerns

Identify **where** and **when** to do instrumentation and monitoring



Hong-Linh Truong, Manfred Halper, „**Characterizing Incidents in Cloud-based IoT Data Analytics**”, Proceedings of The 42nd IEEE International Conference on Computers, Software and Applications (COMPSAC 2018)

# Data quality monitoring at large-scale

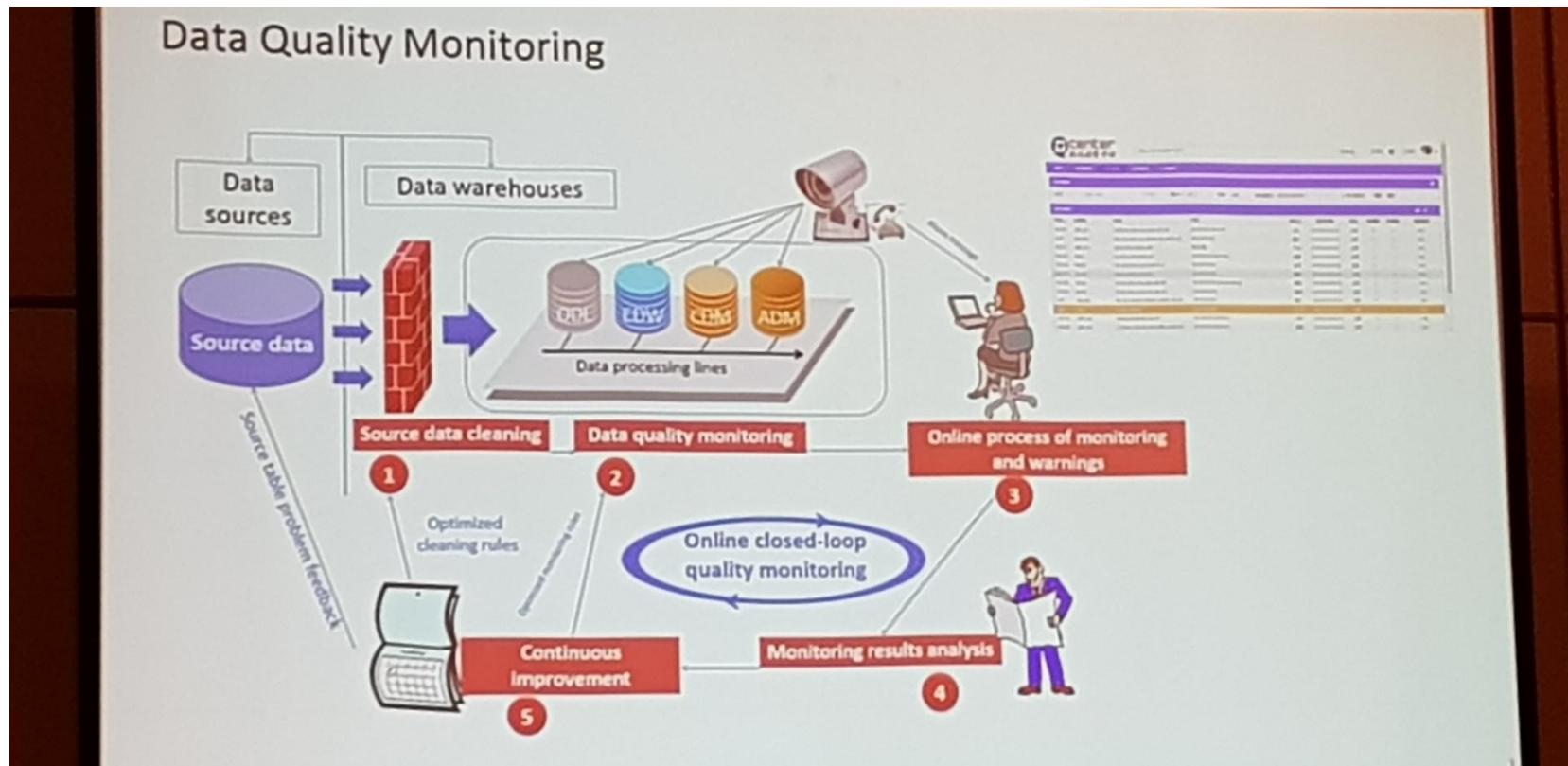


Figure taken from the keynote talk “Data Intelligence and Analytics at Alibaba” of Dr. Jingren Zhou, Alibaba Group at the 2018 International Conference on Cloud Engineering (IC2E 2018)

# How do some big data tools support GDPR?

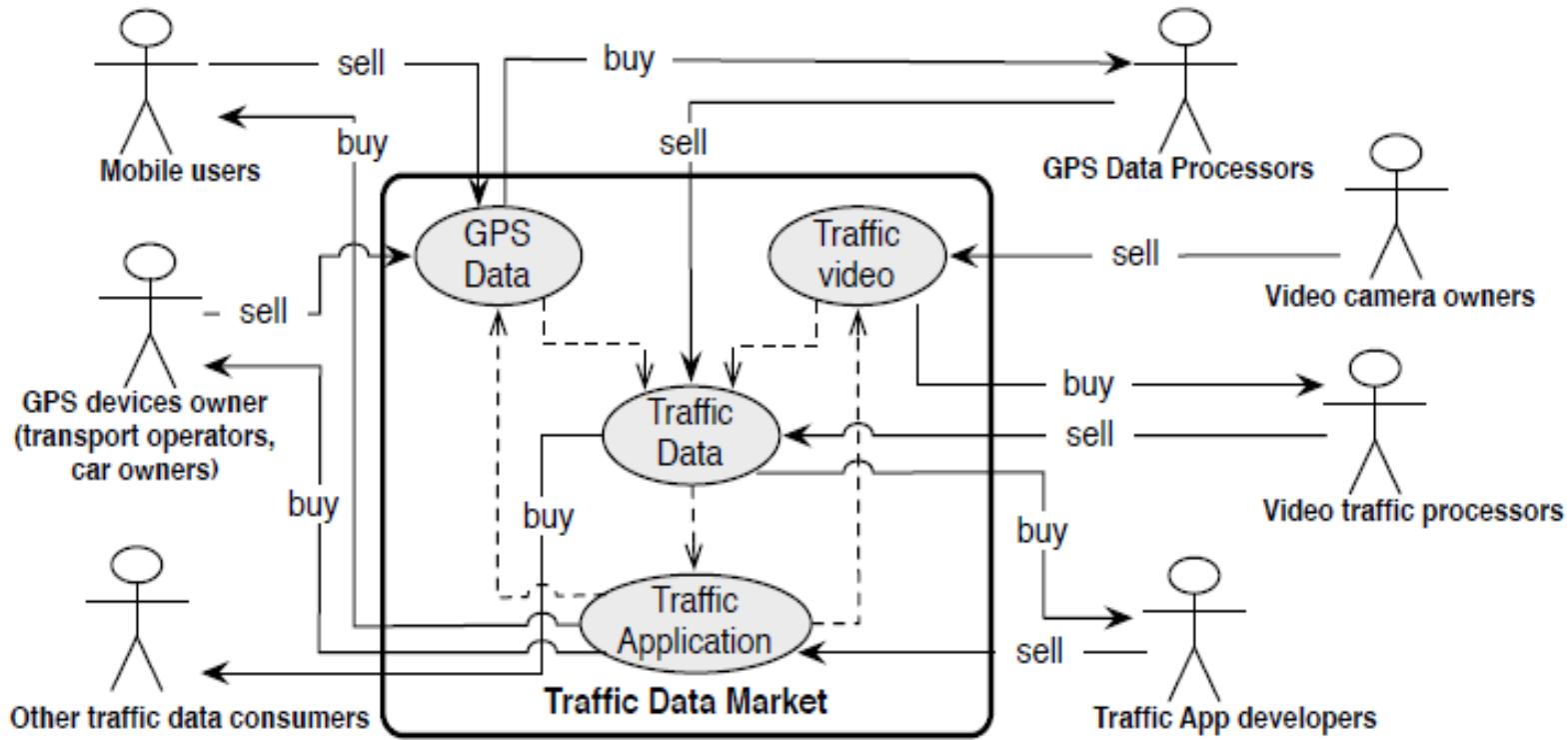
- Big data systems like: ElasticSearch, Hadoop, Bigquery or Apache Nifi
  - Only provide common features (e.g., logs, encryption, data retention rules, ...) so we have to design and deveop suitable features
- Some readings:
  - <https://hortonworks.com/blog/data-protection-takes-center-stage-gdpr/>
  - [https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.6.0/bk\\_data-governance/content/ch\\_hdp\\_data\\_governance\\_overview.html](https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.6.0/bk_data-governance/content/ch_hdp_data_governance_overview.html)
  - <https://www.elastic.co/gdpr>

# DATA MARKETPLACE

# Data marketplaces

- More than just DaaS
  - DaaS focuses on data provisioning features
- Stakeholders in data marketplaces
  - Multiple data providers and consumers
  - Marketplace providers
  - Marketplace authorities
  - Analytics providers
  - Data transportation providers
  - Billing and payment providers

# Example of stakeholders



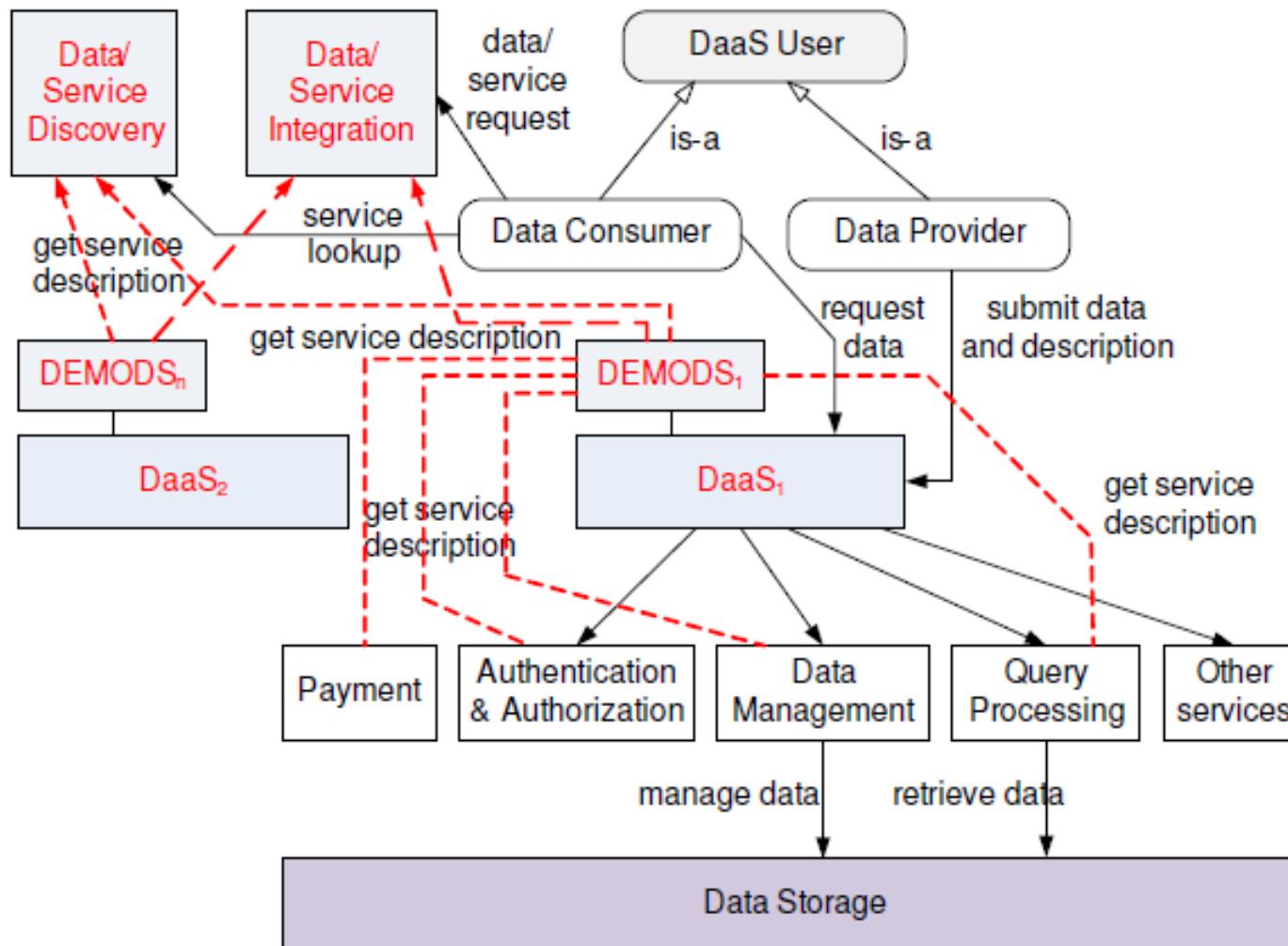
Tien-Dung Cao, Quang-Hieu Vu, Duc-Hung Le, Hong-Linh Truong, Schahram Dustdar: MARSA: A Marketplace for Realtime Human-Sensing Data.  
<http://dungcao.github.io/marsa/>

**Specific data market or generic data market?**

# Technical services, protocols, mechanisms in data marketplaces

- Multiple DaaS provisioning
  - Access models and interfaces
- Complex interactions among DaaS providers, data providers, data consumers, marketplace providers, etc.
  - Data exchange as well as payment
- Complex billing and pricing models
- Market dynamics
- Service and data contracts

# Data marketplaces and related components/services



# Data contracts

- Give a clear information about data usage
- Have a remedy against the consumer for illegal data usage
- Limit the liability of data providers in case of failure of the provided data;
- Specify information on data delivery, acceptance, and payment

# Data contracts

- Well-researched contracts for services but not for DaaS and data marketplaces
  - But **service APIs != data APIs != data assets**
- Several open questions
  - Right to use data? Quality of data in the data agreement? Search based on data contract? Etc.

- Require extensible models
- Capture contractual terms for data contracts
  - Support (semi-)automatic data service/data selection techniques.

Hong-Linh Truong, Marco Comerio, Flavio De Paoli, G.R. Gangadharan, Schahram Dustdar, "**Data Contracts for Cloud-based Data Marketplaces**", International Journal of Computational Science and Engineering, 2012 Vol.7, No.4, pp.280 - 295

# Study of main data contract terms

- Data rights
  - Derivation, Collection, Reproduction, Attribution
- Quality of Data (QoD)
  - Not mentioned, Not clear how to establish QoD metrics
- Regulatory Compliance
  - Sarbanes-Oxley, EU data protection directive, etc.
- Pricing model
  - Different models, pricing for data APIs and for data assets
- Control and Relationship
  - Evolution terms, support terms, limitation of liability, etc

Most information is in human-readable form

# Representing data contract terms

- Contract term: (termName,termValue)
  - Term name: common terms or user-specific terms
  - Term value: a single value, a set, or a range

<i>Category</i>	<i>Term representation</i>	<i>Examples</i>
Data rights	$\text{termName} = \{\text{val}_1, \text{val}_2, \dots, \text{val}_n\}$	$\text{termName} = \{\text{Derivation}, \text{Collection}, \text{Reproduction}, \text{Attribution}, \text{Noncommercialuse}\}$ , $\text{val}_i = \{\text{Undefined}, \text{Null}, \text{Allowed}, \text{Required}, \text{True}, \text{False}\}$
Quality of data	$\text{val}_l \leq \text{termName} \leq \text{val}_u$	$\text{termName} = \{\text{Accuracy}, \text{Completeness}, \text{Uptodateness}\}$ , $\text{val}_l$ and $\text{val}_u \in [0, 1]$
Compliance	$\text{termName} = \{\text{val}_1, \text{val}_2, \dots, \text{val}_n\}$	$\text{termName}$ and $\text{val}_i$ are any string, e.g., $\text{termName} = \{\text{PrivacyCompliance}\}$ and $\text{termValue} = \{\text{Sarbanes-Oxley (SOX) Act}\}$
Pricing model	$\text{termName} = (\text{cost} = \text{val}_1, \text{usagetime} = \text{val}_2, \text{maximumuse} = \text{val}_3)$	$\text{termName}$ is any string, e.g., $\text{MonthlyPayment}$ ; $\text{val}_1 \in R$ , e.g., $\text{cost} = 50 \text{ €}$ , $\text{val}_2 = \{(\text{end}_t - \text{start}_t); \text{UNLIMITED}\}$ where $\text{end}_t, \text{start}_t \in \text{datetime}$ , e.g., $\text{usagetime} = 30 \text{ days}$ ; $\text{val}_3 \in N$ , e.g., $\text{maximumuse} = 1,000 \text{ calls}$
Control and relationship	$\text{termName} = \text{val}$	$\text{termName}$ and $\text{val}$ are any string, e.g., $\text{termName} = \{\text{Liability}, \text{LawandJurisdiction}\}$ and $\text{val} = \{\text{US}, \text{Austria}\}$

Discussion time

## **HOW DOES NEAR-REALTIME DATA IMPACT ON DATA CONTRACT EXCHANGE?**

# DATA MARKETPLACE IN BLOCKCHAIN AGE

# IoT Data Market without Marketplace?

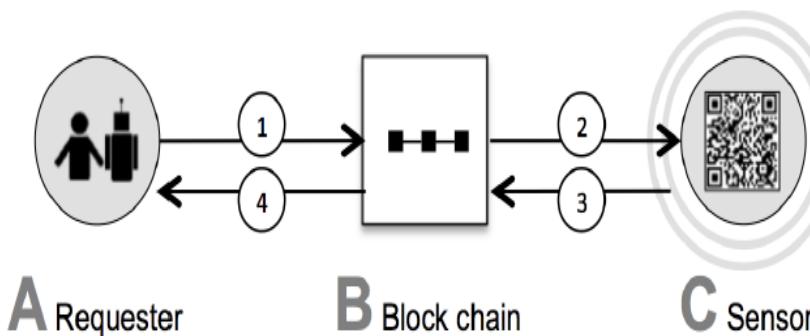


Fig. 1. Schema for the atomic S<sup>2</sup>aaS process of exchanging a single datum for cash using Bitcoin.

Kay Noyen, Dirk Volland, Dominic Wörner, Elgar Fleisch:  
When Money Learns to Fly: Towards Sensing as a Service Applications Using Bitcoin.

But what about data contract? → smart contract

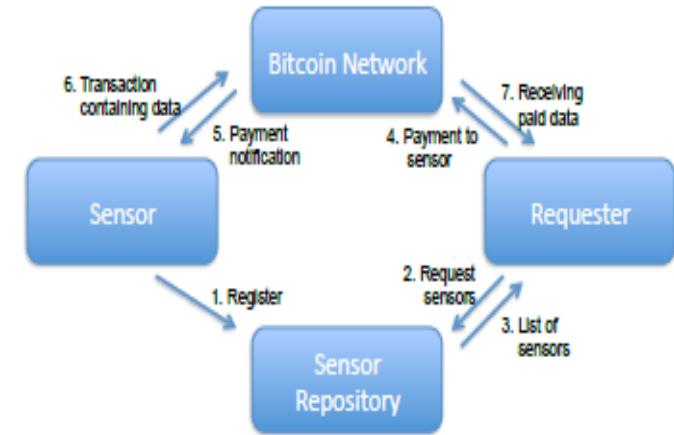
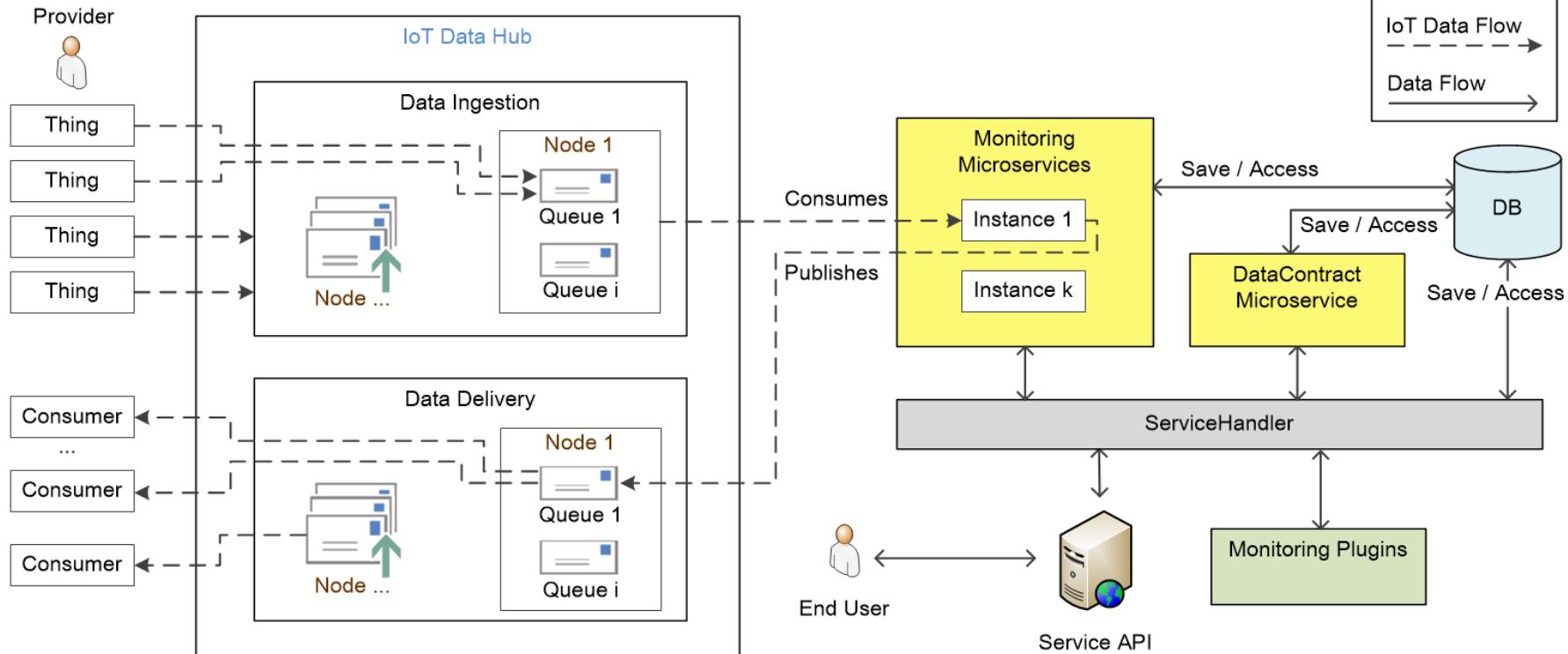


Figure 1: Process for exchanging data for bitcoin.

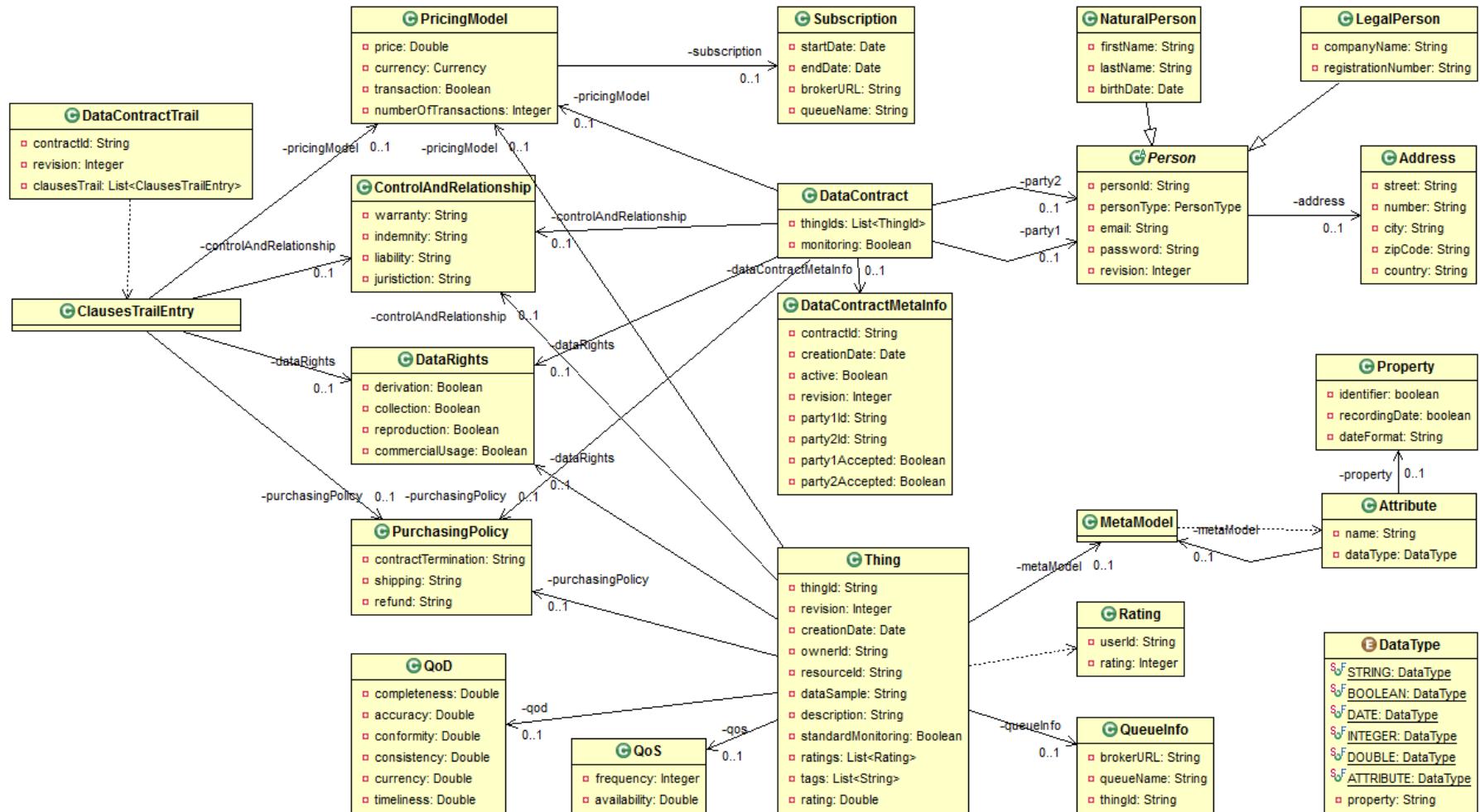
Dominic Wörner and Thomas von Bomhard. 2014. **When your sensor earns money: exchanging data for cash with Bitcoin**. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct). ACM, New York, NY, USA, 295-298.

# Generic Contract-aware Framework Architecture



Florin-Bogdan Balint, Hong-Linh Truong:  
On Supporting Contract-Aware IoT Dataspace Services. MobileCloud 2017: 117-124

# IoT Data Contract Design



# IoTA Data marketplace

- The principle is very much like any other data marketplaces
- Payment via IoTA

## Marketplace Sensors

Click below to view and purchase sensor data stream

<p>Building Management System <b>3for2_API_feed</b></p> <p> Location       Sensor streams: Singapore      21</p> <p>Owner: <b>3for2</b> Data Price: <b>1794I</b></p>	<p>Kodi media center <b>Aorange-kodiFranck</b></p> <p> Location       Sensor streams: Grenoble      5</p> <p>Owner: <b>Orange</b> Data Price: <b>350I</b></p>	<p>Weather Station <b>BerryPi_RuuviTag</b></p> <p> Location       Sensor streams: Rivoli (TO)      8</p> <p>Owner: <b>france193</b> Data Price: <b>295I</b></p>
<p>Temperature <b>BerryPi_temp102</b></p> <p> Location       Sensor streams: Rivoli (TO)      1</p> <p>Owner: <b>france193</b> Data Price: <b>90350I</b></p>	<p>XDK Sensor <b>Connectory-XDK-01</b></p> <p> Location       Sensor streams: Connectory      4</p> <p>Owner: <b>Connectory</b> Data Price: <b>1337I</b></p>	<p>Temp sensor <b>DNVGLTEST</b></p> <p> Location       Sensor streams: Trondheim      1</p> <p>Owner: <b>DNV GL</b> Data Price: <b>359I</b></p>

Figure captured from <https://data.iota.org/>

## Using smart contract contract

### HOW MONETASA DATAMARKET WORKS



Figure captured from <https://monetasa.com/>

# Databroker DAO

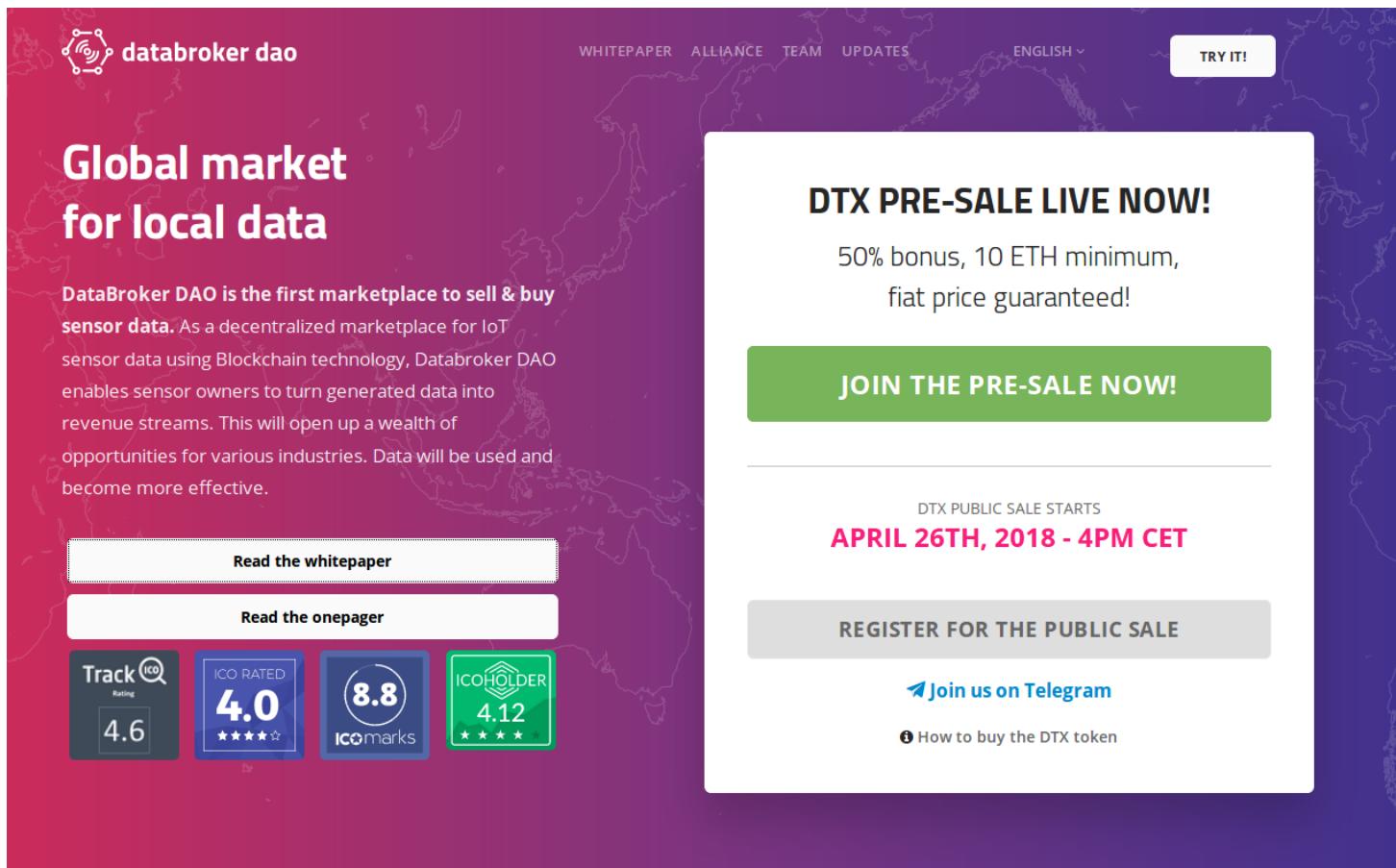


Figure captured from <https://databrokerdao.com/>

# Exercises

- Read mentioned papers
- Check characteristics, service models and deployment models of mentioned DaaS (and find out more)
- Identify services in the ecosystem of some DaaS
- Turn some data to DaaS using existing tools

# Exercises (2)

- For your mini project:
  - Identify and analyze the relationships between data concerns evaluation tools and types of data
  - Analyze trade-offs between on-line and off-line evaluation and when we can combine them
  - Analyze how to utilize evaluated data concerns for optimizing data compositions
  - Analyze situations when software cannot be used to evaluate data concerns

# Exercises (3)

- For your mini project:
  - Develop some specific data contracts for open government data
  - Work on some algorithms for checking data contract compatibility
  - Incorporate data marketplaces concepts into your scenario

# CASE STUDY – DESIGN DATA MARKETPLACE

MARSA: A Marketplace for Realtime Human-Sensing Data

Cao, Tien-Dung ; Pham, Tran-Vu ; Vu, Quang-Hieu ; Le, Duc-Hung  
; Truong, Hong-Linh ; Dustdar, Schahram

ACM Transactions on Internet Technology, 2016

<http://dungcao.github.io/marsa/>

# Traffic problems in HoChiMinh City

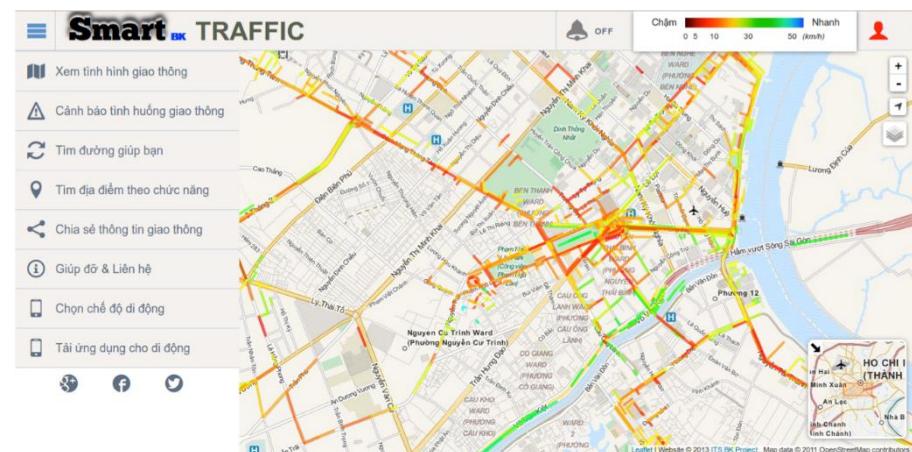


- Crowded and unpredictable
- Needs a lot of data to understand traffics
- Lack infrastructures for collecting traffic information
- Common problems in developing countries

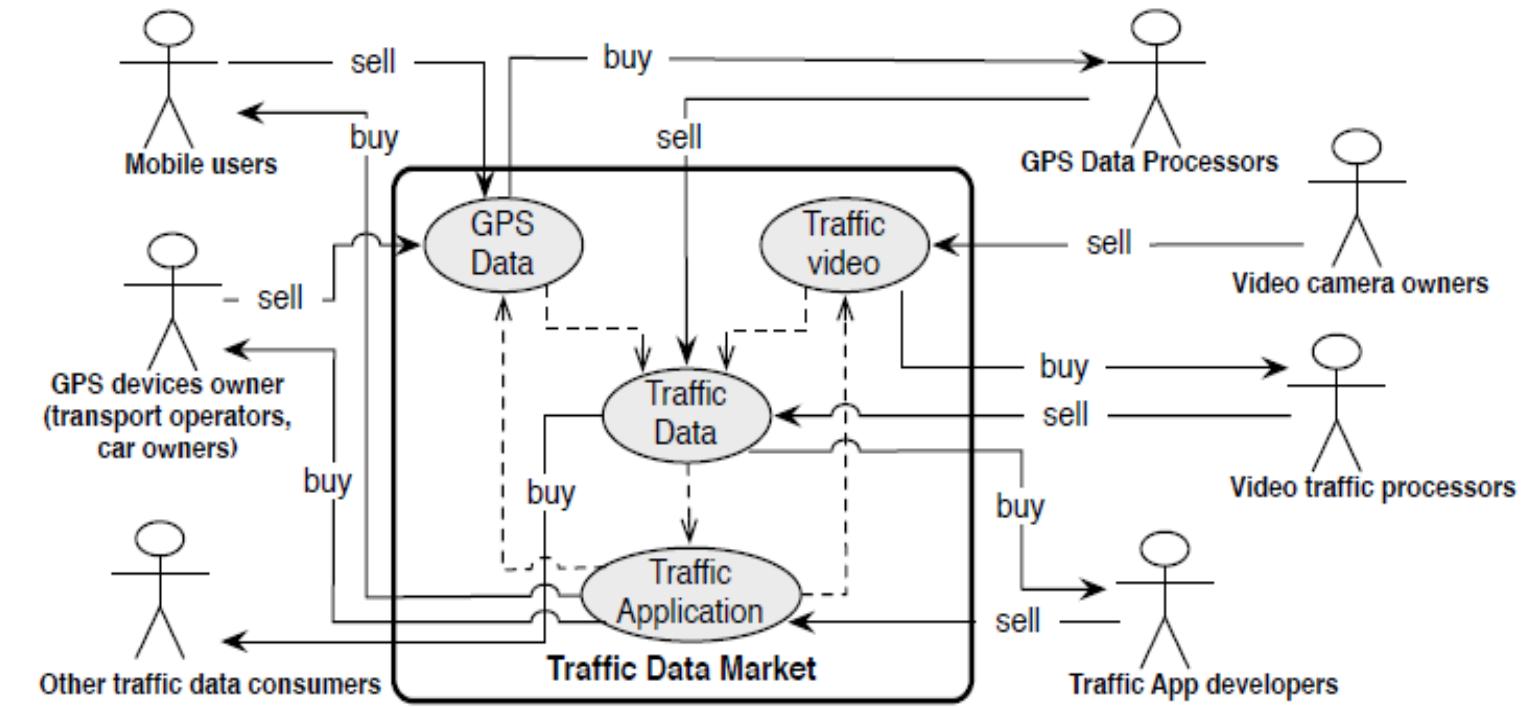
Figure sources: Internet

Cannot buy  
expensive traffic  
data collection  
systems!

ASE Fudan FIST, Summer 2018



# Market-oriented View of traffic data scenarios



4000 citybus fleet, 0.25MB per day per bus (7.5MB/month/bus), 30GB for the fleet

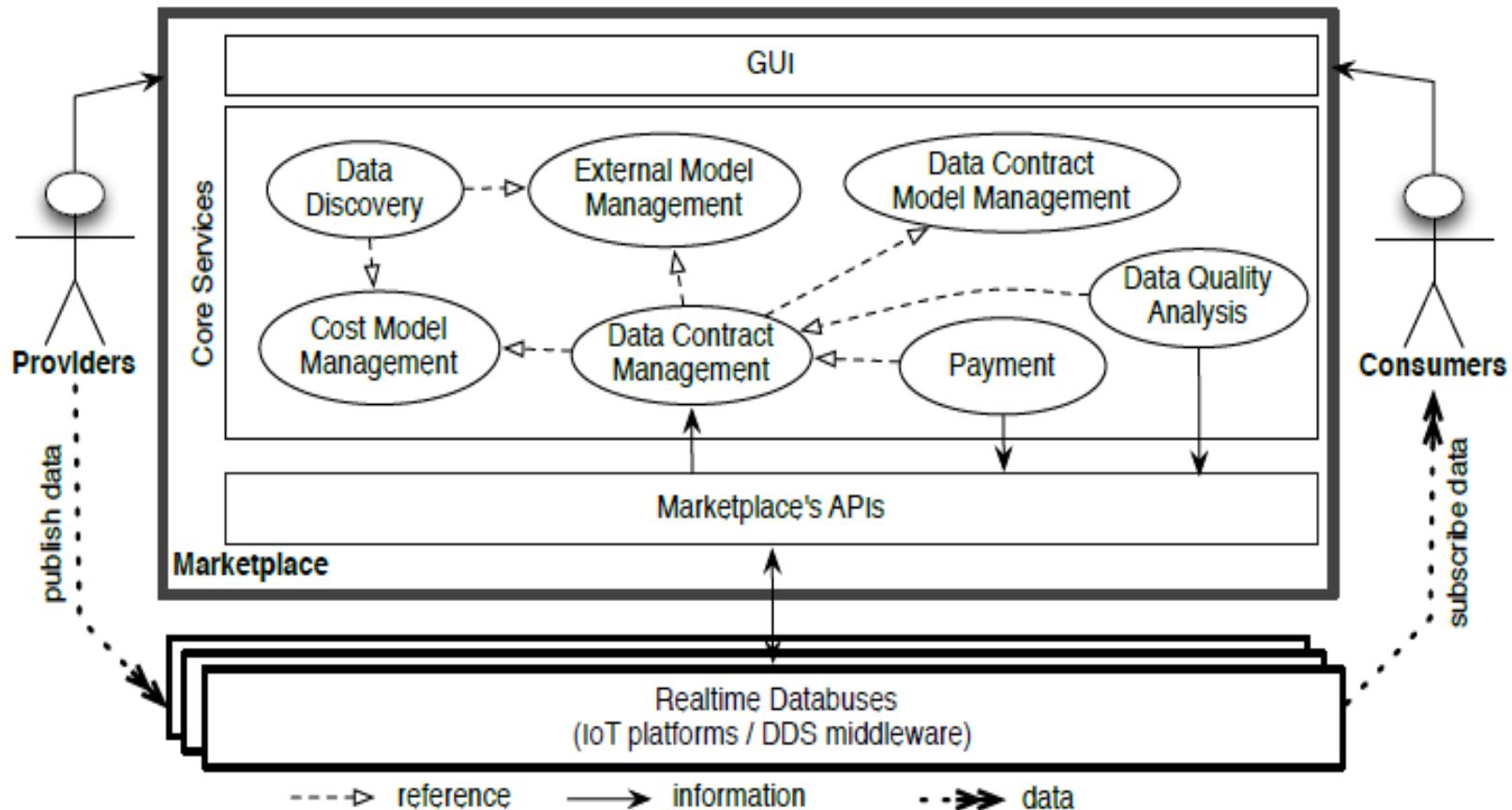
1MB of GPS data = 20 USD cent → 6000 USD for the fleet operators

A mobile phone, like a bus, can receive 1.5 USD per month →  $\frac{1}{2}$  of 3G data bill

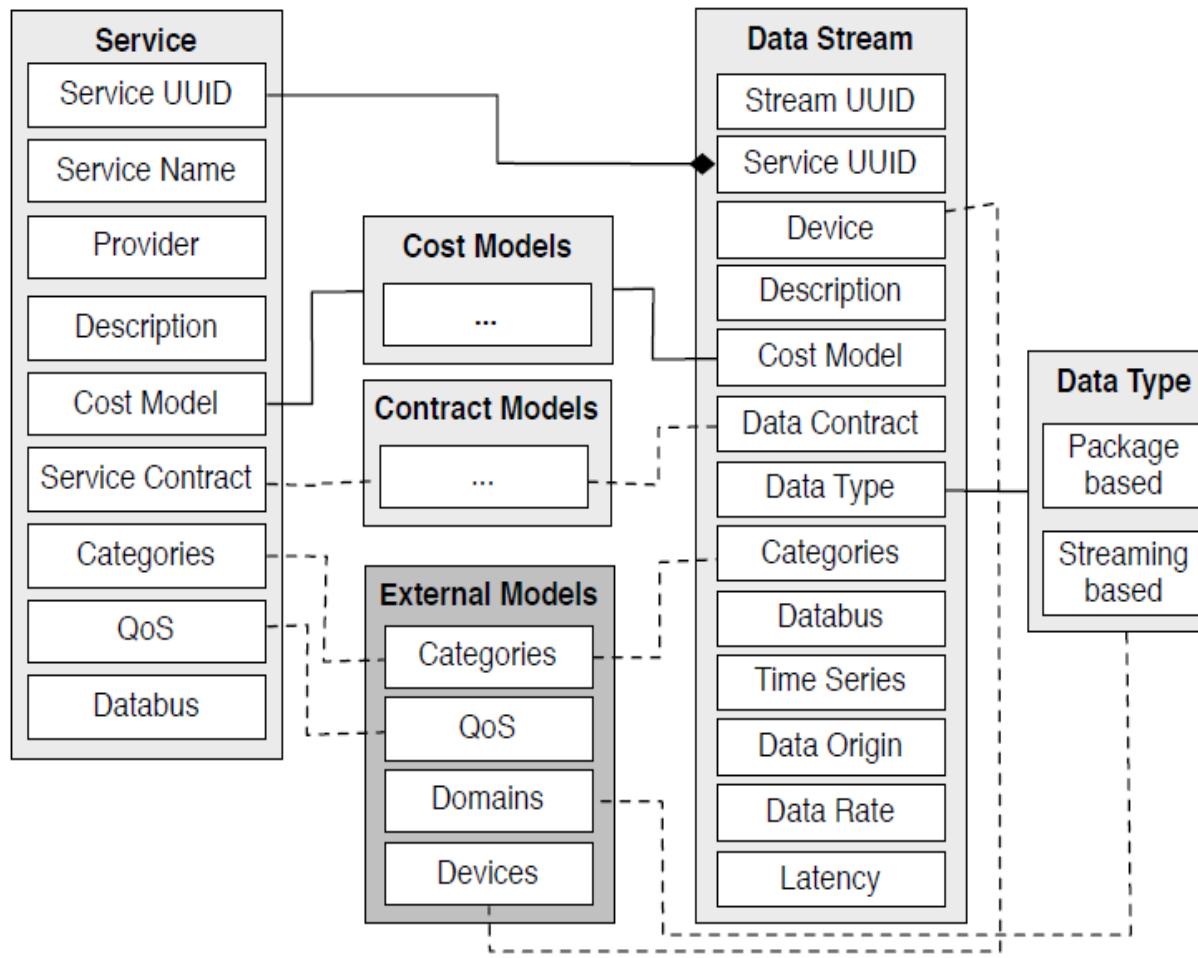
# Costs and benefits

Parties	Costs of collecting raw data	Benefits from processed traffic data
Bus, taxi and truck operators	GPS devices, Internet and mobile network subscription fees, acquiring and maintaining data at servers	Able to track status of their buses, knowledge of current traffic conditions to better provide services to commuters
Private car owners	GPS devices, mobile network subscription fees	Knowledge of current traffic conditions to better navigate in cities
Mobile device owners	Mobile devices (e.g. smartphones, tablets), mobile network subscription fees and device battery time	Knowledge of current traffic conditions to better navigate in cities
Video camera owners	Video cameras and network connections to video cameras	Selling of video data and traffic information
Data processors	Cost of raw data, infrastructures for collecting and processing raw data	Selling traffic data
Traffic data users	Buying traffic data	Knowledge of current traffic conditions to better navigate in cities

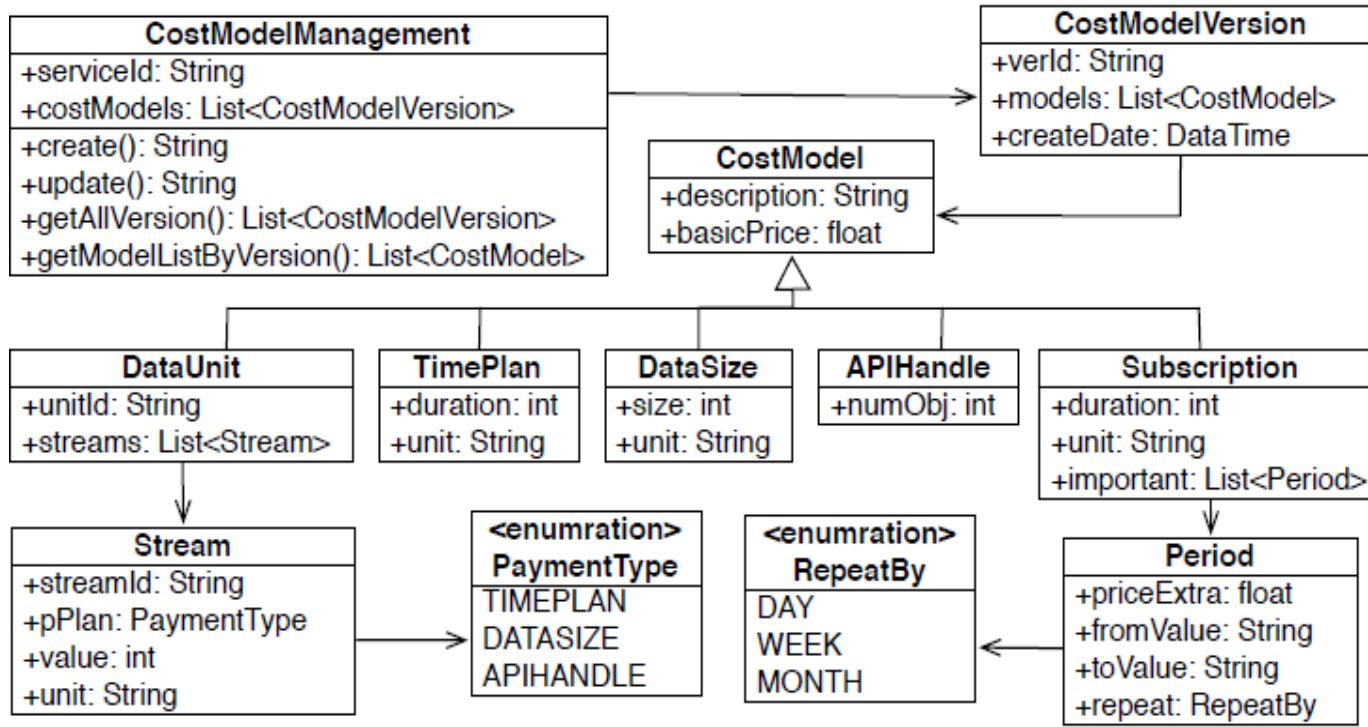
# MARSA Design overview



# MARSA description for human-sensing data marketplace

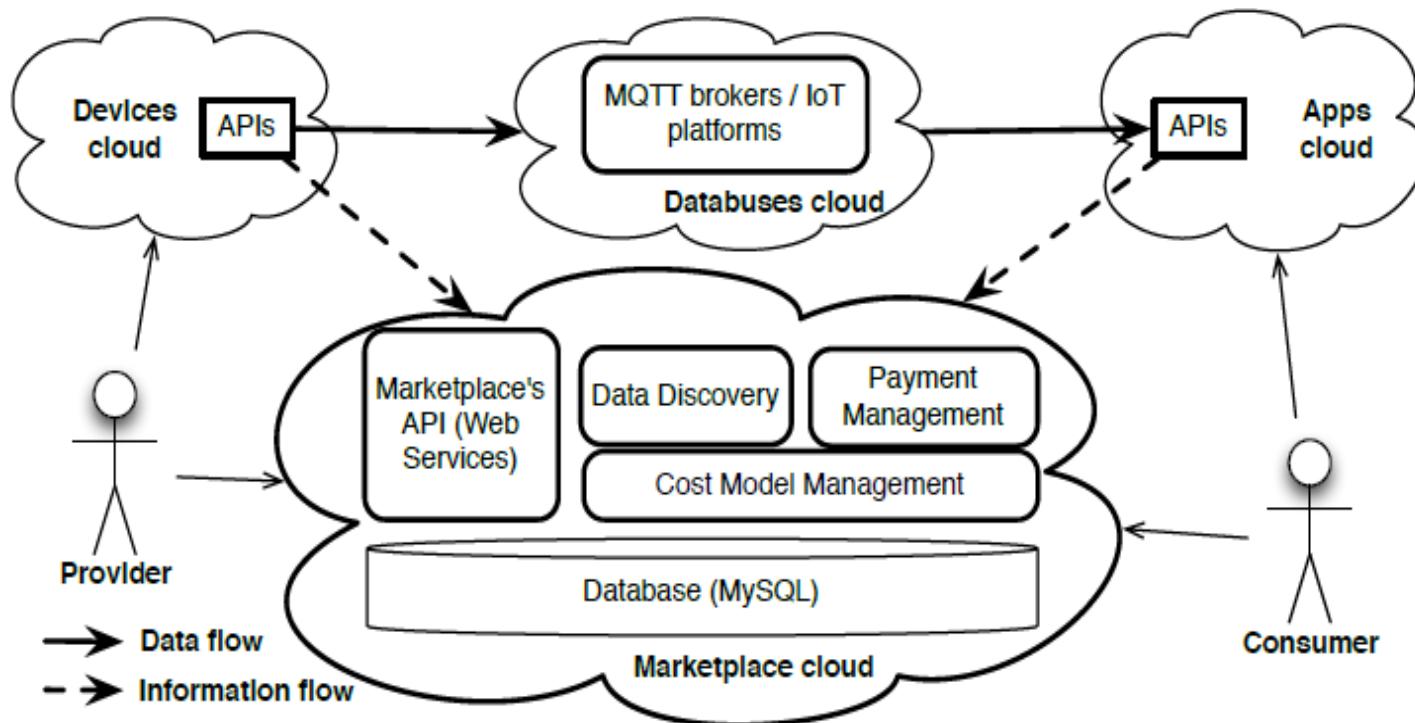


# Cost model

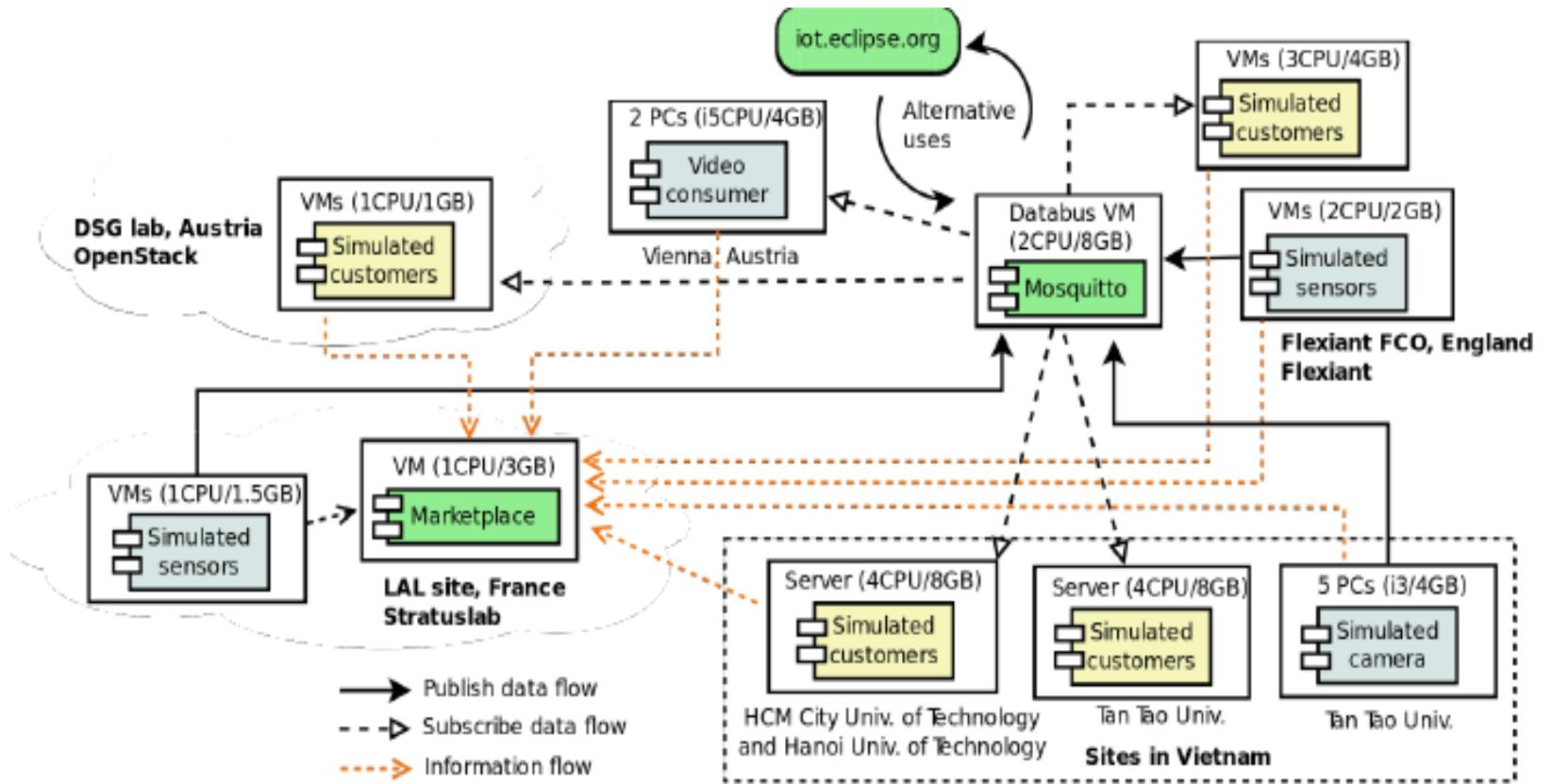


Quality of data has not supported yet

# Implementation



# Testbed



# Example of bills

**Bill No.: 2015/03-5.1**

From date: 2015-03-30 12:39:53 To date: 2015-03-30 18:40:57

Status: Not Payment

Payment on DATA\_SIZE (5.0 \$ / 1 GB)

List of streams

No.	Stream UUID	Size	Price
1	suuid1427702254973/sid1	0.219 GB	\$ 1.1
2	suuid1427702254973/sid2	0.0217 GB	\$ 0.11
3	suuid1427702254973/sid3	0.0550 GB	\$ 0.28
4	suuid1427702254973/sid4	0.181 GB	\$ 0.9
5	suuid1427702254973/sid5	0.205 GB	\$ 1.02
			<b>Total price: \$ 3.41</b>

Payment on SUBSCRIPTION (2.0 \$ / 1 HOUR)

List of streams

No.	Stream UUID	Size	Price	Size Extra	Price Extra	Sum Price
1	suuid1427702254973/sid11	3.67 HOUR	\$ 7.34	0	\$ 0	\$ 7.34
2	suuid1427702254973/sid12	6.02 HOUR	\$ 12.04	0	\$ 0	\$ 12.04
						<b>Total Price: \$ 19.38</b>

**Total price of contract: \$ 22.79**

# Thanks for your attention

Hong-Linh Truong

Distributed Systems Group, TU Wien

[Hong-linh.truong@tuwien.ac.at](mailto:Hong-linh.truong@tuwien.ac.at)

[@linhsolar](http://www.infosys.tuwien.ac.at/staff/truong)