Advanced Services Engineering,
Summer 2017 – Lecture 4

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

Hong-Linh Truong
Distributed Systems Group, TU Wien

truong@dsg.tuwien.ac.at
http://dsg.tuwien.ac.at/staff/truong
@linhsolar

Distributed Systems Group

Outline

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

From last year projects

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

“Measure and report water quality metrics”

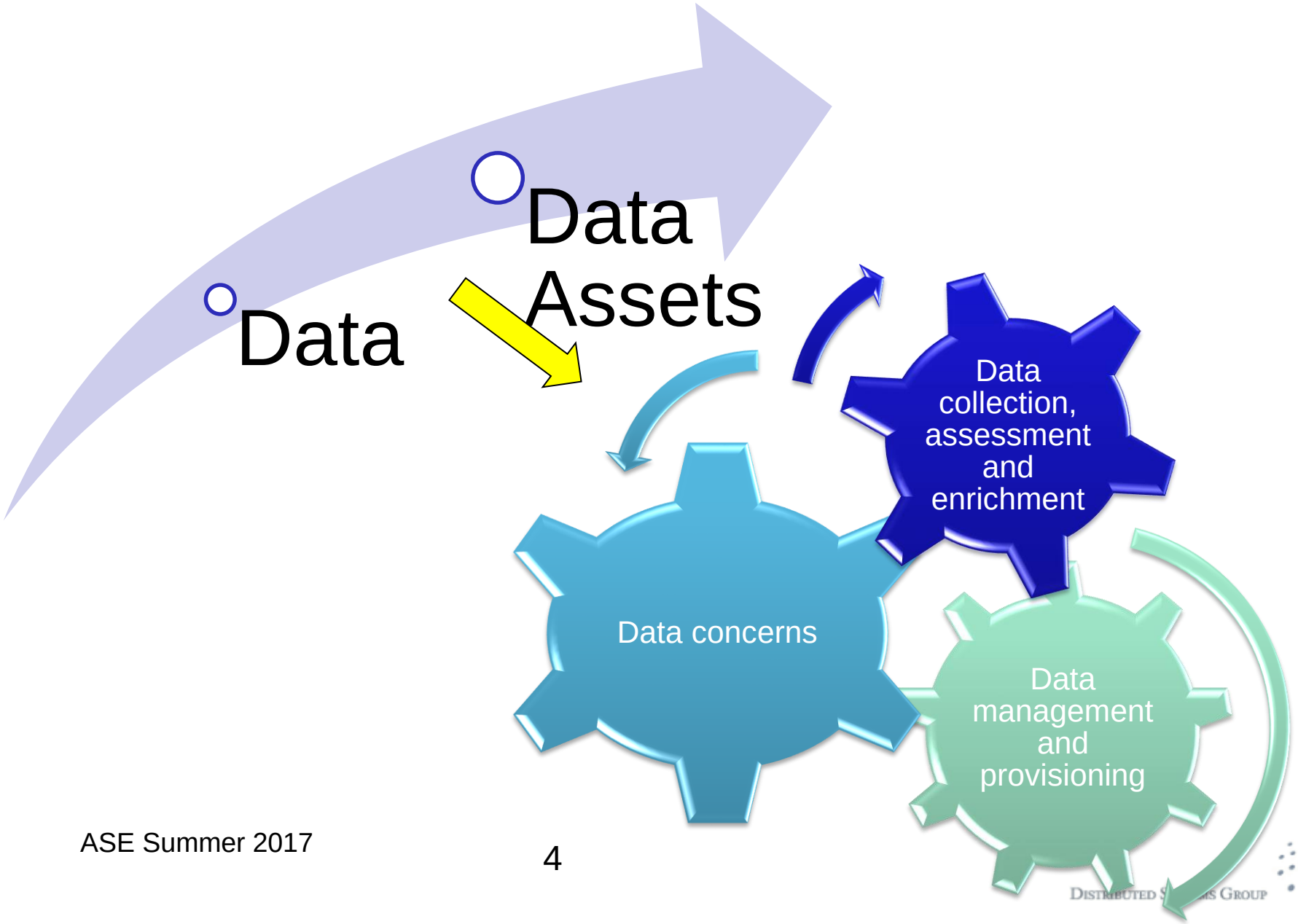
“Latest data on air quality is fetched from London Air API”

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

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

„real time production information from photovoltaic panels”

Data versus data assets



Data provisioning activities and issues

Provisioning Models

Collect
- Data sources
- Ownership
- License
- Quality assessment and enrichment

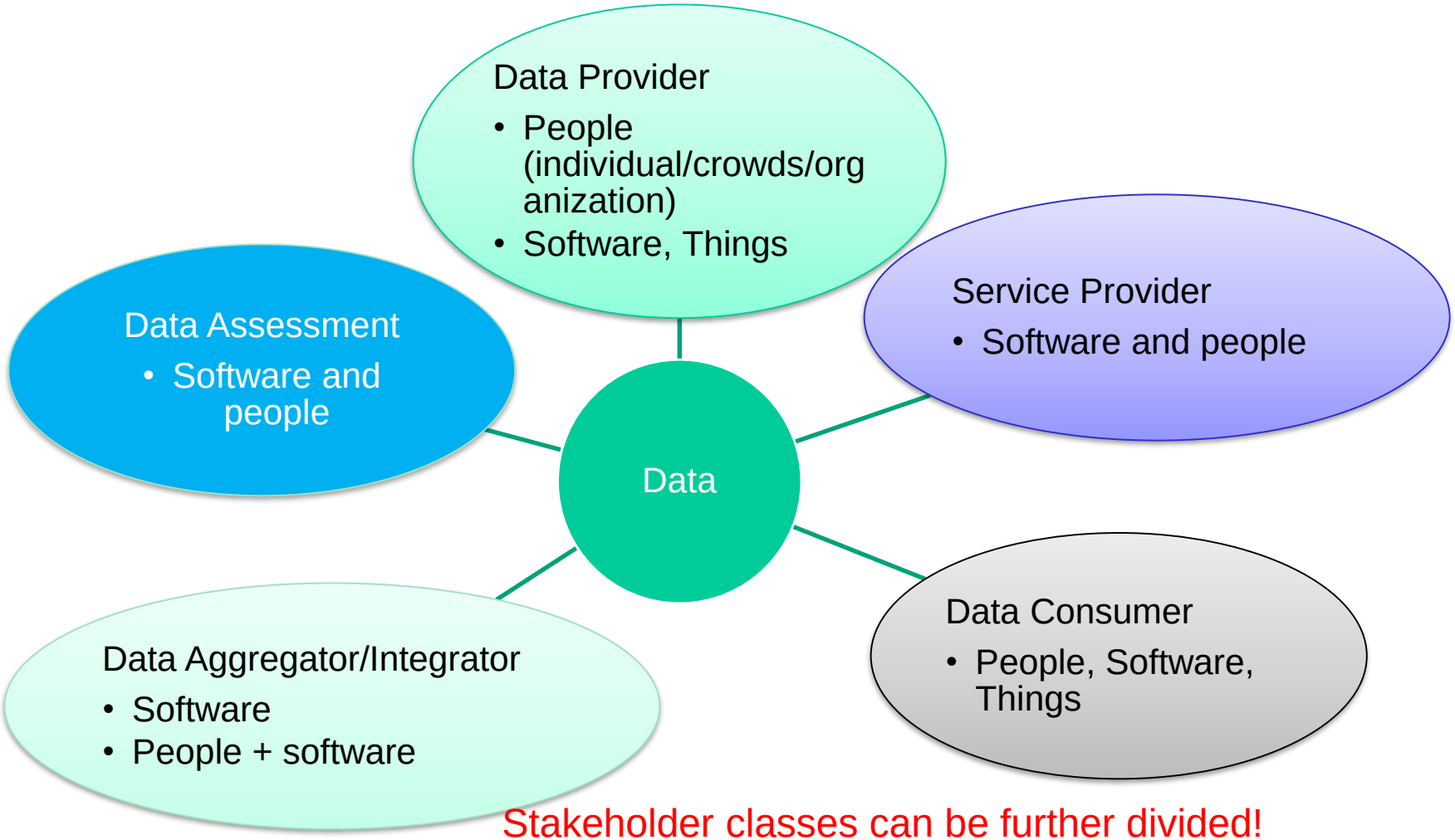
Store
- Query and backup capabilities
- Local versus cloud, distributed versus centralized storage

Access
- Interface
- Public versus private access
- Access granularity
- Pricing and licensing model

Utilize
- Alone or in combination with other data sources
- Redistribution
- Updates

Non-exhausive list! Add your own issues!

Stakeholders in data provisioning



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

Data service unit

Consumption, ownership, provisioning, price, etc.

Service model

Data

Unit Concept

„basic component“/“basic function“ modeling and description

Data service unit

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

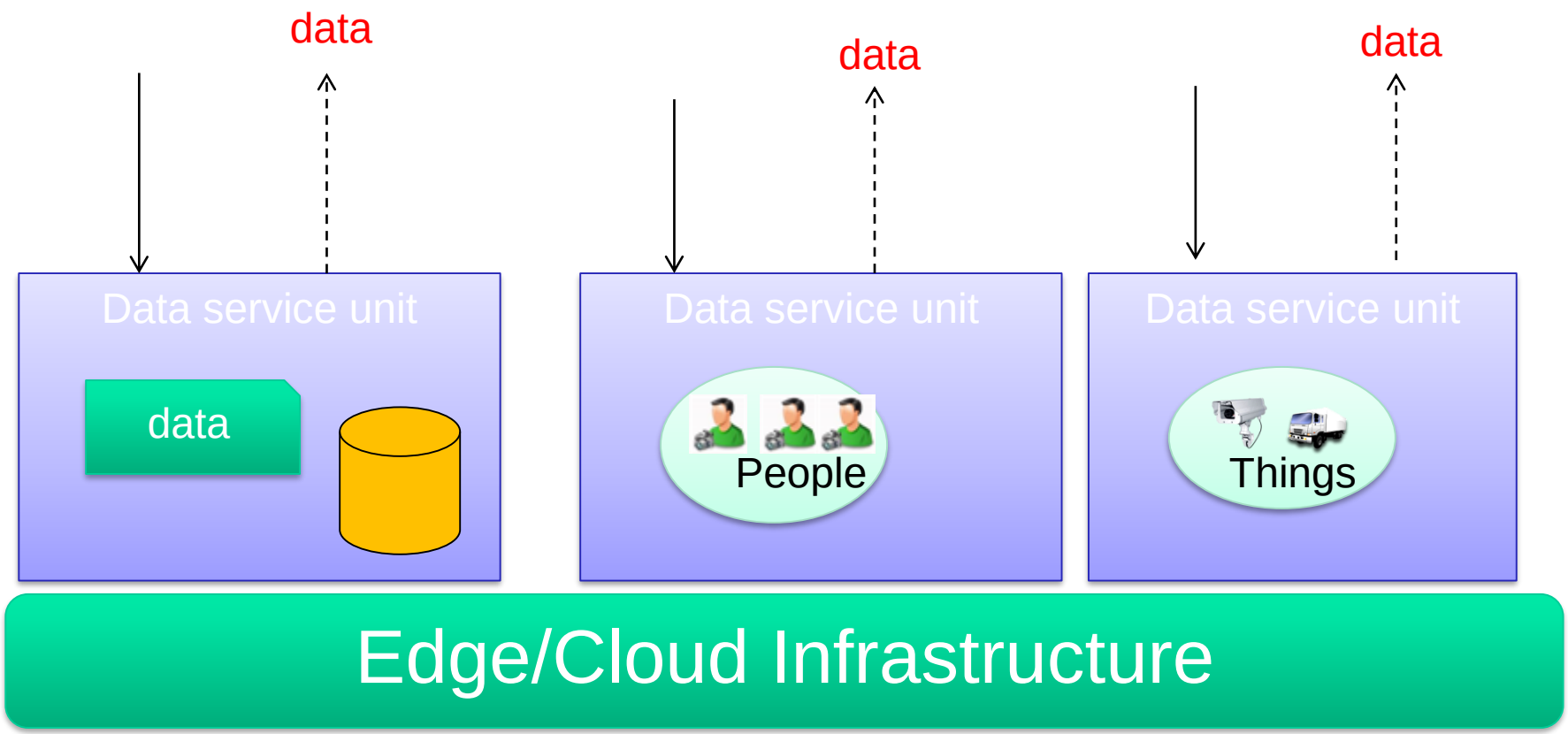
Data service units in clouds

- *Provide data capabilities* rather than provide computation or software capabilities

- Providing data in clouds/internet is an increasing trend
  - In both business and e-science environments

- Now often in a combination of data + analytics of the data → 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 access 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

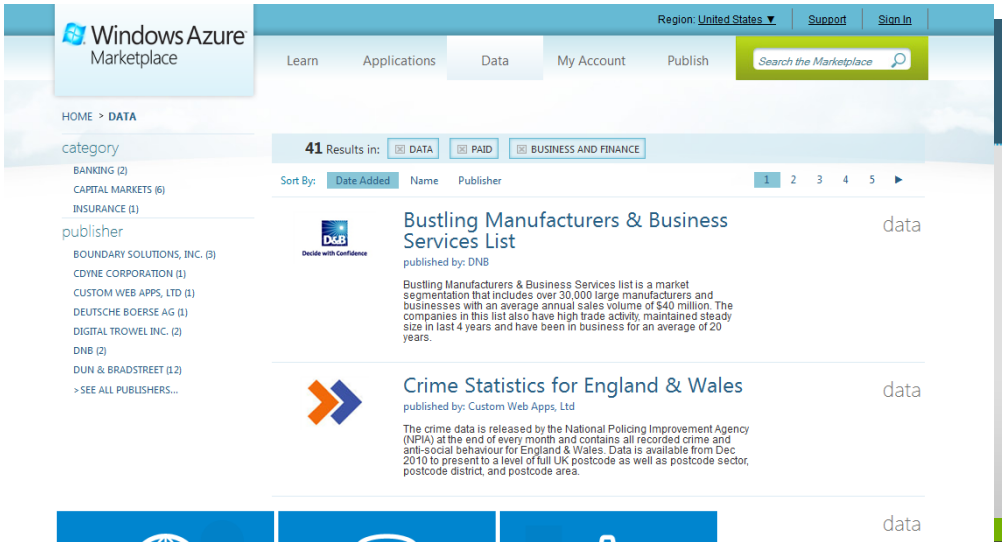
Data-as-a-Service – service models

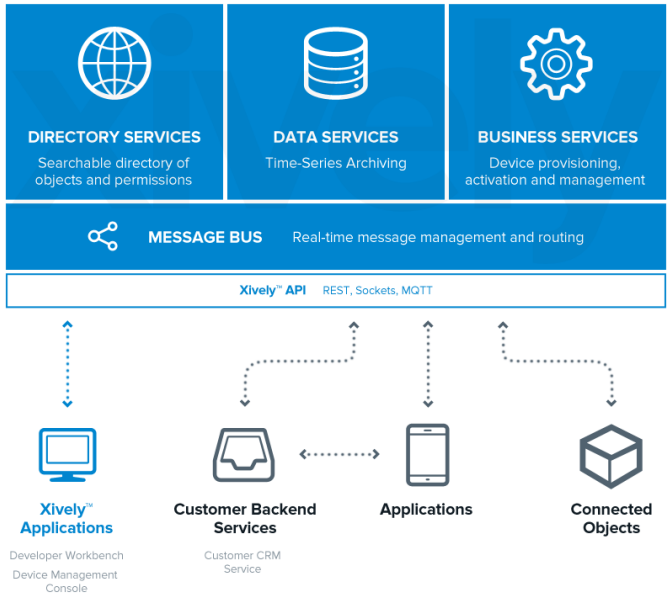
- Data publish/subcription middleware as a service
- Database-as-a-Service (Structured/non-structured querying systems)
- Sensor-as-a-Service
- Storage-as-a-Service (Basic storage functions)

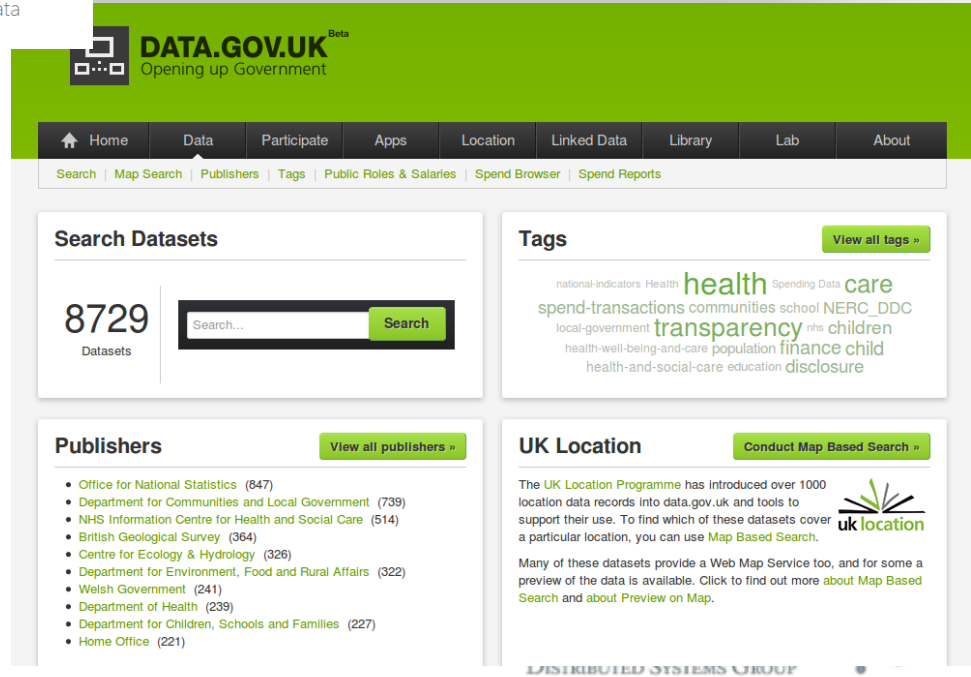
deploy

Edge and/or Cloud Systems

Examples of DaaS





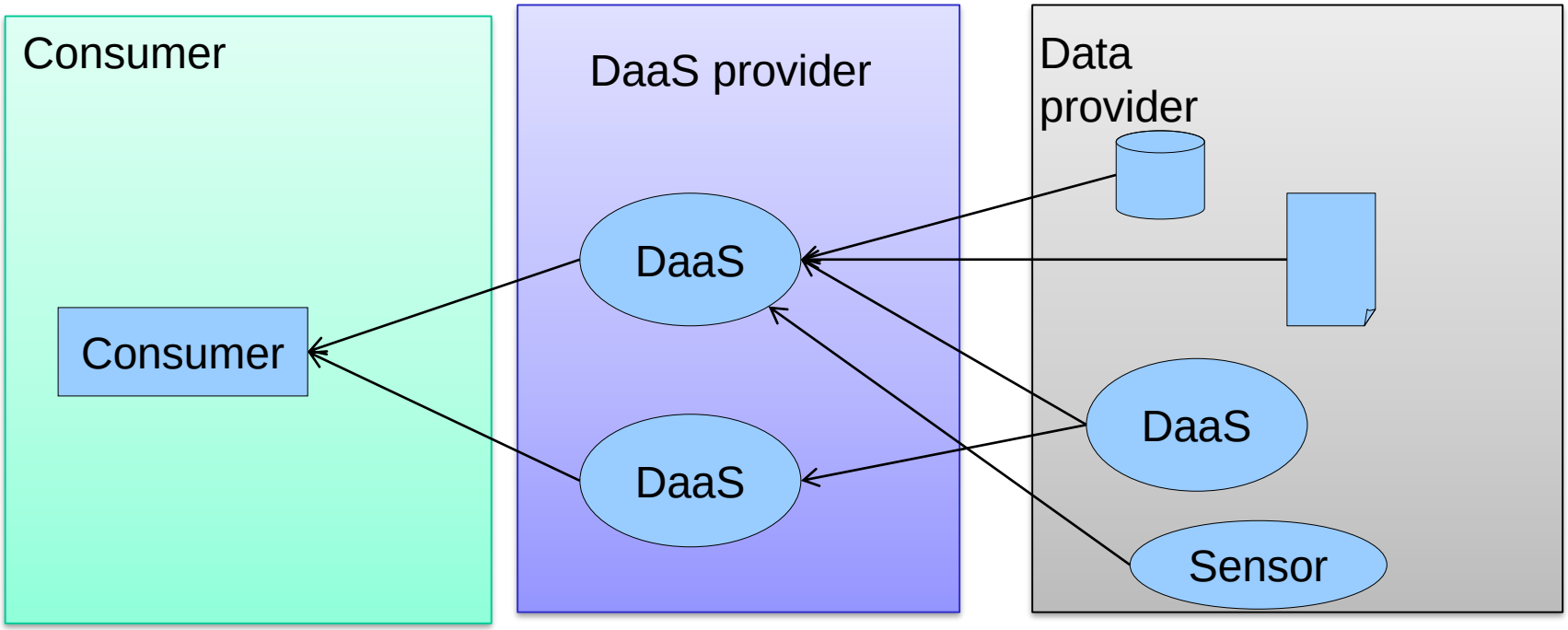


DaaS design & implementation – APIs

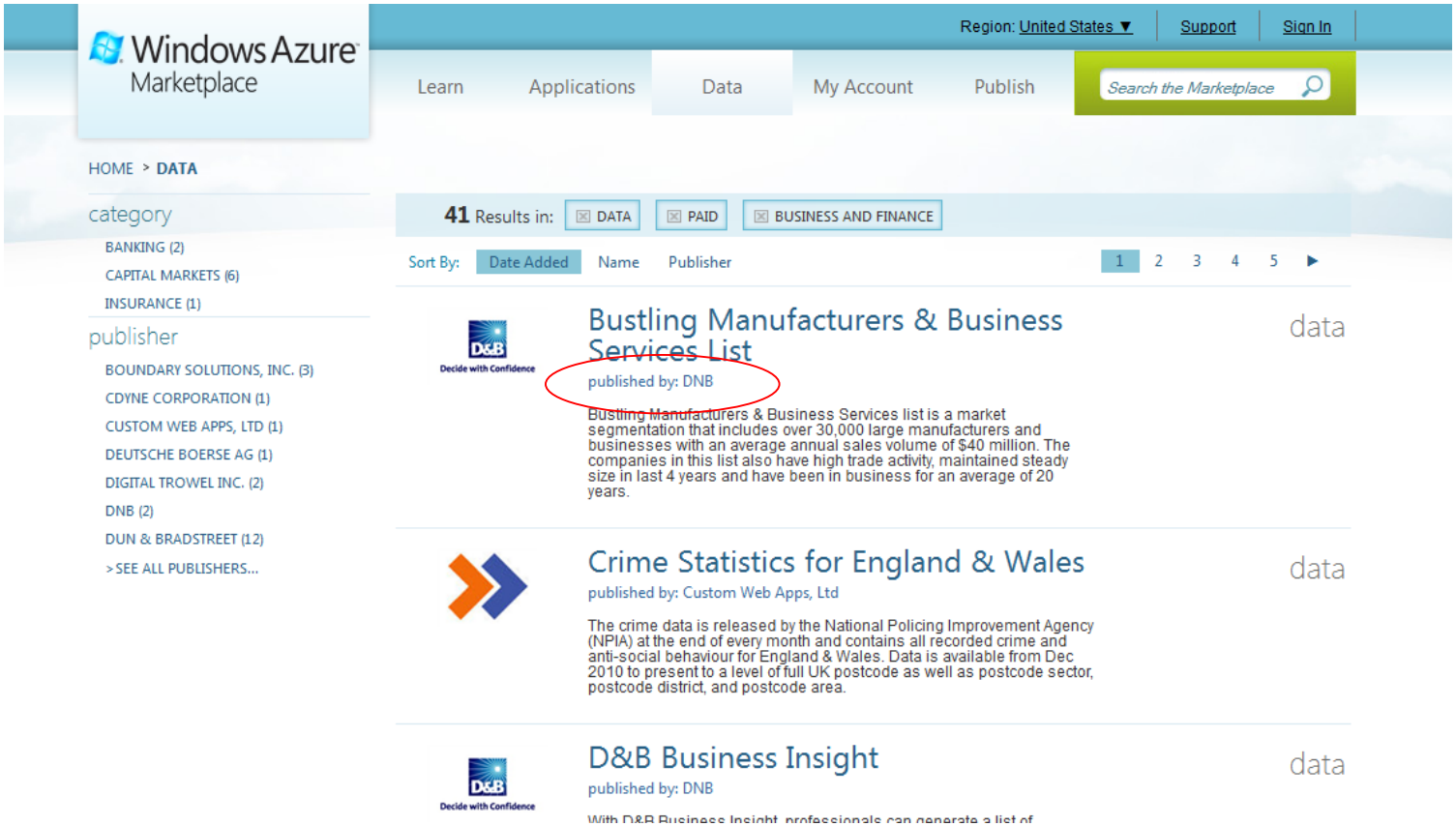
- Read-only DaaS versus CRUD DaaS APIs
- Service APIs versus Data APIs
  - They are not the same wrt data/service concerns
- SOAP versus REST
- Streaming data API

DaaS design & implementation – service provider vs data provider

- The DaaS provider is separated from the data provider

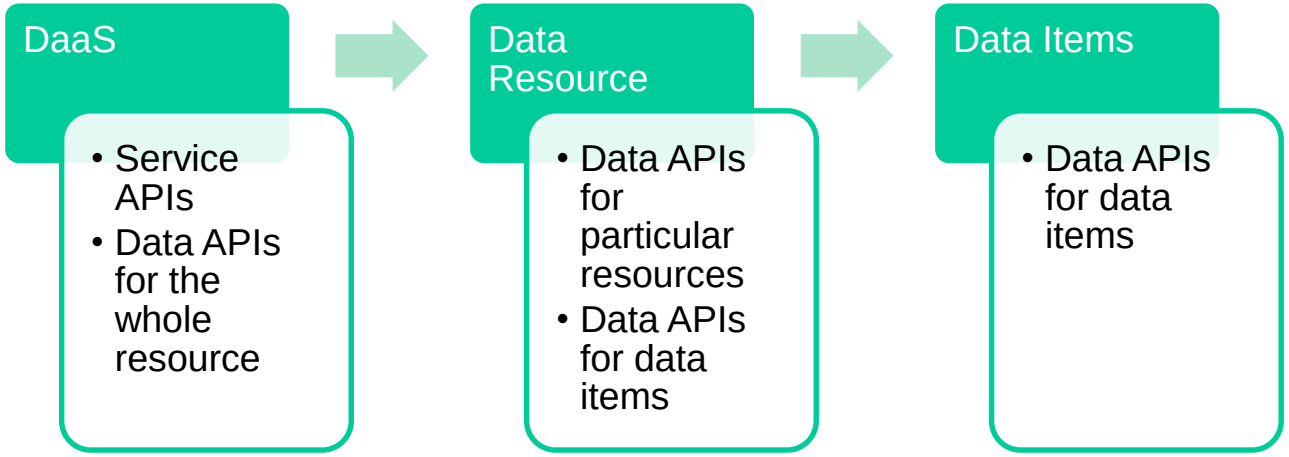


Example: DaaS provider =! data provider



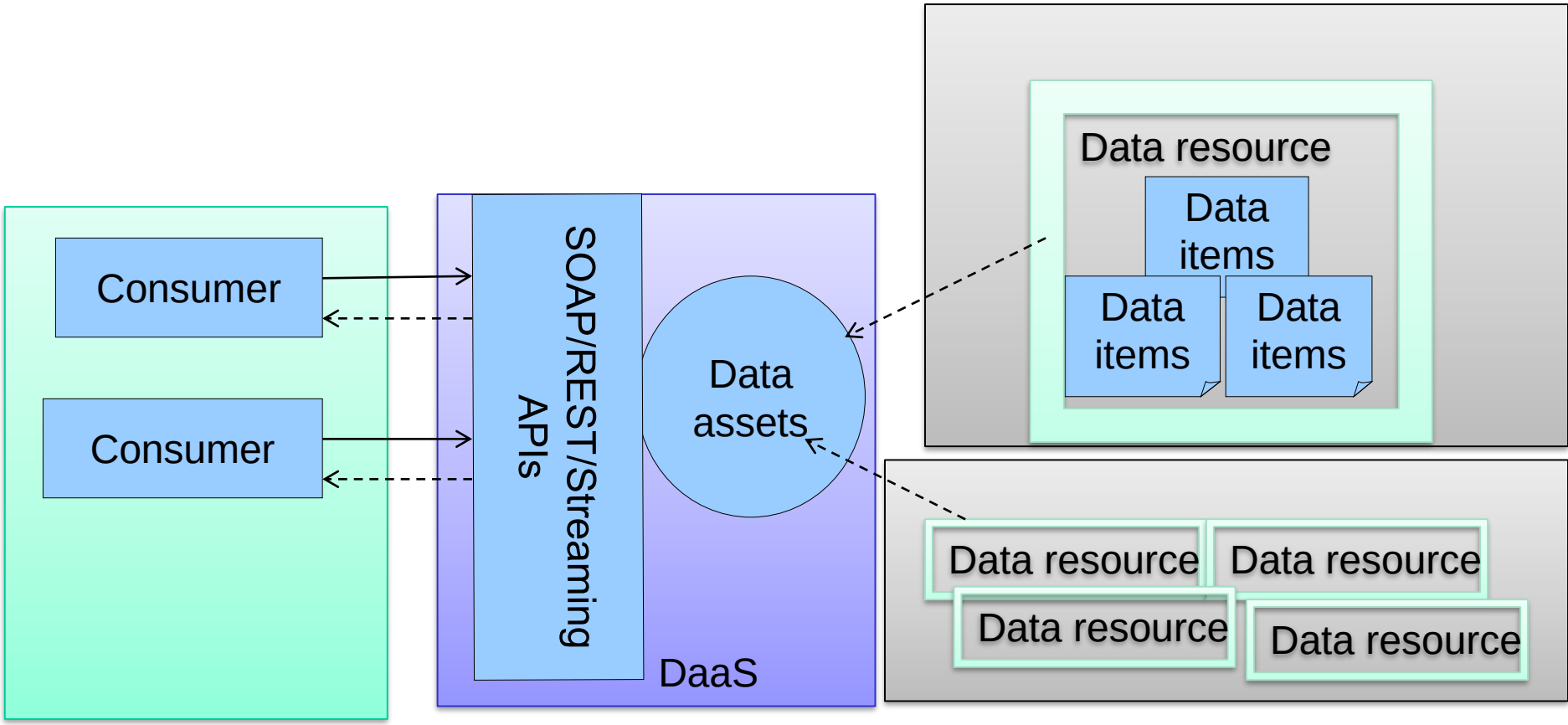
DaaS design & implementation – structures

Three levels

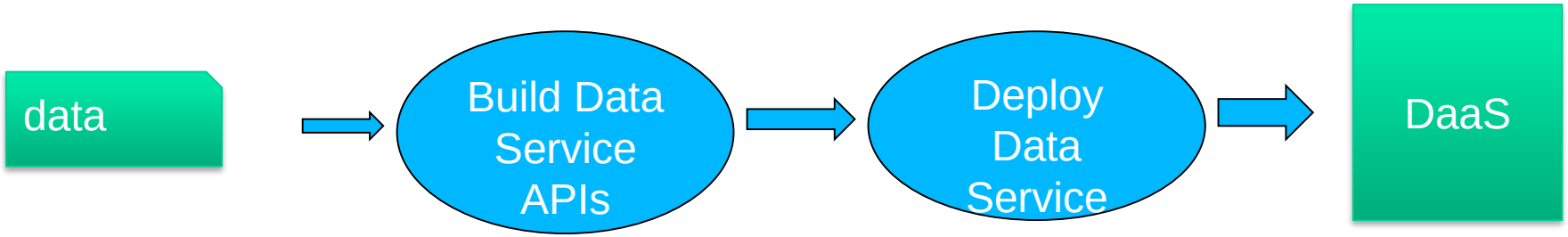


- DaaS and data providers have the right to publish the data

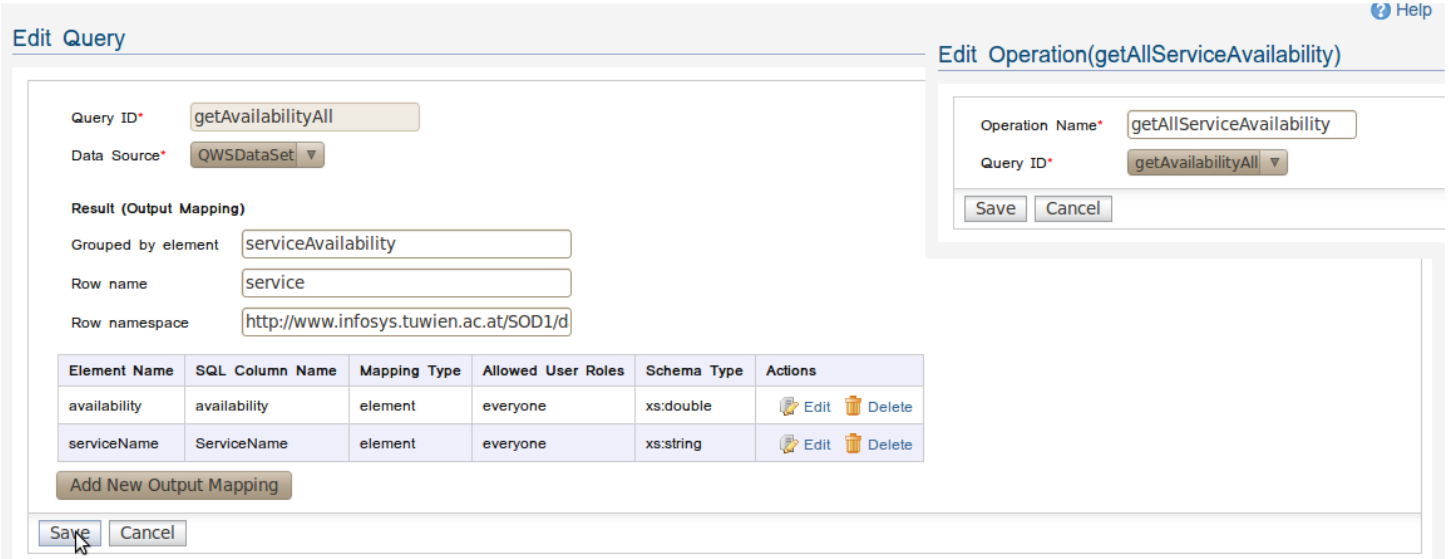
DaaS design & implementation – structures (2)



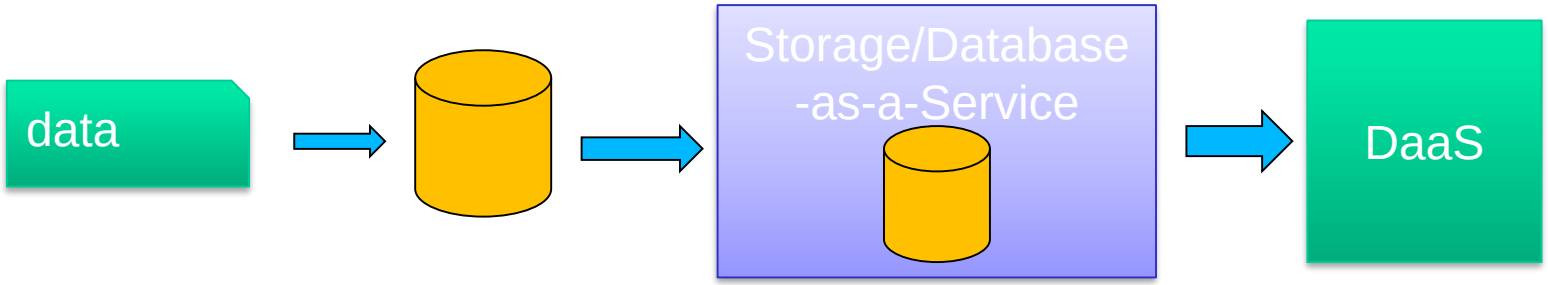
DaaS design & implementation – patterns for „turning data to DaaS“ (1)



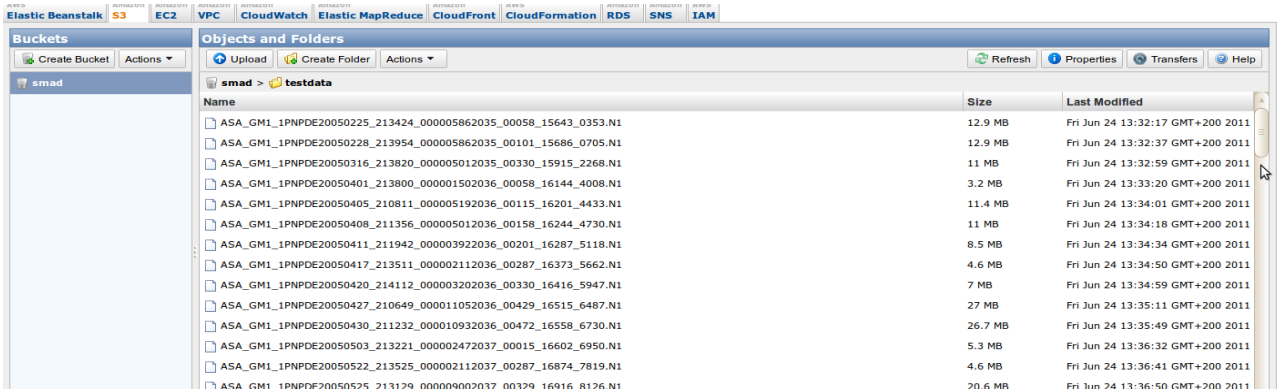
Examples: using WSO2 data service



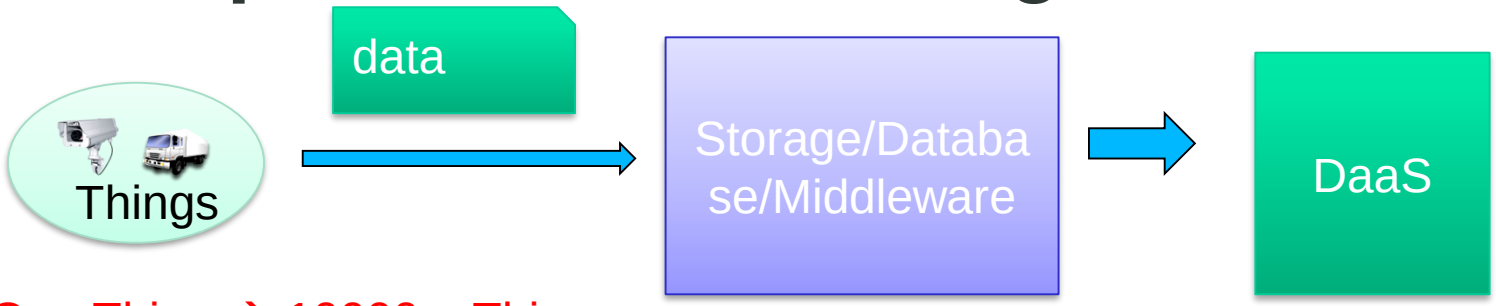
DaaS design & implementation – patterns for „turning data to DaaS“ (2)



Examples: using Amazon S3

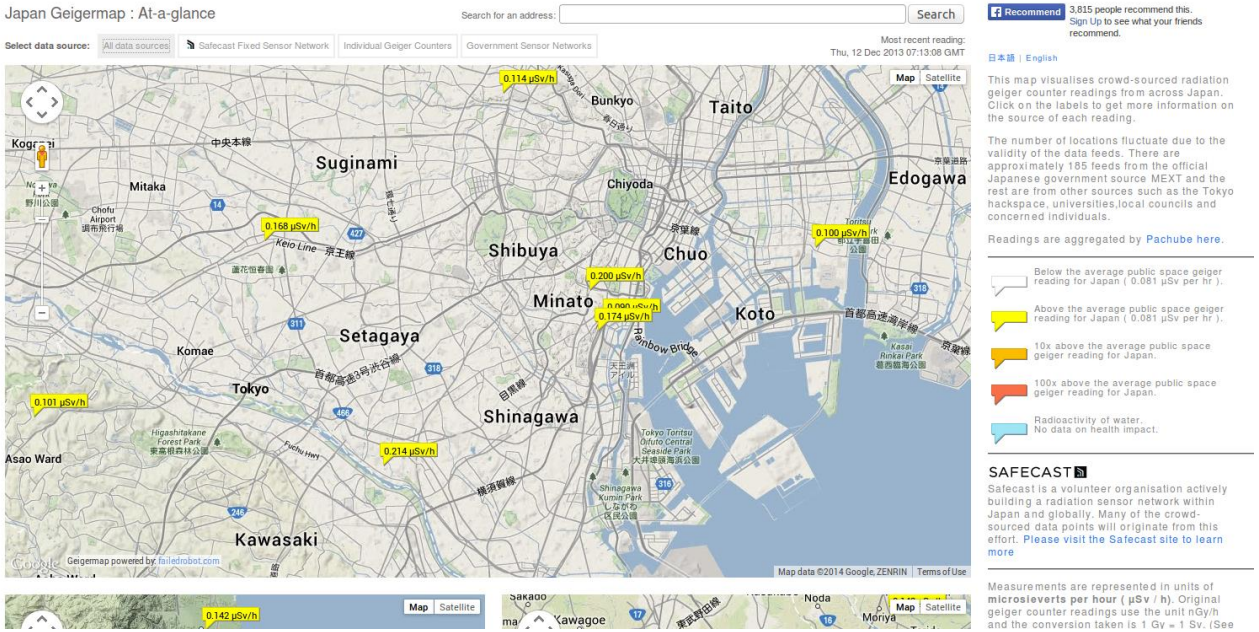


DaaS design & implementation – patterns for „turning data to DaaS“ (3)

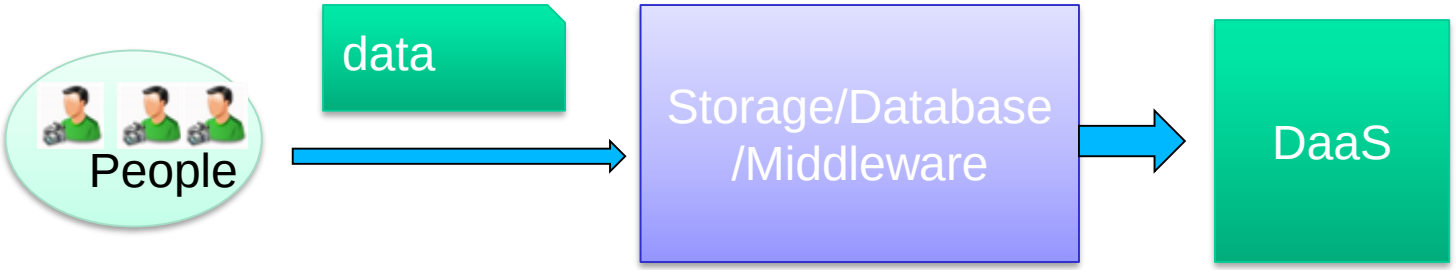


One Thing → 10000... Things

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



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



Examples: using Twitter

Developers API Health Blog Discussions Documentation Search Sign in

Home

REST API v1.1 Resources

Jump to

Timelines

Timelines are collections of Tweets, ordered with the most recent first.

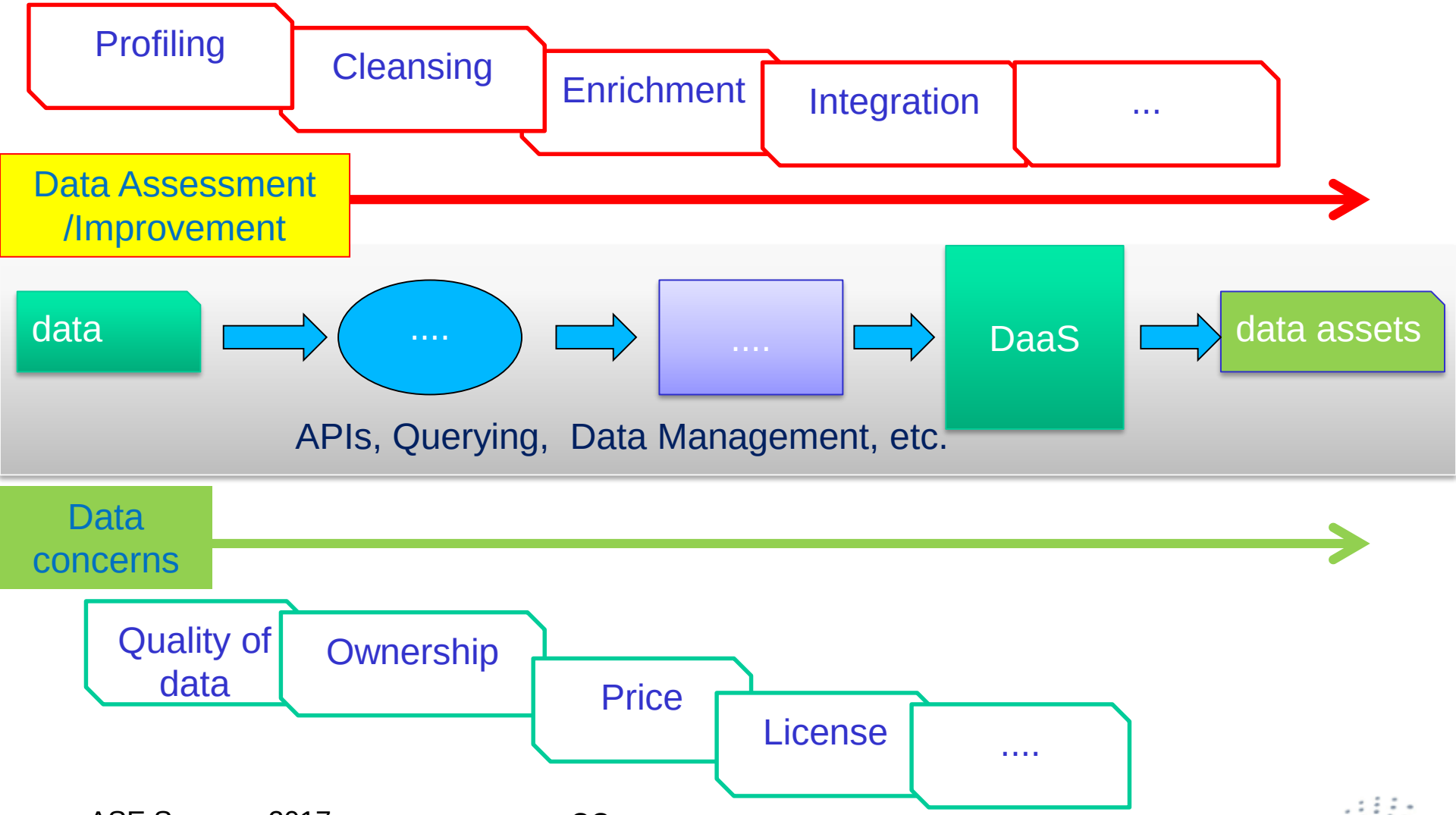
Resource	Description
GET statuses/mentions_timeline	Returns the 20 most recent mentions (tweets containing a users's @screen_name) for the authenticating user. The timeline returned is the equivalent of the one seen when you view your mentions on twitter.com. This method can only return up to 800 tweets. See Working with Timelines for...
GET statuses/user_timeline	Returns a collection of the most recent Tweets posted by the user indicated by the screen_name or user_id parameters. User timelines belonging to protected users may only be requested when the authenticated user either "owns" the timeline or is an approved follower of the owner. The timeline...
GET statuses/home_timeline	Returns a collection of the most recent Tweets and retweets posted by the authenticating user and the users they follow. The home timeline is central to how most users interact with the Twitter service. Up to 800 Tweets are obtainable on the home timeline. It is more volatile for users that follow...
GET statuses/retweets_of_me	Returns the most recent tweets authored by the authenticating user that have been retweeted by others. This timeline is a subset of the user's GET statuses/user_timeline. See Working with Timelines for instructions on traversing timelines.

Tweets

Tweets are the atomic building blocks of Twitter, 140-character status updates with additional associated metadata. People tweet for a variety of reasons about a multitude of topics.

Resource	Description
GET statuses/retweets/:id	Returns a collection of the 100 most recent retweets of the tweet specified by the id parameter.

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

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

Data governance Process

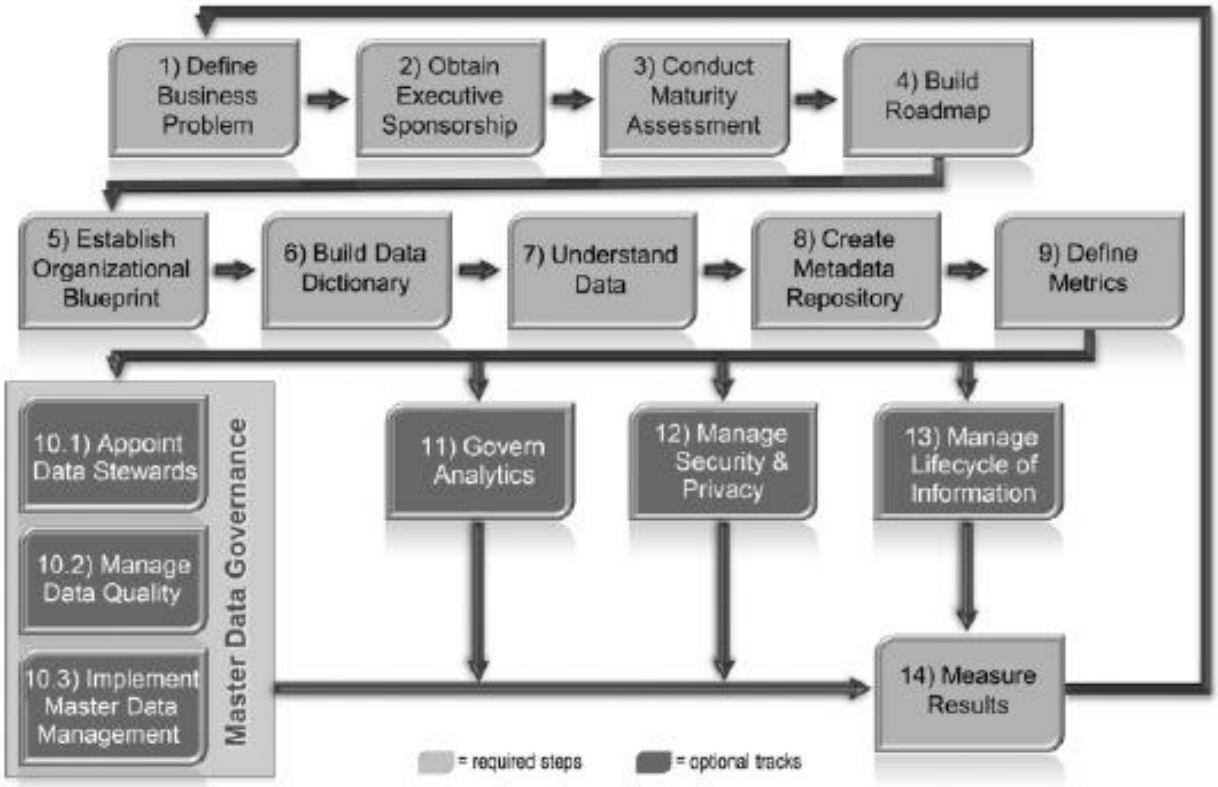


Figure 2.1: An overview of the IBM Data Governance Unified Process.

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.[10]

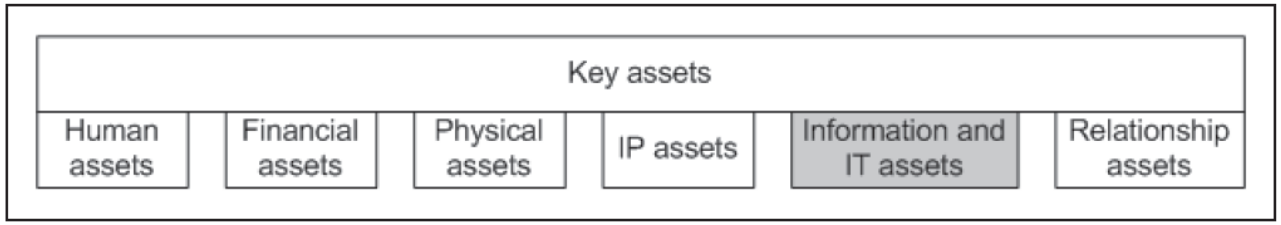
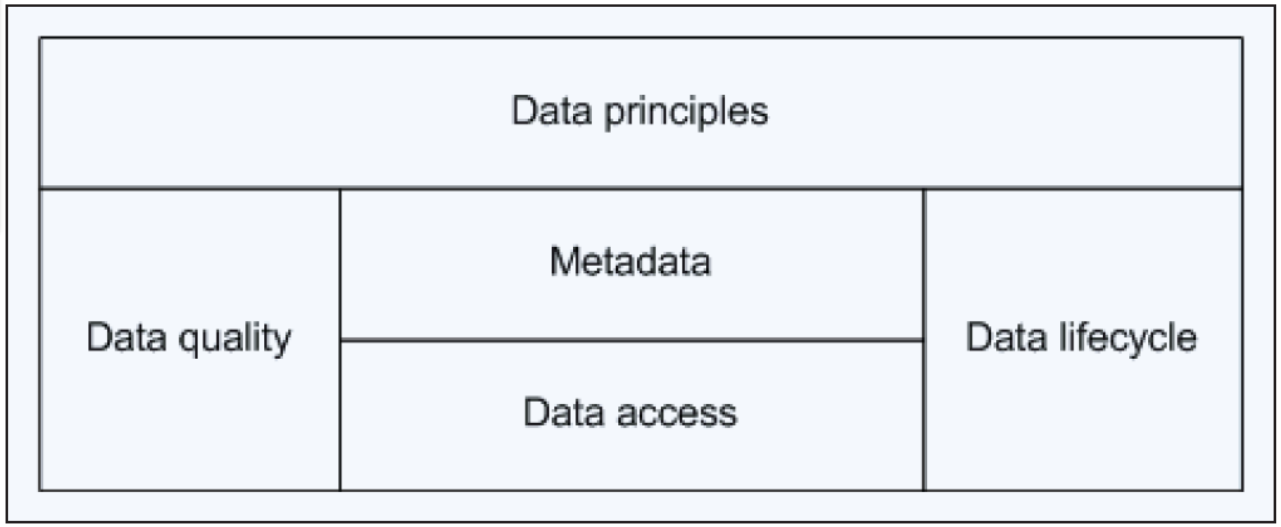


Figure 2: Decision domains for data governance.



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

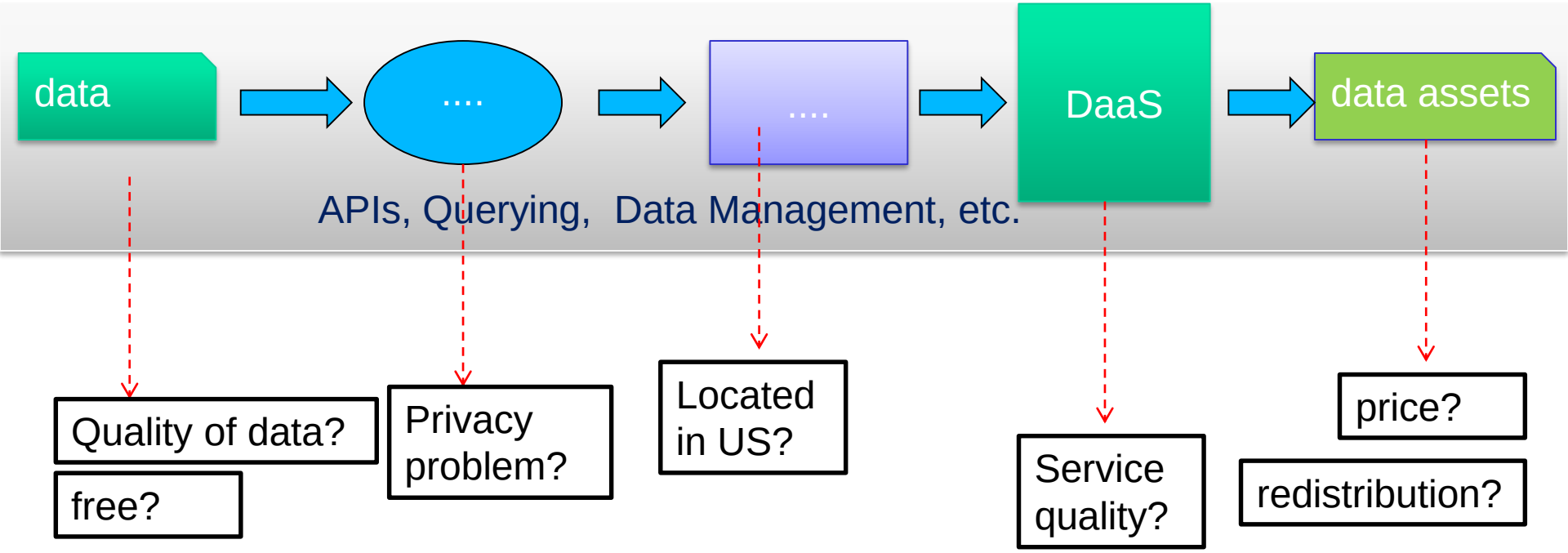
Vijay Khatri and Carol V. Brown. 2010. Designing data governance. Commun. ACM 53, 1 (January 2010), 148-152. DOI=http://dx.doi.org/10.1145/1629175.1629210

Table 1: Framework for data decision domains.

Data Governance Domains	Domain Decisions	Potential Roles or Locus of Accountability
Data Principles • Clarifying the role of data as an asset	• What are the uses of data for the business? • What are the mechanisms for communicating business uses of data on an ongoing basis? • What are the desirable behaviors for employing data as assets? • How are opportunities for sharing and reuse of data identified? • How does the regulatory environment influence the business uses of data?	• Data owner/trustee • Data custodian • Data steward • Data producer/supplier • Data consumer • Enterprise Data Committee/ Council
Data Quality • Establishing the requirements of intended use of data	• What are the standards for data quality with respect to accuracy, timeliness, completeness and credibility? • What is the program for establishing and communicating data quality? • How will data quality as well as the associated program be evaluated?	• Data owner • Subject matter expert • Data quality manager • Data quality analyst
Metadata • Establishing the semantics or "content" of data so that it is interpretable by the users	• What is the program for documenting the semantics of data? • How will data be consistently defined and modeled so that it is interpretable? • What is the plan to keep different types of metadata up-to-date?	• Enterprise data architect • Enterprise data modeler • Data modeling engineer • Data architect • Enterprise Architecture Committee
Data Access • Specifying access requirements of data	• What is the business value of data? • How will risk assessment be conducted on an ongoing basis? • How will assessment results be integrated with the overall compliance monitoring efforts? • What are data access standards and procedures? • What is the program for periodic monitoring and audit for compliance? • How is security awareness and education disseminated? • What is the program for backup and recovery?	• Data owner • Data beneficiary • Chief information security officer • Data security officer • Technical security analyst • Enterprise Architecture Development Committee
Data Lifecycle • Determining the definition, production, retention and retirement of data	• How is data inventoried? • What is the program for data definition, production, retention, and retirement for different types of data? • How do the compliance issues related to legislation affect data retention and archiving?	• Enterprise data architect • Information chain manager

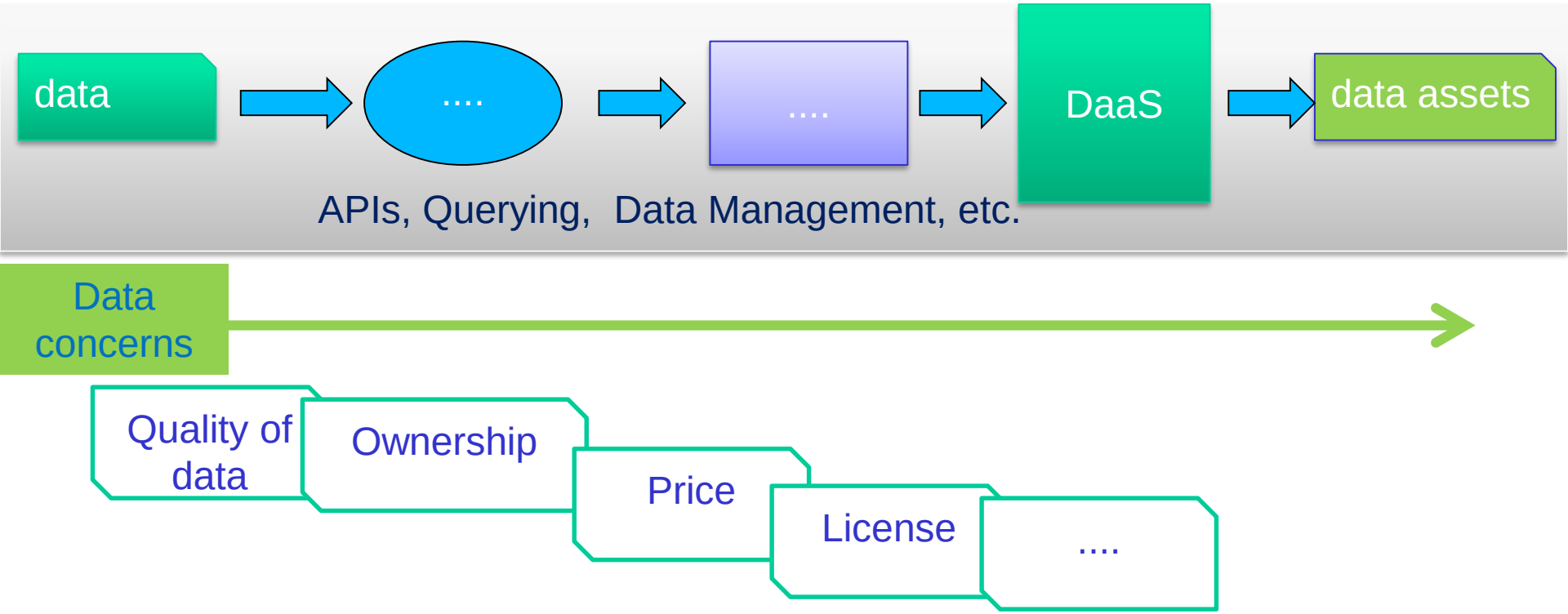
DATA CONCERNS

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?

- Too much 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

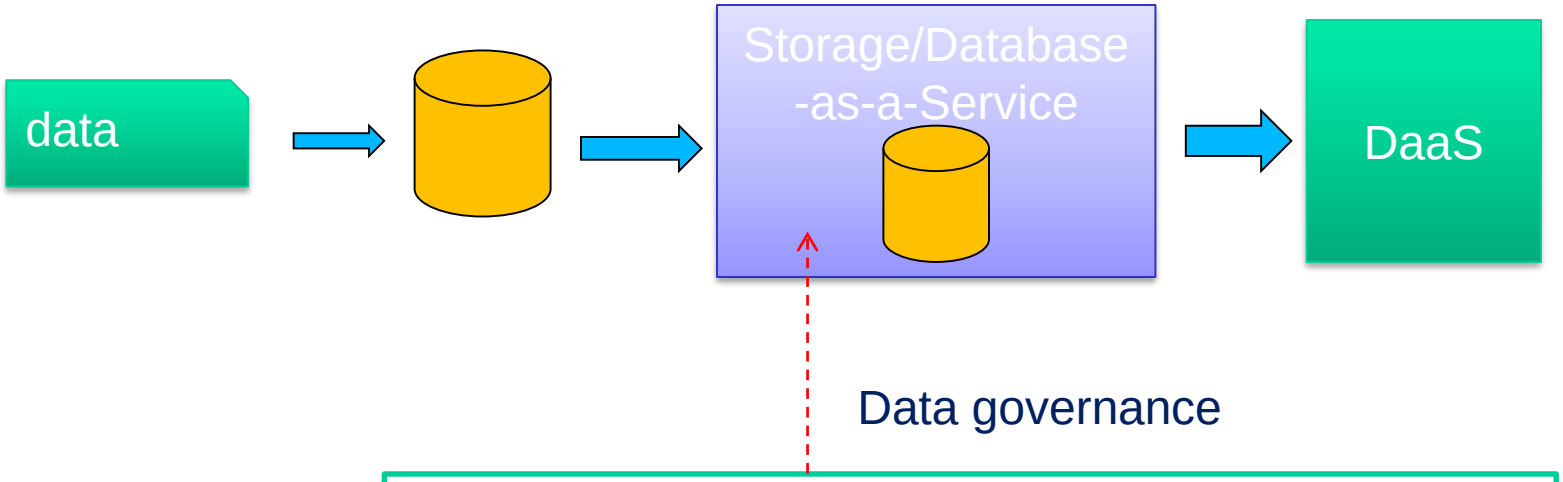
Ultimate goal: to provide *relevant* data with *acceptable constraints on data concerns in different provisioning models*

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



Important factor, for example, the security and privacy compliance, data distribution, and auditing

Quality of data

Read-only DaaS

- Important factor for the selection of DaaS.
- For example, the accurary and compleness of the data, whether the data is up-to-date

CRUD DaaS

- Expected some support to control the quality of the data in case the data is offered to other consumers

Data and service usage

Read-only DaaS

- Important factor, in particular, price, data and service APIs licensing, law enforcement, and Intellectual Property rights

CRUD DaaS

- Important factor, in paricular, price, service APIs licensing, and law enforcement

Quality of service

Read-only DaaS

- Important factor, in particular availability and response time

CRUD Daas

- Important factor, in particular, availability, response time, dependability, and security

Contextual information

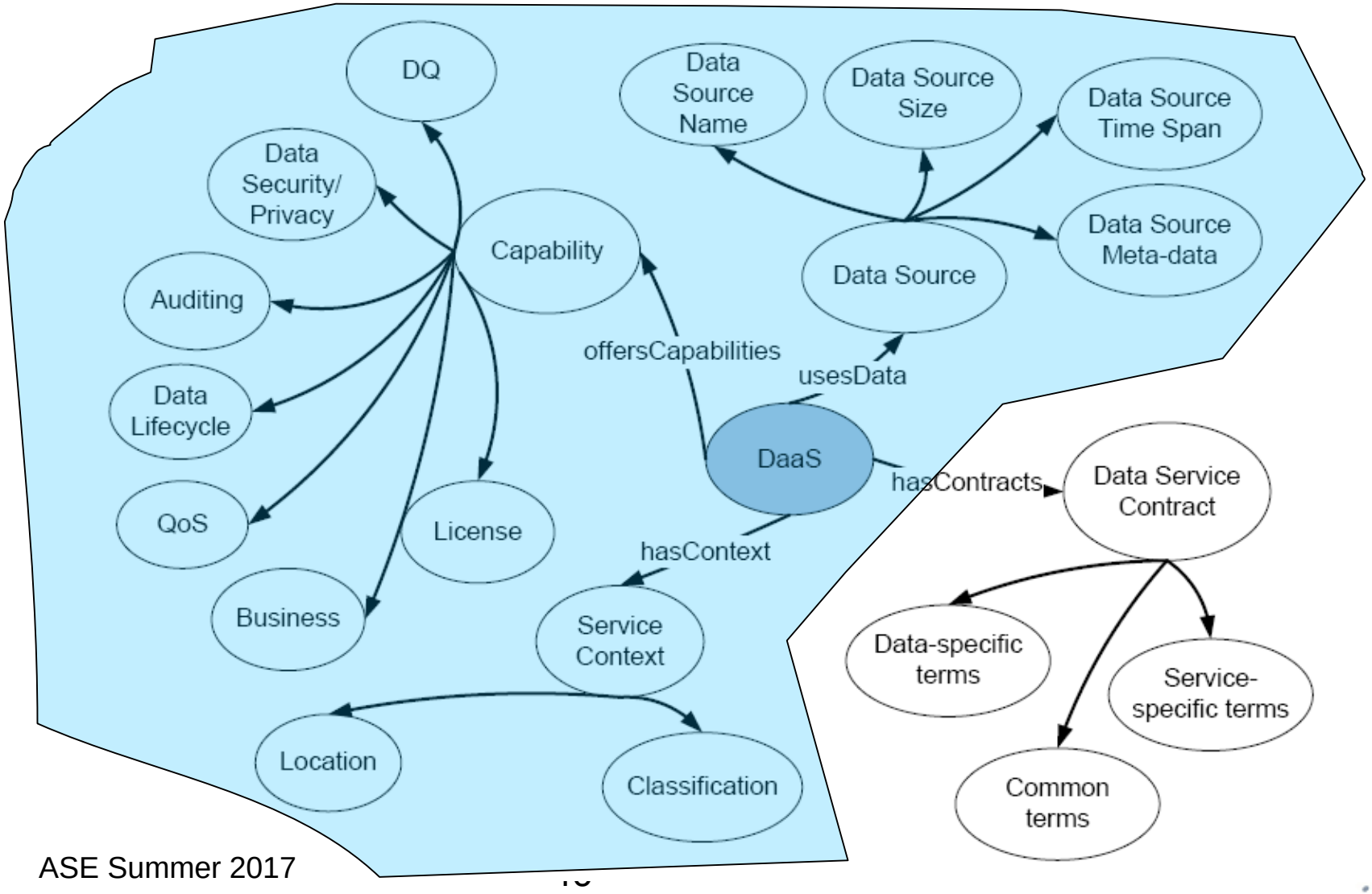
Read-only DaaS

- Useful factor, such as classification and service type (REST, SOAP), location

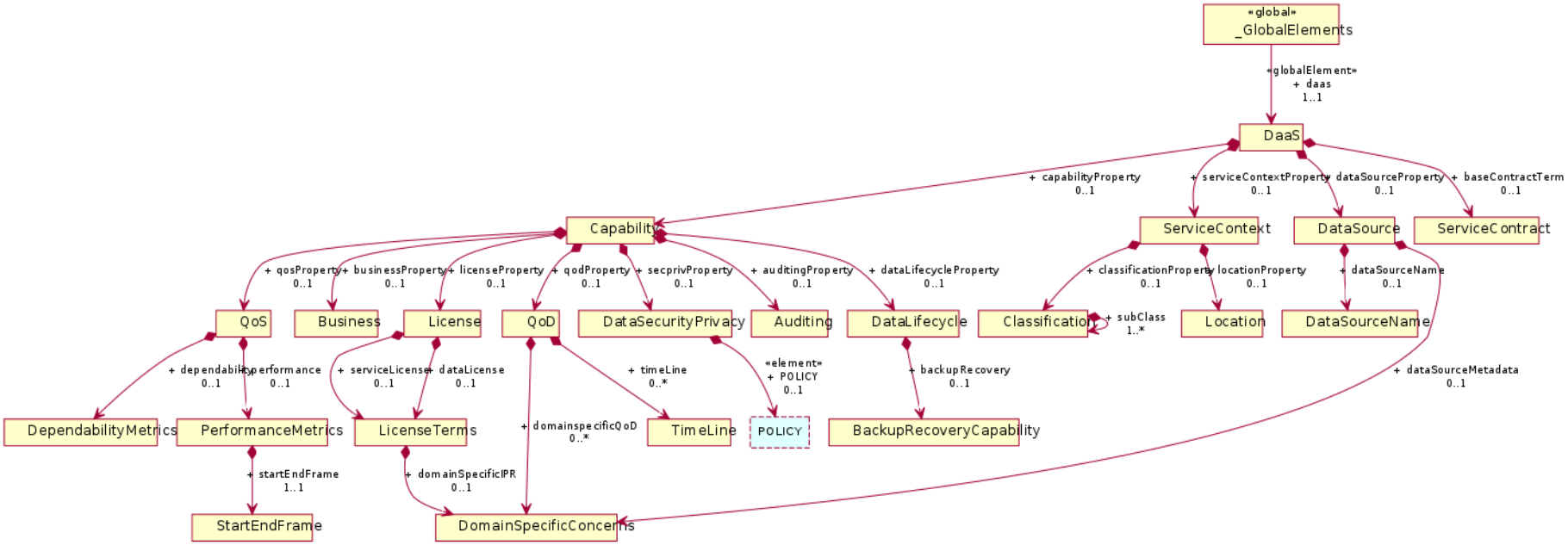
CRUD DaaS

- Important factor, e.g. location (for regulation compliance) and versioning

Conceptual model for DaaS concerns and contracts



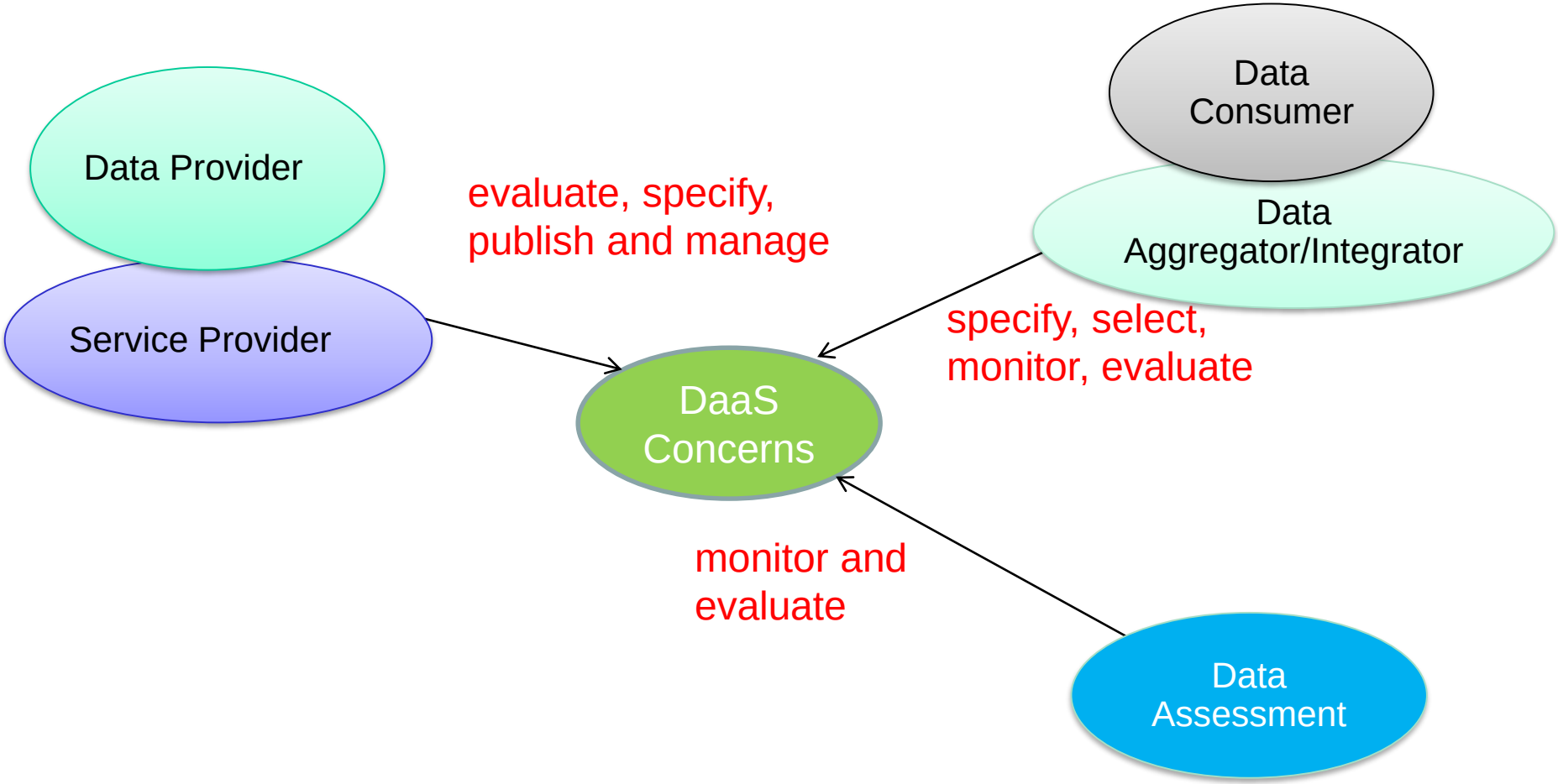
Implementation (1)



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

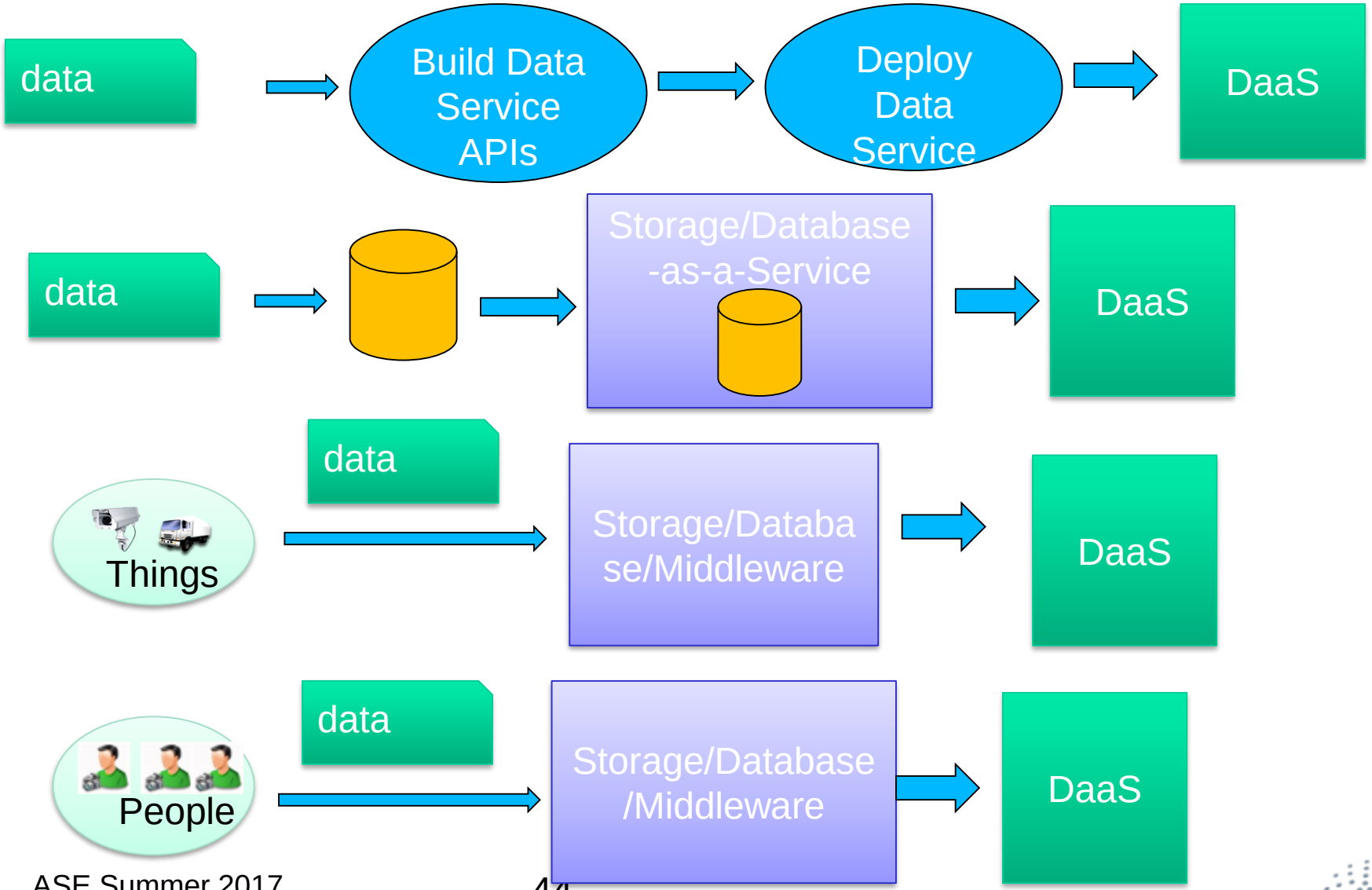
Populating DaaS concerns

The role of stakeholders in the most trivial view



HOW TO EVALUATE DATA CONCENRS FOR DATA ASSETS IN DAAS?

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

Evaluating data concerns → Describing data concerns → Populating data concerns

What do we need in order to perform these activities?

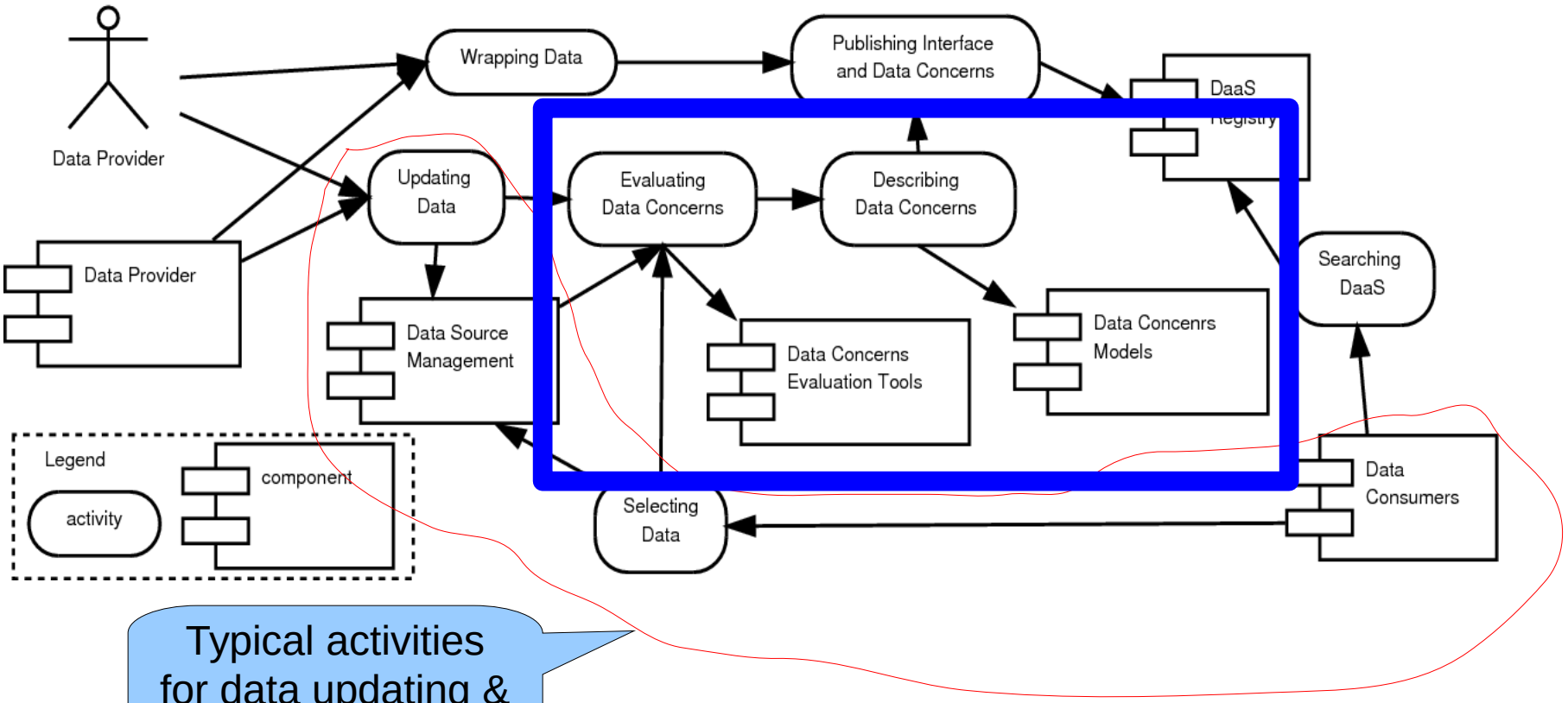
Data Concerns Evaluation Tools

Data Concerns Representation Models

Publishing services

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?

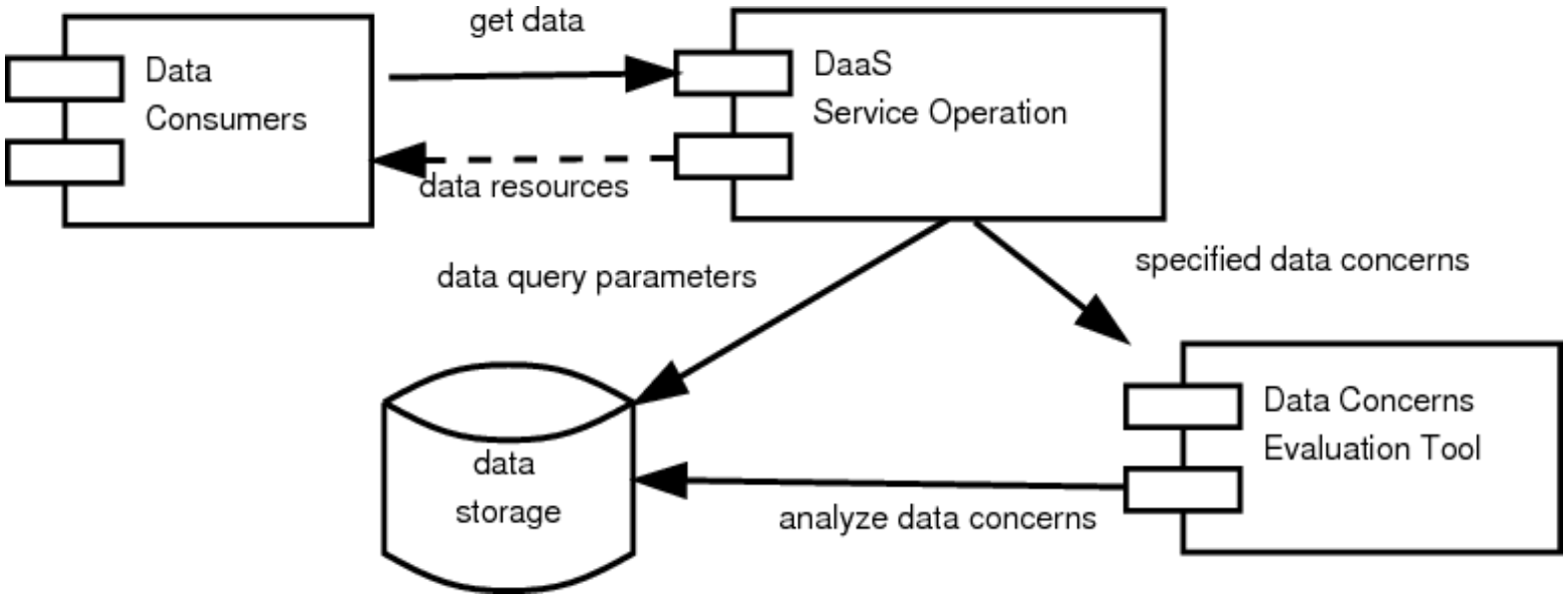
evaluation modes
- When the evaluation is done?

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

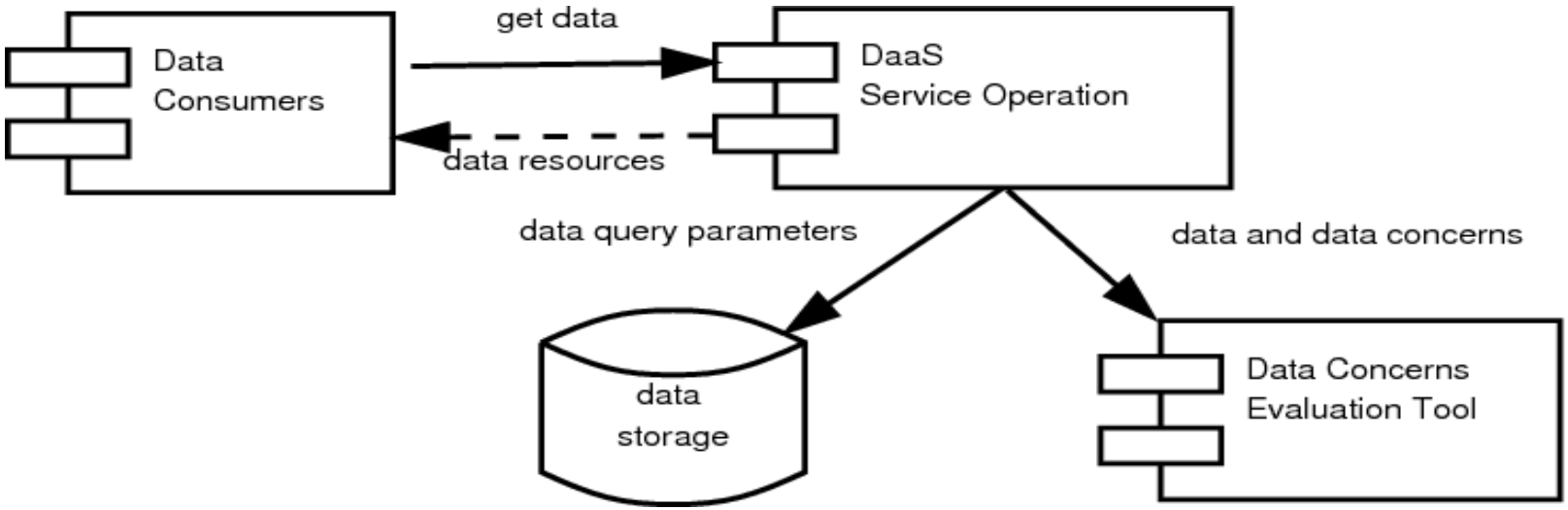
Evaluating data concerns – some patterns (1)

Pull, pass-by-references



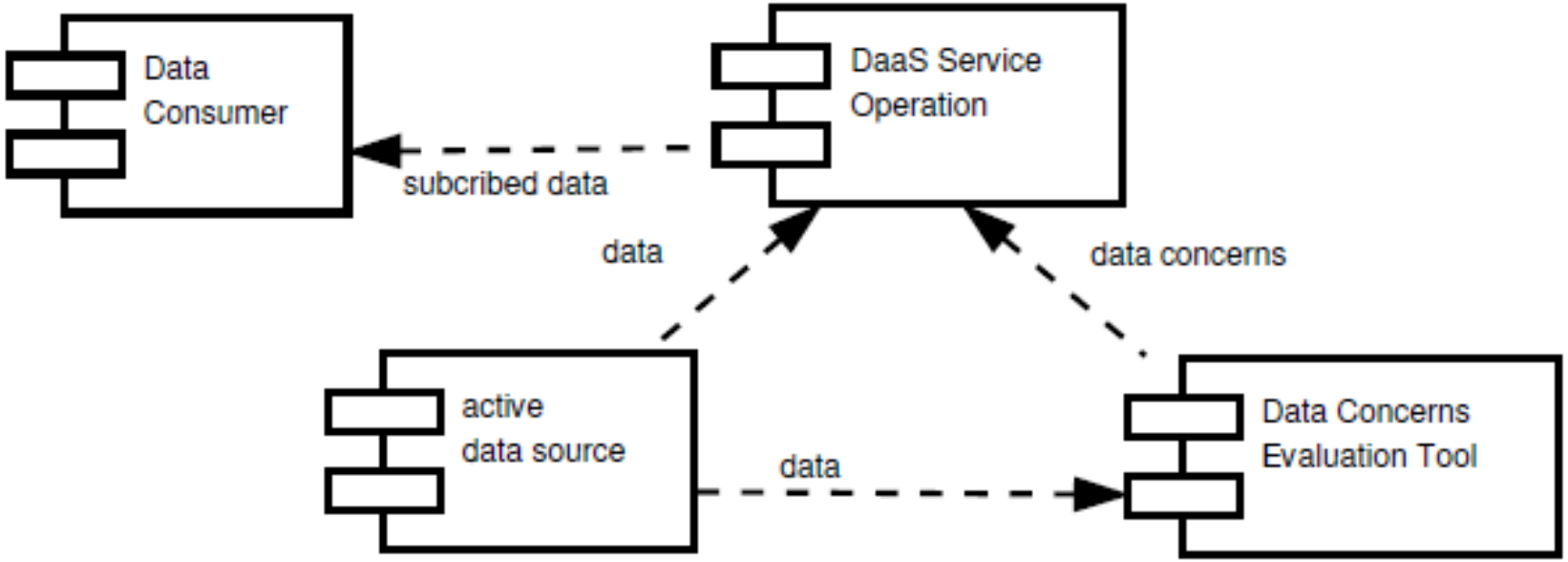
Evaluating data concerns – some patterns (2)

Pull, pass-by-values



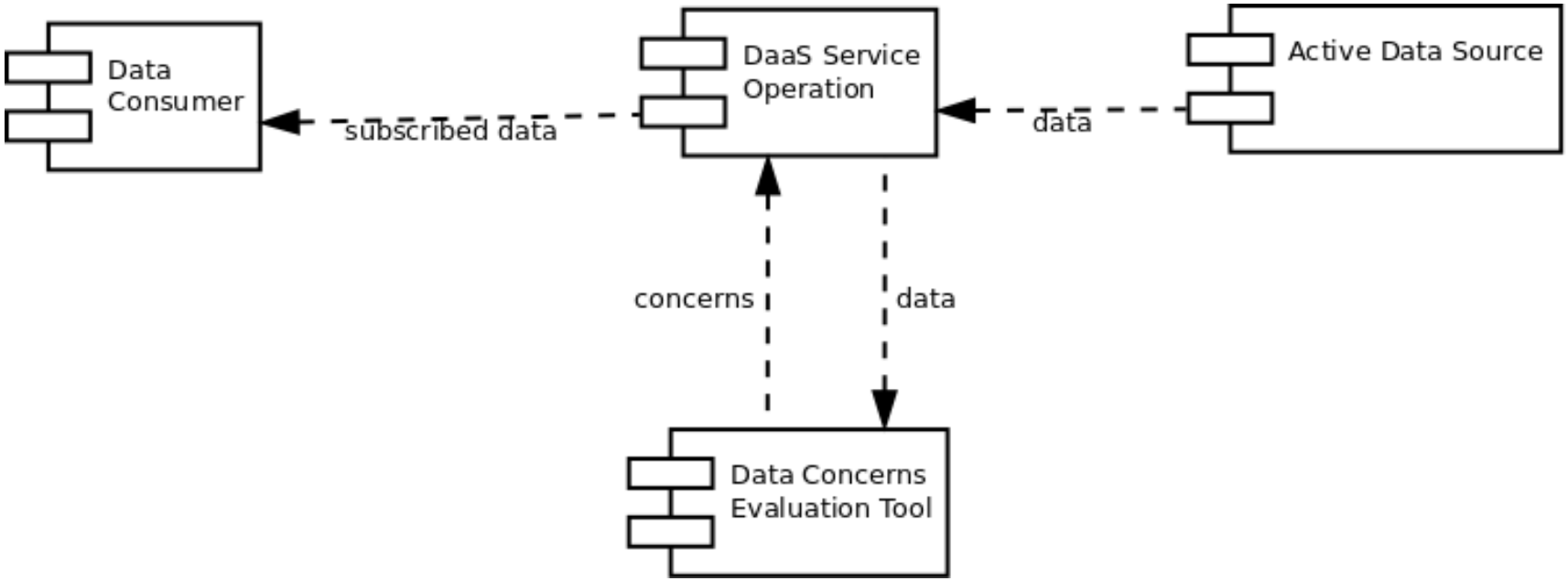
Evaluating data concerns – some patterns (3)

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

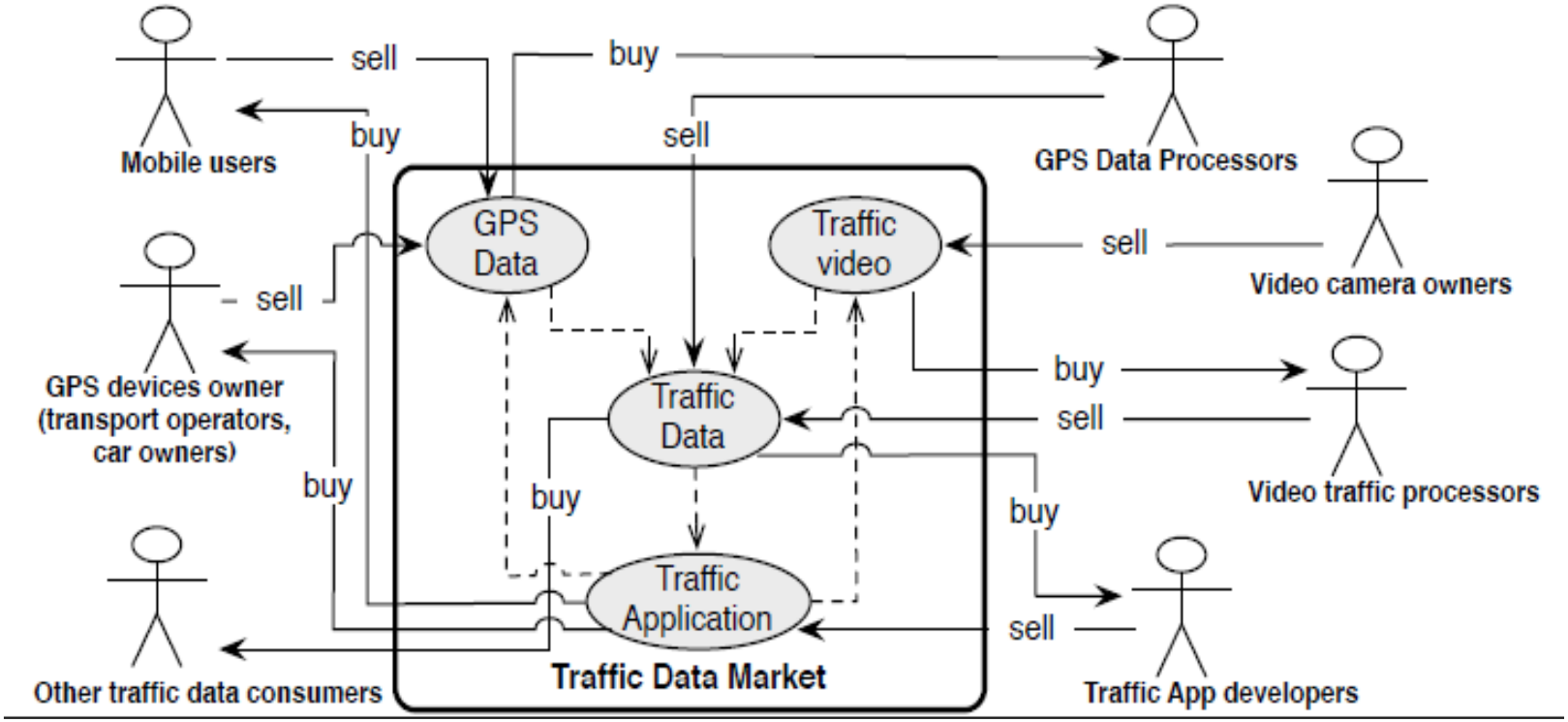
Óscar Figuerola Salas, Velibor Adzic, Akash Shah, and Hari Kalva. 2013. Assessing internet video quality using crowdsourcing. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (CrowdMM '13). ACM, New York, NY, USA, 23-28. DOI=10.1145/2506364.2506366 http://doi.acm.org/10.1145/2506364.2506366

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/

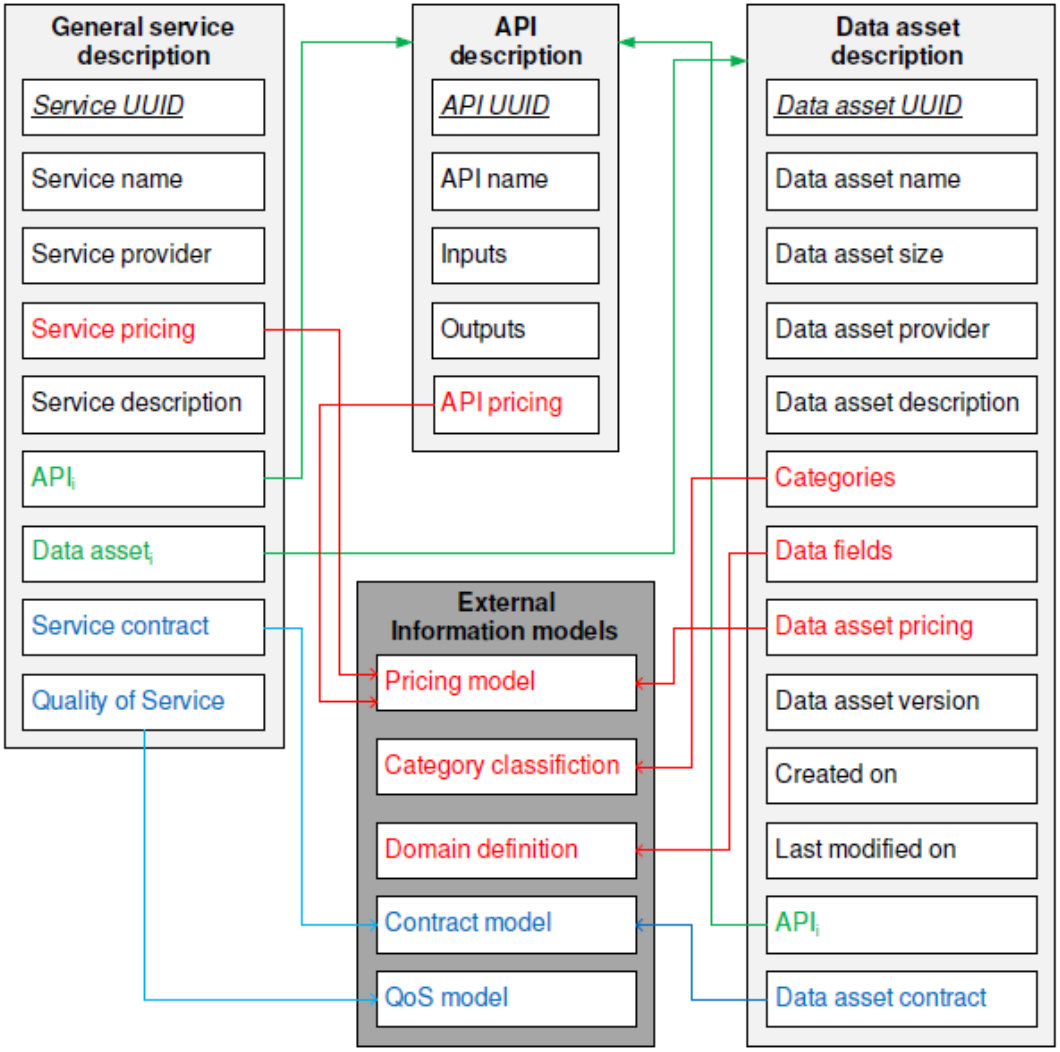
Specific data market or generic data market?

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

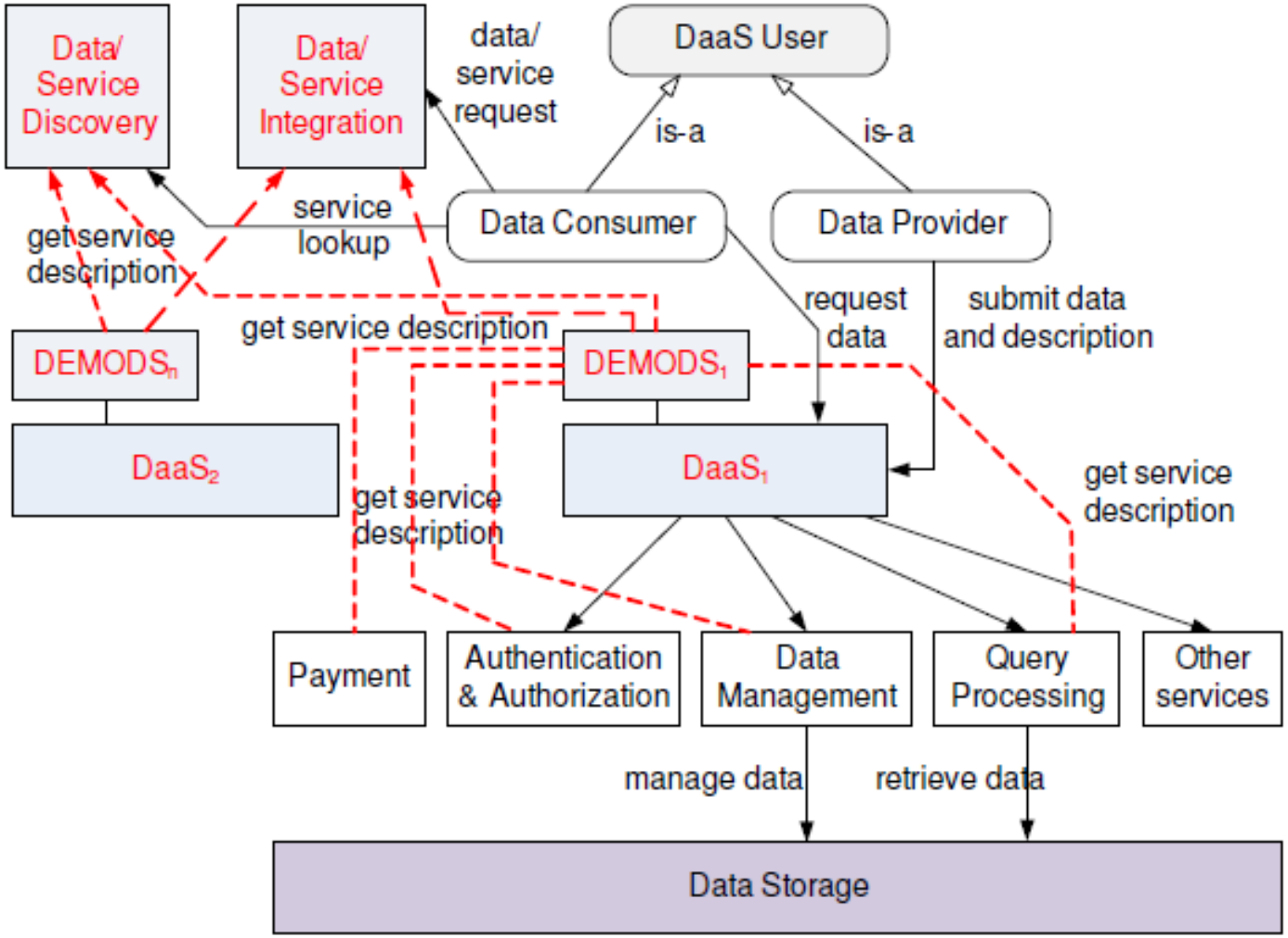
DEMODS – a description model for Data-as-a-Service

Quang Hieu Vu, Tran Vu Pham, Hong Linh Truong,, Schahram Dustdar, Rasool Asal: DEMODS: A Description Model for Data-as-a-Service. AINA 2012: 605-612



See prototype:
http://www.infosys.tuwien.ac.at/prototype/SOD1/demods/

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

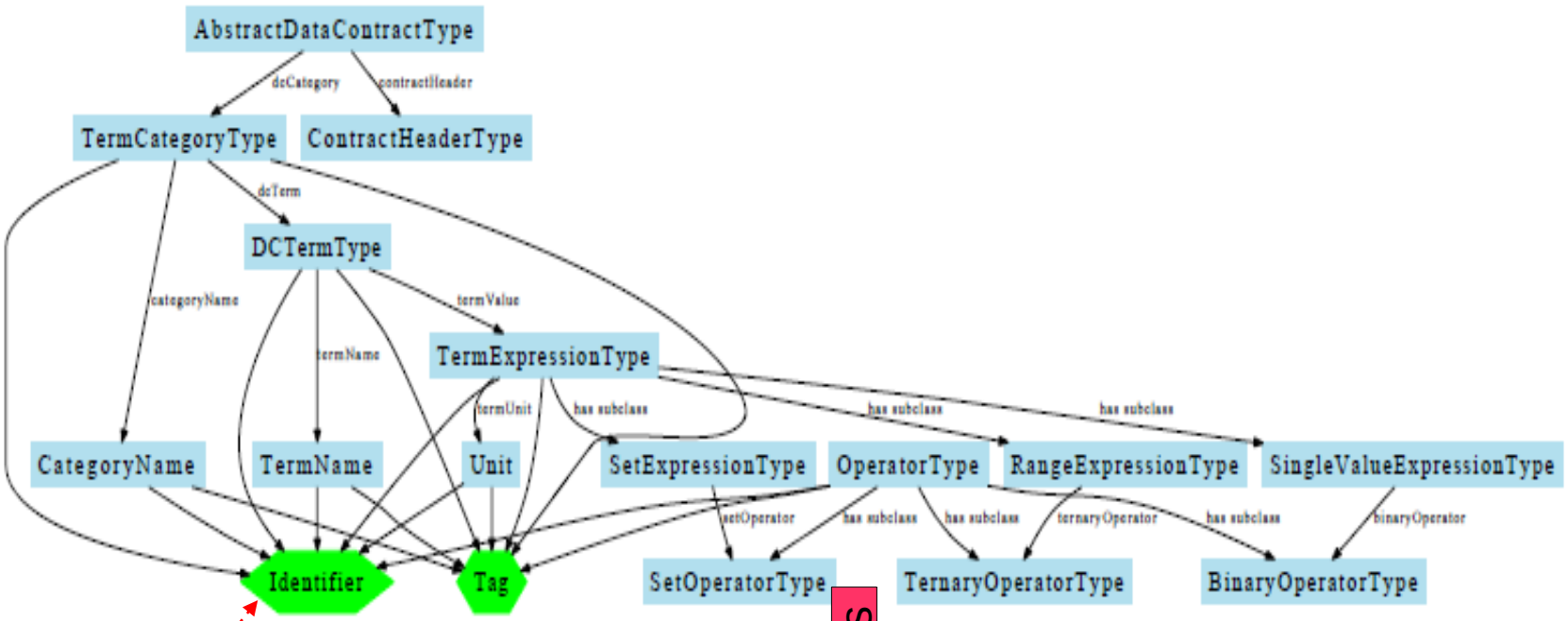
Most information is in human-readable form

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

Category	*Term representation*	*Examples*
Data rights	$termName = \{val_1, val_2, \cdots, val_n\}$	$termName$={Derivation, Collection, Reproduction, Attribution, Noncommercialuse}, val_i = {Undefined, Null, Allowed, Required, True, False}
Quality of data	$val_l \leq termName \leq val_u$	$termName$={Accuracy, Completeness, Uptodateness}, val_l and $val_u \in [0,1]$
Compliance	$termName = \{val_1, val_2, \cdots, val_n\}$	$termName$ and val_i are any string, e.g., $termName$={PrivacyCompliance} and $termValue$={Sarbanes-Oxley (SOX) Act}
Pricing model	$termName = (cost = val_1, usagetime = val_2, , maximumuse = val_3)$	$termName$ is any string, e.g., MonthlyPayment; $val_1 \in R$, e.g., $cost = 50$ €, $val_2 = \{(end_t - start_t); UNLIMITED\}$ where $end_t, start_t \in datetime$, e.g., $usagetime = 30$ days; $val_3 \in N$, e.g, $maximumuse = 1,000$ calls
Control and relationship	$termName = val$	$termName$ and val are any string, e.g., $termName$={Liability, LawandJurisdiction} and $val = \{US, Austria\}$

Structuring abstract data contracts



Use Identifiers and Tags for identifying and searches

generates

Concrete data contracts can be in RDF, XML or JSON

Discussion time

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

Data Market without Marketplace?

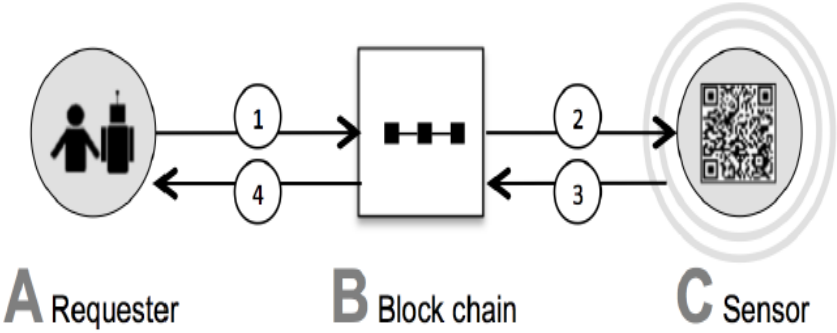


Fig. 1. Schema for the atomic S^2aaS process of exchanging a single datum for cash using Bitcoin.

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

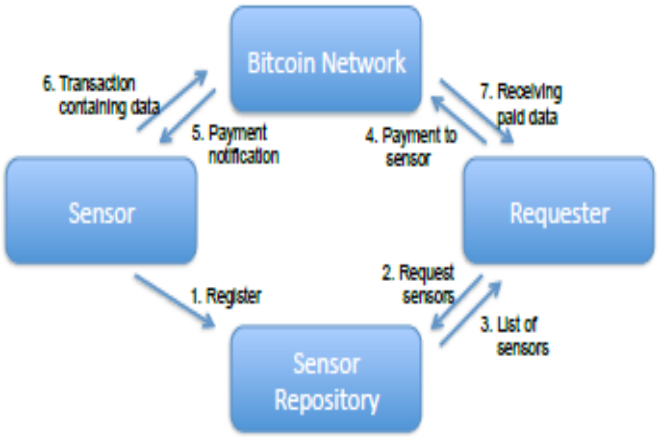


Figure 1: Process for exchanging data for bitcoin.

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.

But what about data contract? → smart contract

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

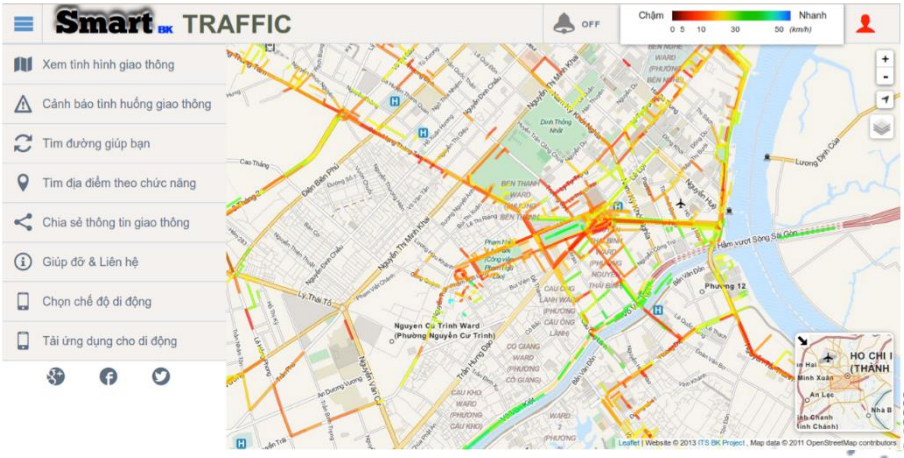
http://dungcao.github.io/marsa/

Traffic problems in HoChiMinh City

- Crowded and unpredictable
- Needs a lot of data to understand traffics

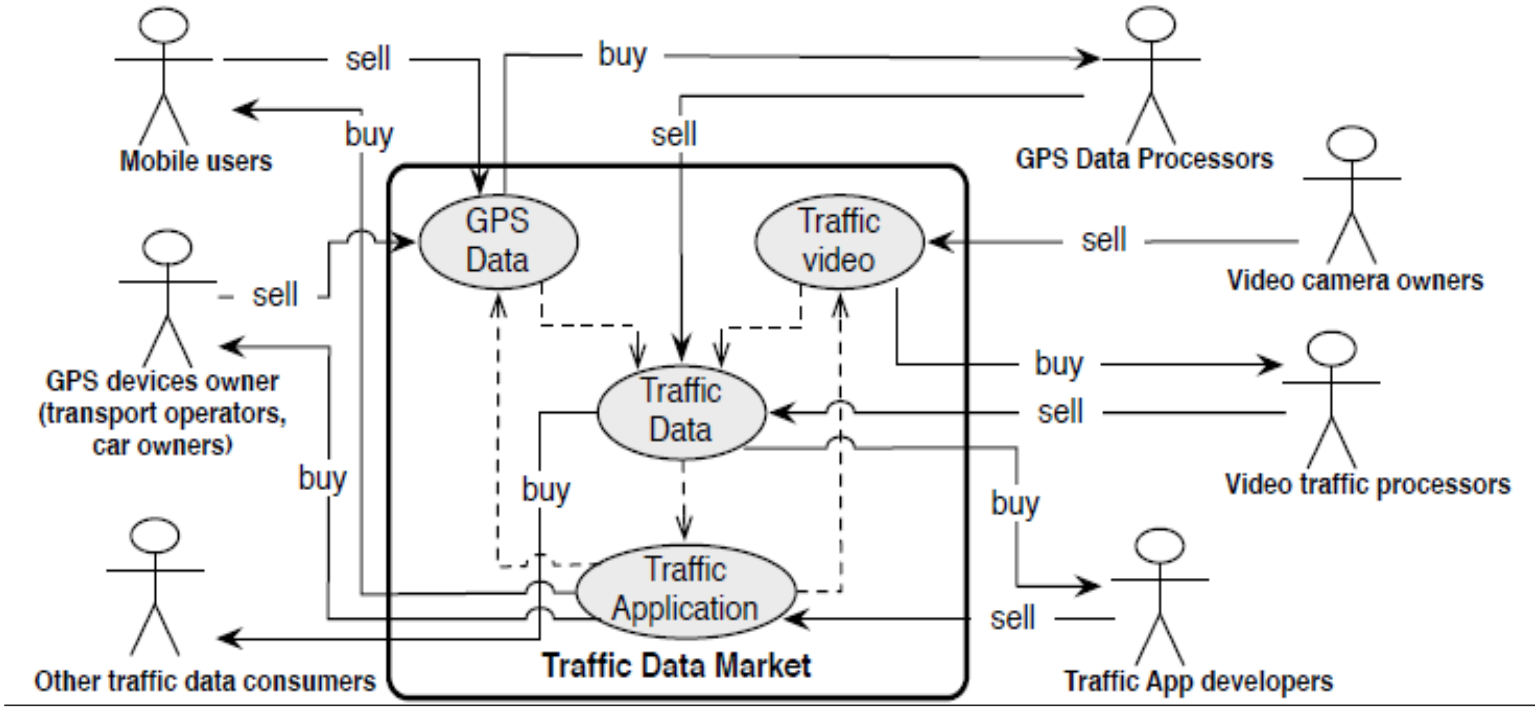
- Lack infrastructures for collecting traffic information
- Common problems in developing countries

Figure sources: Internet



Cannot buy expensive traffic data collection systems!

Market-oriented View of traffic data scenarios



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

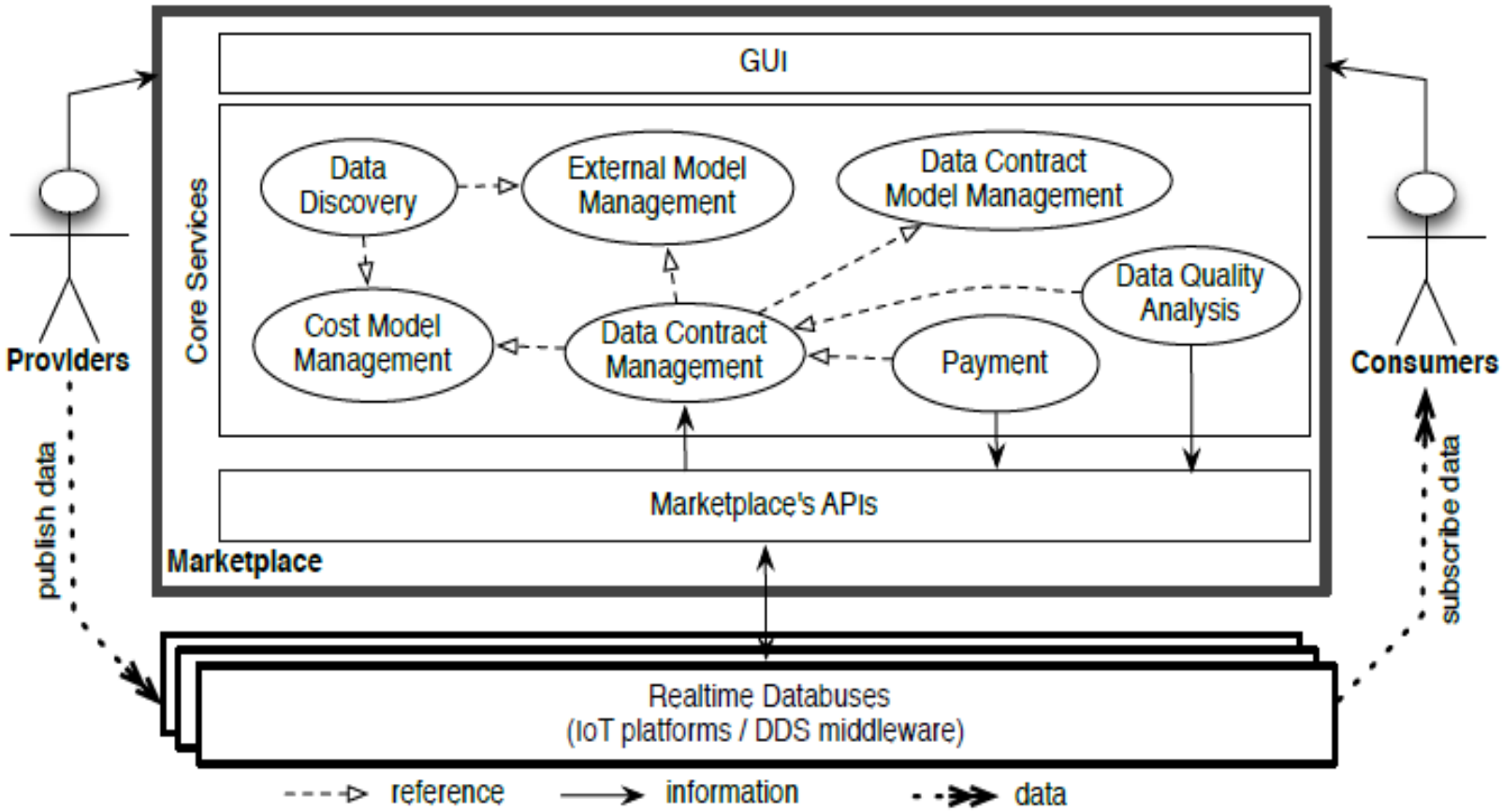
1MB of GPS data =20 USD cent → 6000 USD for the fleet operators

A mobile phone, like a bus, can receive 1.5 USD per month → ½ of 3G data bill

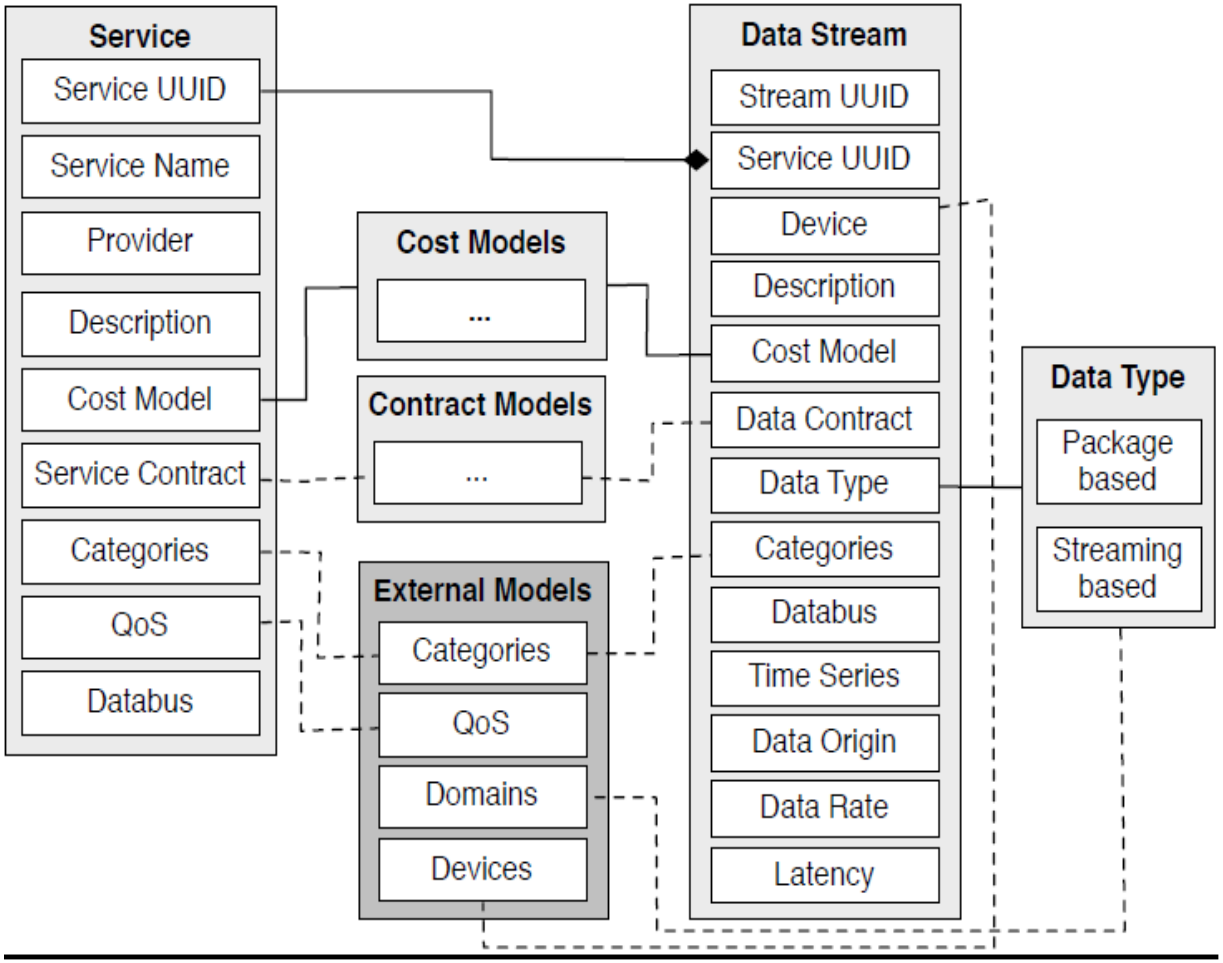
Costs and benefits

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

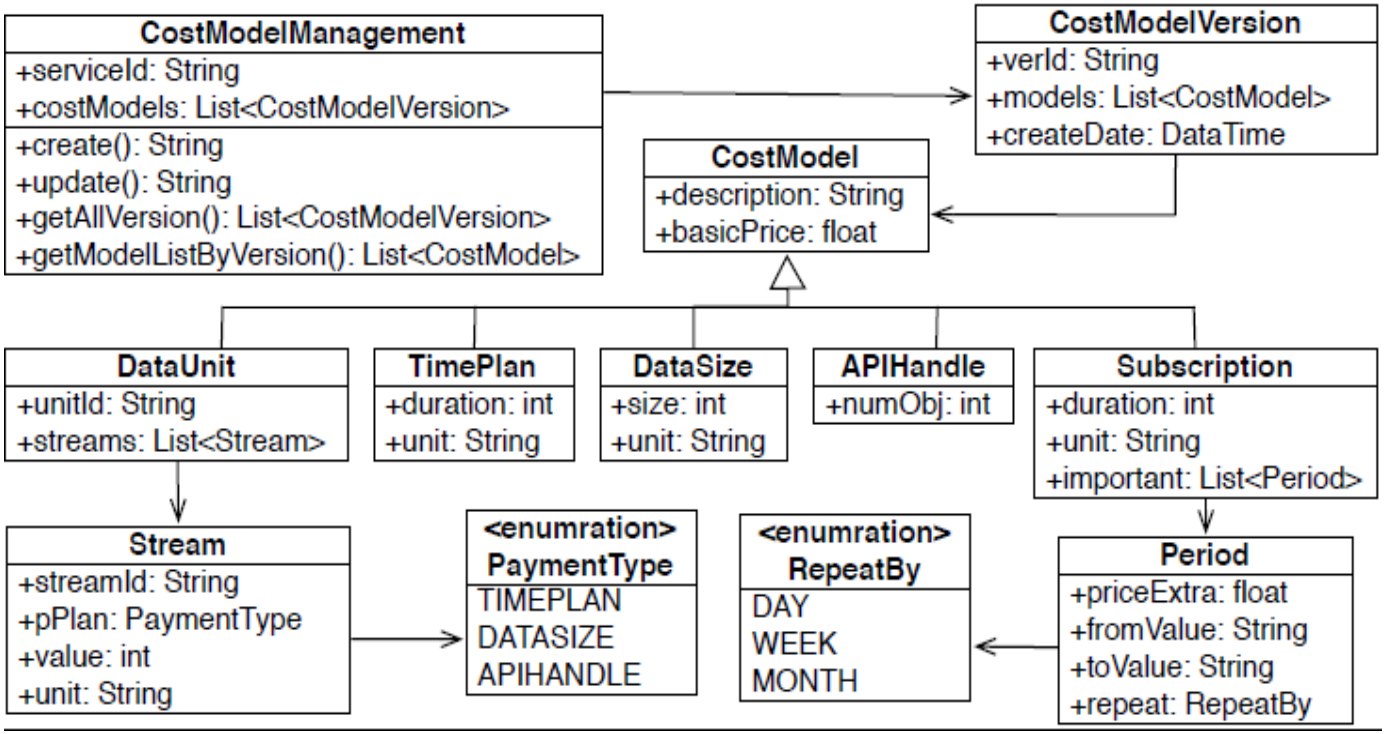
MARSA Design overview



MARSA description for human-sensing data marketplace

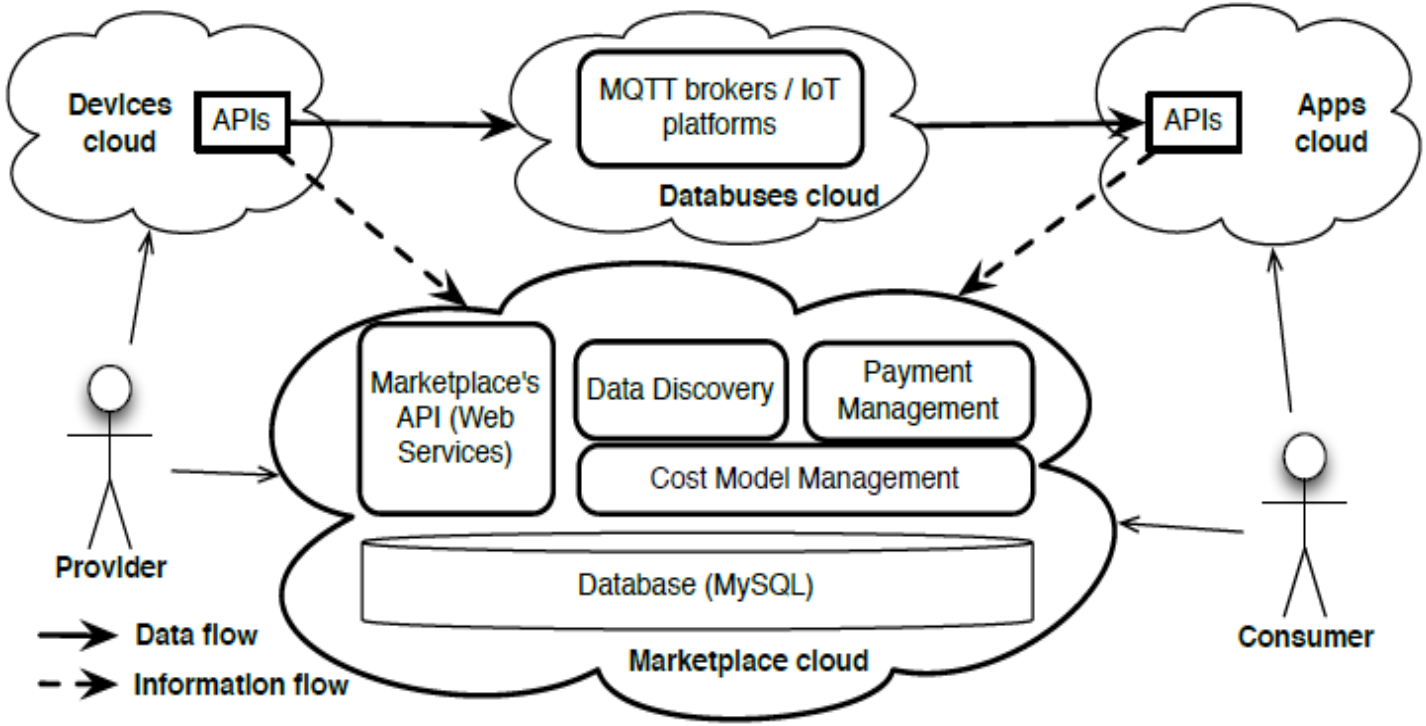


Cost model

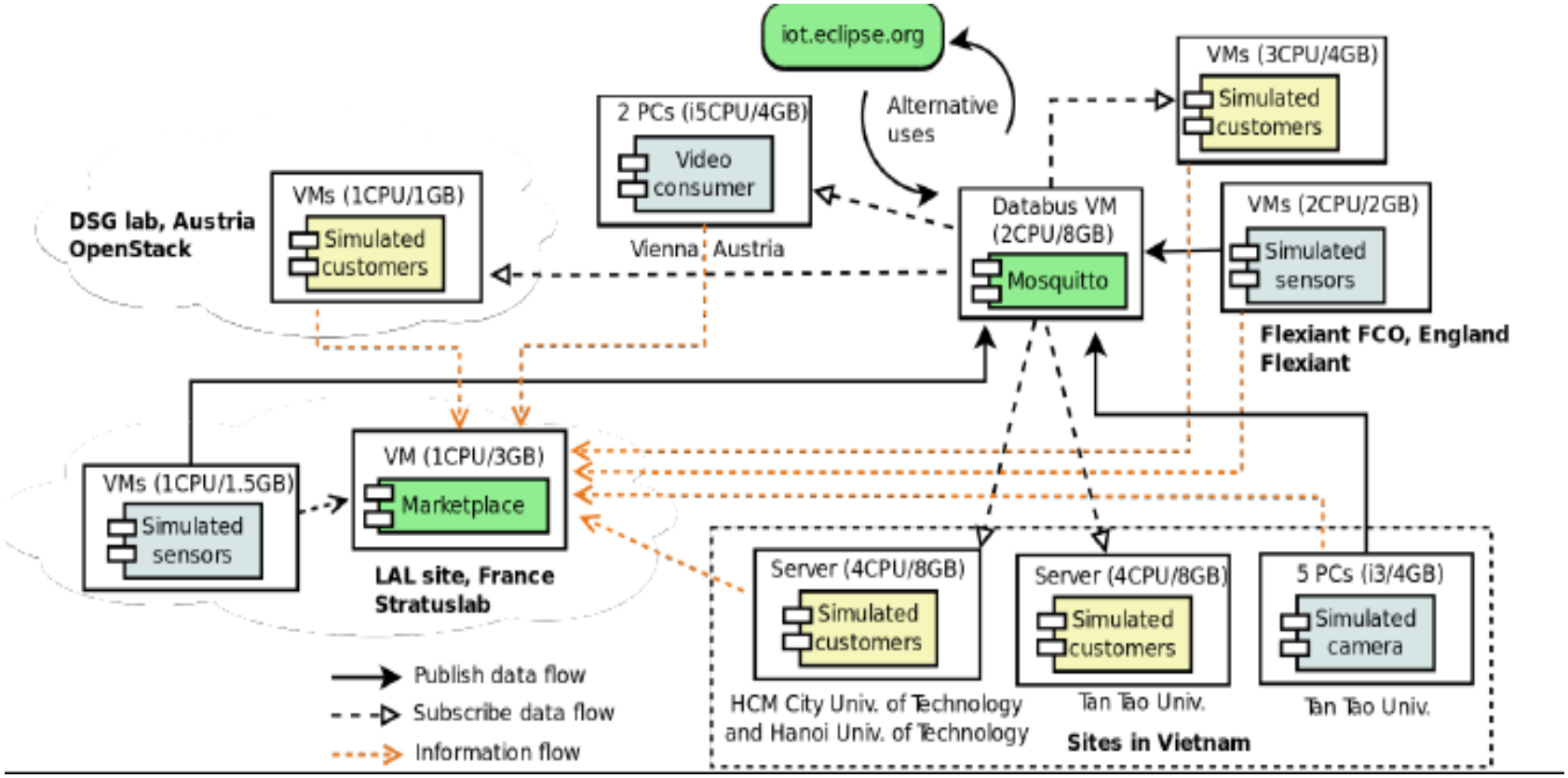


Quality of data has not supported yet

Implementation



Testbed


iot.eclipse.org
Alternative uses
VMs (3CPU/4GB)
Simulated customers
2 PCs (i5CPU/4GB)
Video consumer
Vienna, Austria
DSG lab, Austria
OpenStack
VMs (1CPU/1GB)
Simulated customers
Databus VM (2CPU/8GB)
Mosquitto
VMs (2CPU/2GB)
Simulated sensors
Flexiant FCO, England
Flexiant
VMs (1CPU/1.5GB)
Simulated sensors
VM (1CPU/3GB)
Marketplace
LAL site, France
Stratuslab
Server (4CPU/8GB)
Simulated customers
HCM City Univ. of Technology and Hanoi Univ. of Technology
Server (4CPU/8GB)
Simulated customers
Tan Tao Univ.
5 PCs (i3/4GB)
Simulated camera
Tan Tao Univ.
Sites in Vietnam
Publish data flow
Subscribe data flow
Information flow

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
			Total price: $ 3.41

Payment on SUBSCRIPTION (2.0 $ / 1 HOUR)

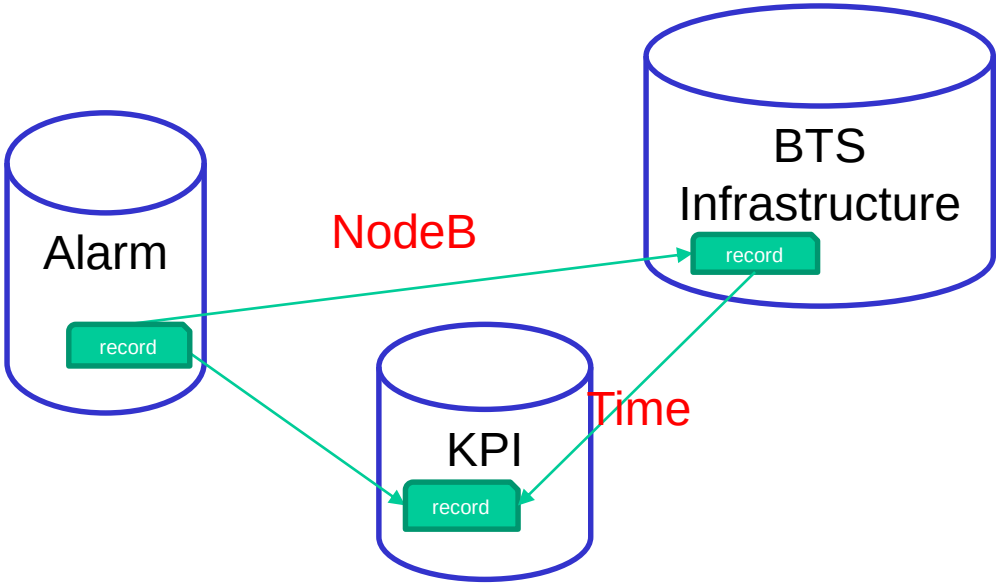
List of streams

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
						Total Price: $ 19.38

Total price of contract: $ 22.79

DATALAKE

Example: Linking data in telco management



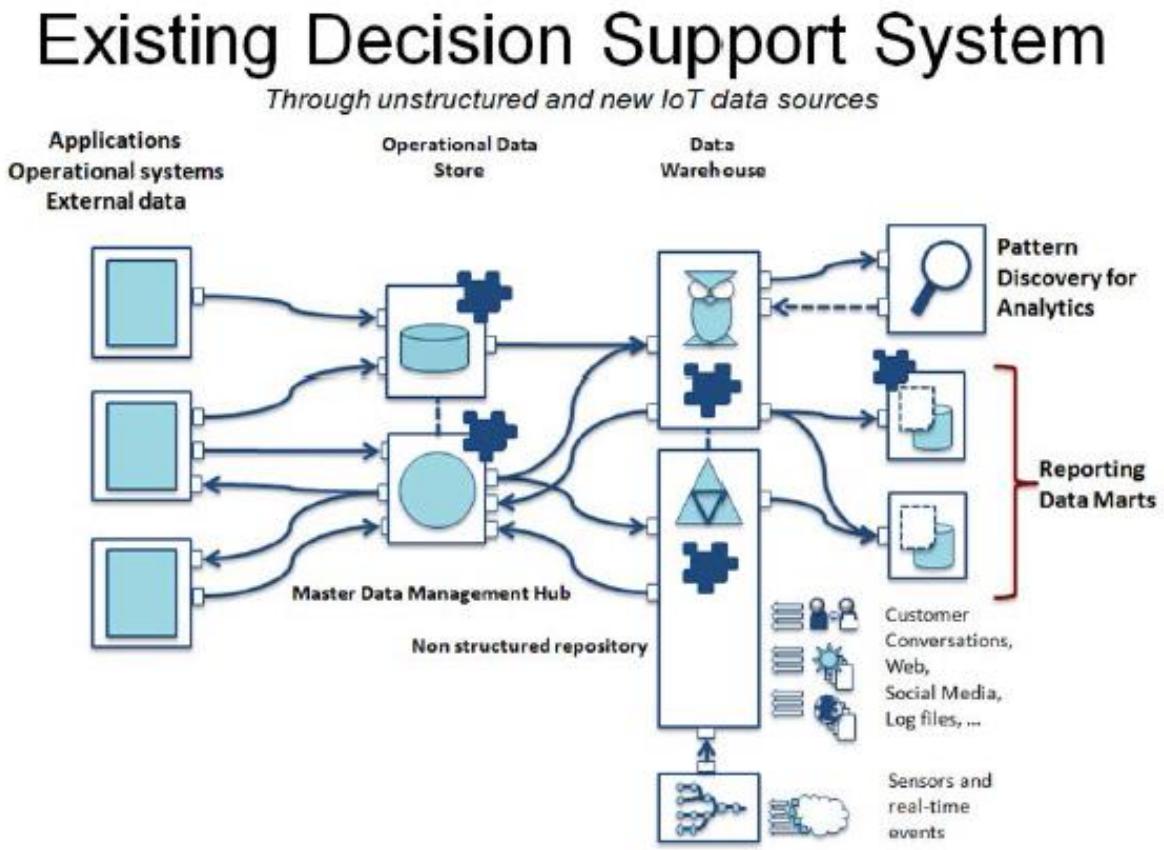
You can continue to have different data sources like that but you need to make sure they are linked

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 take is a concept so you can implement it based on your requirements and needs

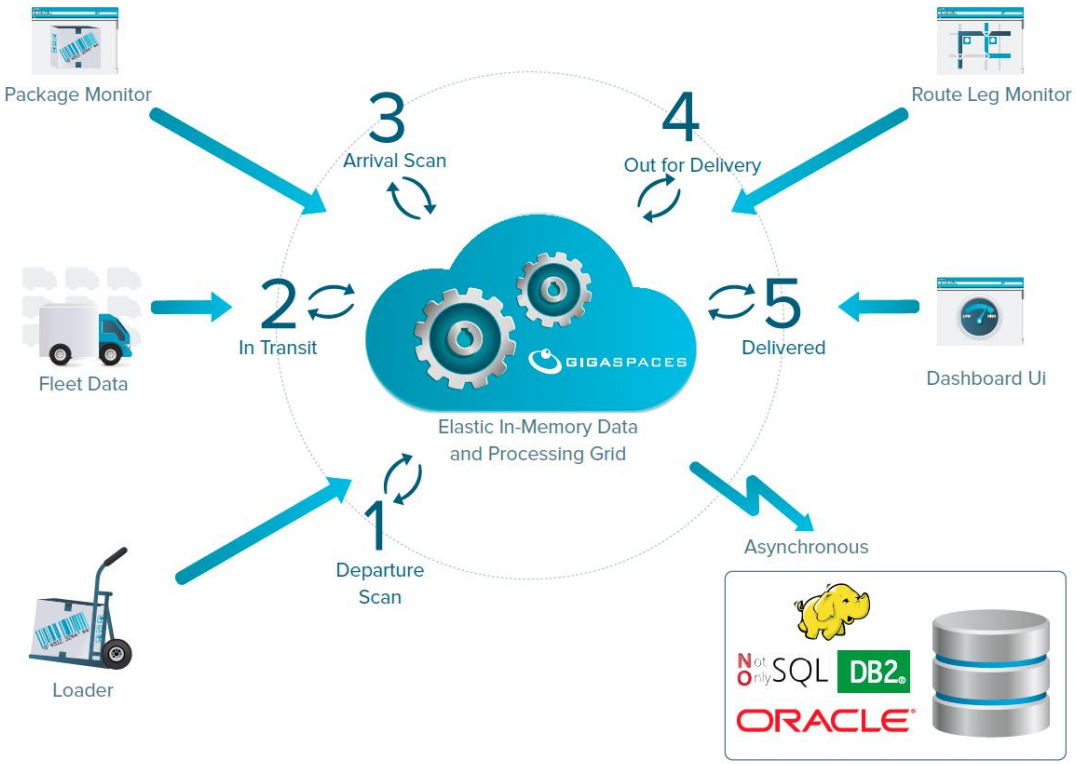
Example



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

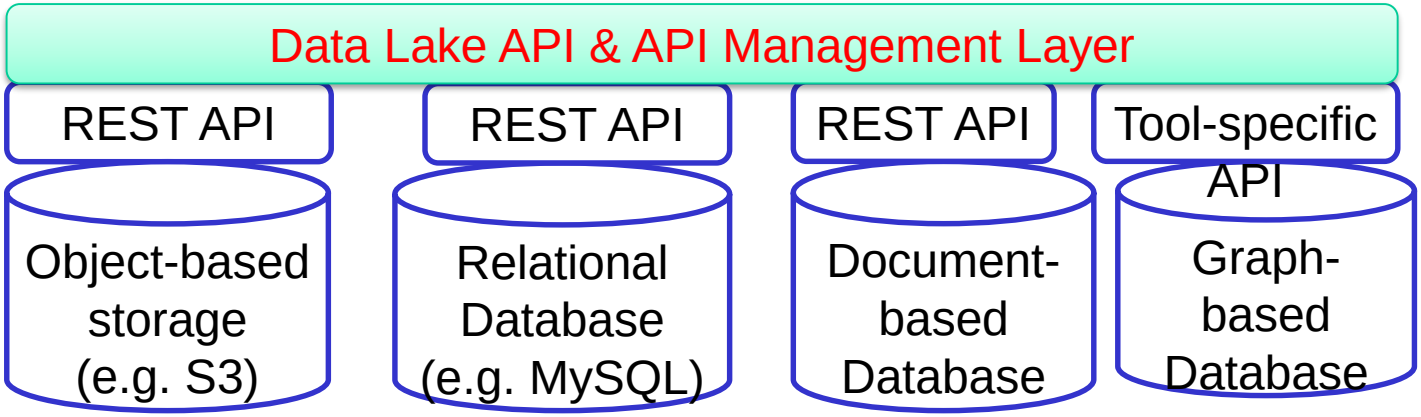
Implementation

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

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)

- 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)

- 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
- Build your own mini data marketplace
- Build your own datalake

Thanks for your attention

Hong-Linh Truong
Distributed Systems Group, TU Wien
truong@dsg.tuwien.ac.at
http://dsg.tuwien.ac.at/staff/truong
@linhsolar