Data as a Service – Models and Engineering

Hong-Linh Truong
Distributed Systems Group,
Vienna University of Technology

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

- Data provisioning and data service units
- Data-as-a-Service concepts
- Data concerns for DaaS
- Evaluating data concerns
What is the common point here?

„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-exhaustive list! Add your own issues!
Stakeholders in data provisioning

Data Provider
- People (individual/crowds/organization)
- Software, Things

Service Provider
- Software and people

Data Consumer
- People, Software, Things

Data Aggregator/Integrator
- Software
- People + software

Data Assessment
- Software and people

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

Consumption, ownership, provisioning, price, etc.

Data

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

Unit Concept

Service model

Data service unit

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

What about the granularity of the unit?
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
- Sensor-as-a-Service
- Database-as-a-Service (Structured/non-structured querying systems)
- Storage-as-a-Service (Basic storage functions)

deploy

Edge and/or Cloud Systems

ASE Summer 2016
Examples of DaaS

- **Xively Cloud Services™**
  - [https://xively.com/](https://xively.com/)

- **Directory Services**
  - Searchable directory of objects and permissions

- **Data Services**
  - Time-Series Archiving

- **Business Services**
  - Device provisioning, activation and management

- **Message Bus**
  - Real-time message management and routing

**Examples**

- **Bustling Manufacturers & Business Services List**
  - Published by DNB
  - Xively

- **Crime Statistics for England & Wales**
  - Published by Custom Web Apps, Ltd.

**Gnip**

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**

Opening up Government

- Search Datasets
  - 8729 Datasets

- **Tags**
  - nationalindicators, health, healthdata, transparency, children, health, health-and-social-care, education, children

- **UK Location**
  - 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 a 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 wrt data/service concerns
- SOAP versus REST
- Streaming data API
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**
  - Service APIs
  - Data APIs for the whole resource

- **Data Resource**
  - Data APIs for particular resources
  - Data APIs for data items

- **Data Items**
  - Data APIs for data items
DaaS design & implementation – structures (2)

DaaS

Consumer

Consumer

SOAP/REST/Streaming APIs

Data assets

Data resource

Data items

Data items

Data resource

Data items

Data resource

Data resource

Data resource

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

Examples: using WSO2 data service

 TU W IEN !

ASE Summer 2016
DaaS design & implementation – patterns for „turning data to DaaS“ (2)

Examples: using Amazon S3
DaaS design & implementation – patterns for „turning data to DaaS“

- DaaS design & implementation – patterns for „turning data to DaaS“

Examples:
- using Crowd-sourcing with Pachube (the predecessor of Xively)

One Thing → 10000... Things
DaaS design & implementation – patterns for „turning data to DaaS“ (4)

Examples: using Twitter

Storage/Database /Middleware

REST API v1.1 Resources

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

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET statuses/mentions_timeline</td>
<td>Returns the 20 most recent mentions (tweets containing a user'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...</td>
</tr>
<tr>
<td>GET statuses/user_timeline</td>
<td>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 &quot;owns&quot; the timeline or is an approved follower of the owner. The timeline...</td>
</tr>
<tr>
<td>GET statuses/home_timeline</td>
<td>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...</td>
</tr>
<tr>
<td>GET statuses/retweets_of_me</td>
<td>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.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET statuses/retweets/id</td>
<td>Returns a collection of the 100 most recent retweets of the tweet specified by the id parameter.</td>
</tr>
</tbody>
</table>
DaaS design & implementation – not just „functional“ aspects (1)

Data Assessment /Improvement

Profiling → Cleansing → Enrichment → Integration → ...

Data Assessment /Improvement

data → .... → .... → DaaS → data assets

APIs, Querying, Data Management, etc.

Data concerns

Quality of data → Ownership → Price → License → ....

ASE Summer 2016
DaaS design & implementation – not just „functional“ aspects (2)

Understand the DaaS ecosystem

Specifying, Evaluating and Provisioning Data concerns and Data Contract
Example -
http://www.strikeiron.com/
DATA CONCERNS
What are data concerns?

APIs, Querying, Data Management, etc.

Quality of data? free?
Privacy problem?
Located in US?
Service quality?
price?
redistribution?

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?
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 accuracy and completeness 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 particular, 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

- Data Service Contract
  - Data-specific terms
  - Service-specific terms
  - Common terms
Implementation (1)

Check http://www.infosys.tuwien.ac.at/prototyp/SOD1/dataconcerns
Implementation (2)

- Data privacy concerns are annotated with WSDL and MicroWSMO
Implementation (3)

Michael Mrissa, Salah-Eddine Tbahriri, Hong Linh Truong: Privacy Model and Annotation for DaaS. ECOWS 2010: 3-10
Populating DaaS concerns

The role of stakeholders in the most trivial view

- Data Provider
- Service Provider
- Data Aggregator/Integrator
- Data Consumer
- Data Assessment

DaaS Concerns:
- evaluate, specify, publish and manage
- specify, select, monitor, evaluate
- monitor and evaluate
Data concerns in multi-dimensional elasticity

Simple dependency flows

(increase nr. of services)

(increase)

(increase cost)

(increase response time)

How do we maintain our systems to deal with such complex dependencies?
HOW TO EVALUATE DATA CONCERNS FOR DATA ASSETS IN DAAS?
Patterns for „turning data to DaaS“

1. Data
2. Build Data Service APIs
3. Deploy Data Service
4. DaaS
5. Data
6. Data
7. Storage/Database-as-a-Service
8. DaaS
9. Data
10. Storage/Database/Middleware
11. DaaS
12. Data
13. Storage/Database/Middleware
14. DaaS
15. Things
16. People
17. Data
18. Data
19. DaaS

ASE Summer 2016
Data-related activities

Typical activities for data wrapping and publishing

Wrapping data → Publishing DaaS interface → Provisioning data

Typical activities for data updating & retrieval

Updating data → data ← Selecting data
Typical data concern evaluation

What do we need in order to perform these activities?

- Evaluating data concerns
- Describing data concerns
- Populating data concerns

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

<table>
<thead>
<tr>
<th>Evaluation Scope</th>
<th>At which level the evaluation is performed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Modes</td>
<td>When the evaluation is done?</td>
</tr>
<tr>
<td>Integration Model</td>
<td>How the evaluation tool is invoked?</td>
</tr>
</tbody>
</table>

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

Diagram:
- Data Consumers
- DaaS Service Operation
- Data storage
- Data and data concerns
- Data Concerns Evaluation Tool

Flow:
- Data Consumers → DaaS Service Operation
- DaaS Service Operation → Data Concerns Evaluation Tool
- Data Query Parameters
- Data Resources
- Get Data
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, Strikelron,
- 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


QoD framework: publishing concerns (1)

- Off-line data concern publishing, e.g.
  - a common data concern publication specification
  - a tool for providing data concerns according to the specification
  - supported by external service information systems
QoD framework: publishing concerns (2)

- On-the-fly querying data concerns associated with data resources, e.g.,
  - Using REST parameter convention
  - Based on metric names in the data concern specification
QoD framework: publishing concerns (3)

- Specifying requests by using utilizing query parameters the form of metricName=value

```text
GET/resource?crq.accuracy="0.5"&crq.location="Europe"
```

- Obtaining context and quality by using context and quality parameters without specifying value conditions

```text
curl http://localhost:8080/UNDataService/data/query/Population annual growth rate (percent)?crq.qod
{"crq.qod" : {
 "crq.dataelementcompleteness ": 0.8654708520179372,
 "crq.datasetcompleteness": 0.7356502242152466,
 ... 
}}
```
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
- Write small programs to test public DaaS, such as Xively, Microsoft Azure and Infochimps
- 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
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