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
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”

„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”

„Measure and report water quality metrics”

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

Data collection, assessment and enrichment

Data concerns

Data management and provisioning
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

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

- Consumption, ownership, provisioning, price, etc.
- "basic component"/"basic function" modeling and description

Data model
Unit concept
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

Edge/Cloud Infrastructure
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/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
Examples of DaaS

Bustling Manufacturers & Business Services List
published by DDB

Crime Statistics for England & Wales
published by Custom Web Apps, Ltd.

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

Gnip

Twitter Feeds

Get Started!

DATA.GOV.UK

Opening up Government

Search Datasets

Search Tags

Tags

Search

8729

Distributed Systems Group
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
  - 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 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)

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

Data Assessment /Improvement

Profiling → Cleansing → Enrichment → Integration → ...

 APIs, Querying, Data Management, etc.

Data concerns

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

DaaS

data

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

Understand the DaaS ecosystem

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

“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.

### Decision domains for data governance

**Figure 1: Key organizational assets to be governed; adapted from Weill and Ross.**

<table>
<thead>
<tr>
<th>Key assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human assets</td>
</tr>
<tr>
<td>Financial assets</td>
</tr>
<tr>
<td>Physical assets</td>
</tr>
<tr>
<td>IP assets</td>
</tr>
<tr>
<td>Information and IT assets</td>
</tr>
<tr>
<td>Relationship assets</td>
</tr>
</tbody>
</table>

**Figure 2: Decision domains for data governance.**

- **Data principles**
  - Metadata
  - Data access
  - Data lifecycle

Table 1: Framework for data decision domains.

<table>
<thead>
<tr>
<th>Data Governance Domains</th>
<th>Domain Decisions</th>
<th>Potential Roles or Locus of Accountability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Principles</td>
<td>• 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?</td>
<td>• Data owner/trustee • Data custodian • Data steward • Data producer/supplier • Data consumer • Enterprise Data Committee/Council</td>
</tr>
<tr>
<td>Data Quality</td>
<td>• 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?</td>
<td>• Data owner • Subject matter expert • Data quality manager • Data quality analyst</td>
</tr>
<tr>
<td>Metadata</td>
<td>• 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?</td>
<td>• Enterprise data architect • Enterprise data modeler • Data modeling engineer • Data architect • Enterprise Architecture Development Committee</td>
</tr>
<tr>
<td>Data Access</td>
<td>• 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?</td>
<td>• Data owner • Data beneficiary • Chief information security officer • Data security officer • Technical security analyst • Enterprise Architecture Development Committee</td>
</tr>
<tr>
<td>Data Lifecycle</td>
<td>• 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?</td>
<td>• Enterprise data architect • Information chain manager</td>
</tr>
</tbody>
</table>
What are data concerns?

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

APIs, Querying, Data Management, etc.

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

Storage/Database-as-a-Service
### Quality of data

<table>
<thead>
<tr>
<th>Read-only DaaS</th>
<th>CRUD DaaS</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Important factor for the selection of DaaS.</td>
<td>• Expected some support to control the quality of the data in case the data is offered to other consumers</td>
</tr>
<tr>
<td>• For example, the <strong>accurary</strong> and <strong>completeness</strong> of the data, whether the data is <strong>up-to-date</strong></td>
<td></td>
</tr>
</tbody>
</table>
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
Check http://www.infosys.tuwien.ac.at/prototyp/SOD1/dataconcerns
Populating DaaS concerns

The role of stakeholders in the most trivial view

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

- evaluate, specify, publish and manage
- specify, select, monitor, evaluate
- monitor and evaluate
HOW TO EVALUATE DATA CONCERNS FOR DATA ASSETS IN DAAS?
Patterns for „turning data to DaaS“

1. **Data**
   - **Build Data Service APIs**
   - **Deploy Data Service**
   - **DaaS**

2. **Things**
   - **Storage/Database as-a-Service**
   - **DaaS**

3. **People**
   - **Storage/Database/Middleware**
   - **DaaS**

---

ASE Summer 2017
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
- Selecting data
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

Diagram:
- Data Consumers
- DaaS Service Operation
- Data Storage
- Data Concerns Evaluation Tool

Connections:
- Data Consumers to DaaS Service Operation: data resources
- DaaS Service Operation to Data Storage: get data, data query parameters
- Data Storage to Data Concerns Evaluation Tool: data and data concerns
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


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

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

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

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

ASE Summer 2017
Data marketplaces and related components/services

Diagram showing interactions between data and services, including service discovery, integration, consumer and provider interactions, and related components like DaaS, DEMODS, and data storage.
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.


ASE Summer 2017
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**: \((\text{termName}, \text{termValue})\)
  - **Term name**: common terms or user-specific terms
  - **Term value**: a single value, a set, or a range

<table>
<thead>
<tr>
<th>Category</th>
<th>Term representation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rights</td>
<td>(\text{termName} = {\text{val}_1, \text{val}_2, \cdots, \text{val}_n})</td>
<td>(\text{termName} = {\text{Derivation, Collection, Reproduction, Attribution, Noncommercial use}},) (\text{val}_i = {\text{Undefined, Null, Allowed, Required, True, False}})</td>
</tr>
<tr>
<td>Quality of data</td>
<td>(\text{val}_l \leq \text{termName} \leq \text{val}_u)</td>
<td>(\text{termName} = {\text{Accuracy, Completeness, Uptodateness}}, \text{val}_l) and (\text{val}_u) (\in [0, 1])</td>
</tr>
<tr>
<td>Compliance</td>
<td>(\text{termName} = {\text{PrivacyCompliance}})</td>
<td>(\text{termName}) and (\text{val}_i) are any string, (\text{e.g., } \text{termName} = {\text{Sarbanes-Oxley (SOX) Act}})</td>
</tr>
<tr>
<td>Pricing model</td>
<td>(\text{termName} = {\text{cost} = \text{val}_1,))</td>
<td>(\text{termName}) is any string, (\text{e.g., MonthlyPayment; val}_1 \in \mathbb{R}, \text{e.g., cost} = 50) €, (\text{usagetime} = \text{val}_2,) (\text{maximumuse} = \text{val}_3))</td>
</tr>
<tr>
<td>Control and relationship</td>
<td>(\text{termName} = \text{val})</td>
<td>(\text{termName}) and (\text{val}) are any string, (\text{e.g., }\text{Liability, LawandJurisdiction}}) and (\text{val} = {\text{US, Austria}})</td>
</tr>
</tbody>
</table>
Structuring abstract data contracts

Concrete data contracts can be in RDF, XML or JSON

Use Identifiers and Tags for identifying and searches
Discussion time

HOW DOES NEAR-REALTIME DATA IMPACT ON DATA CONTRACT EXCHANGE?
Data Market without Marketplace?

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


But what about data contract? \( \rightarrow \) 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

Cannot buy expensive traffic data collection systems!

Figure sources: Internet
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

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

<table>
<thead>
<tr>
<th>No.</th>
<th>Stream UUID</th>
<th>Size</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>suuid1427702254973/sid1</td>
<td>0.219 GB</td>
<td>$ 1.1</td>
</tr>
<tr>
<td>2</td>
<td>suuid1427702254973/sid2</td>
<td>0.0217 GB</td>
<td>$ 0.11</td>
</tr>
<tr>
<td>3</td>
<td>suuid1427702254973/sid3</td>
<td>0.0550 GB</td>
<td>$ 0.28</td>
</tr>
<tr>
<td>4</td>
<td>suuid1427702254973/sid4</td>
<td>0.181 GB</td>
<td>$ 0.9</td>
</tr>
<tr>
<td>5</td>
<td>suuid1427702254973/sid5</td>
<td>0.205 GB</td>
<td>$ 1.02</td>
</tr>
</tbody>
</table>

Total price: $ 3.41

**Payment on SUBSCRIPTION (2.0 $ / 1 HOUR)**

**List of streams**

<table>
<thead>
<tr>
<th>No.</th>
<th>Stream UUID</th>
<th>Size</th>
<th>Price</th>
<th>Size Extra</th>
<th>Price Extra</th>
<th>Sum Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>suuid1427702254973/sid11</td>
<td>3.67 HOUR</td>
<td>$ 7.34</td>
<td>0</td>
<td>$ 0</td>
<td>$ 7.34</td>
</tr>
<tr>
<td>2</td>
<td>suuid1427702254973/sid12</td>
<td>6.02 HOUR</td>
<td>$ 12.04</td>
<td>0</td>
<td>$ 0</td>
<td>$ 12.04</td>
</tr>
</tbody>
</table>

Total Price: $ 19.38

**Total price of contract:** $ 22.79

ASE Summer 2017
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
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