Data Mesh Architecture / Designing Data Products

Designing Data Products

According to J. Majchrzak et al., a "*data product is an autonomous, read-optimized, standardized data unit containing at least one dataset (Domain Dataset), created for satisfying user needs*". Conceptually, a mesh is a graph, a network, consisting of nodes and connecting edges. Each node in a data mesh is called **data product**. Every data product exists within a bounded context and one bounded context might contain several data products.

As with every architecture, data architecture also has a sociotechnical perspective. Therefore, we need to consider people and processes in addition to technology for data management. Furthermore, because we have consumers for our data, these consumers expect a certain quality. Therefore, our data need to be accessible, and everyone should know where to find data. Thus, the sociotechnical perspective pushes us to adopt the product thinking philosophy for data - data products.

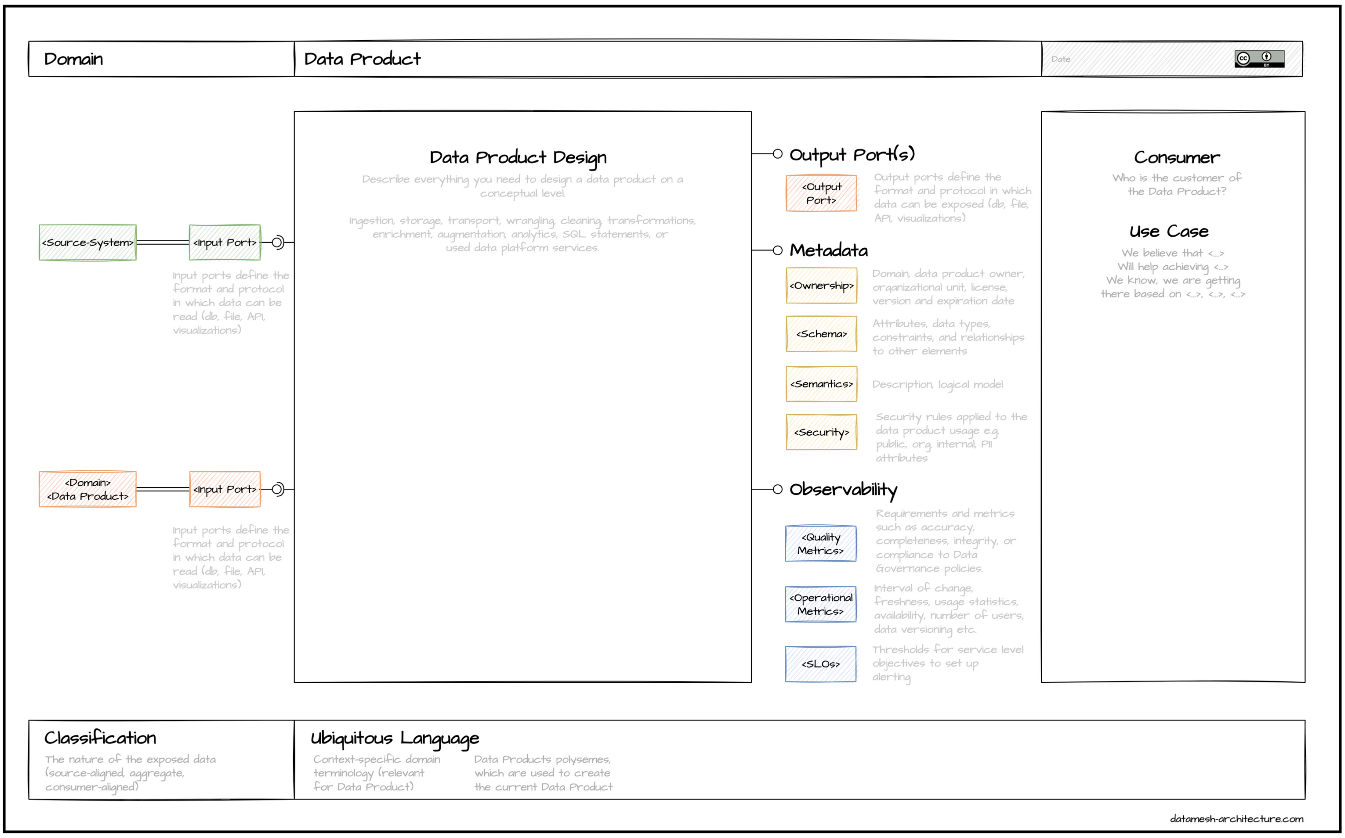
What could be a data product? Generally, any data representation that has value for its consumers can be a good candidate. In the following, you will find a list of examples of possible data products:

- A database table or view
- Raw unstructured files: e.g., images or videos uploaded by users of a video portal; however, to be valuable to consumers, they should be provided with metadata
- Data stream of data entities from a transaction system
- Data stream representing the history of changes to the application: For example, events that relate to changes made within a billing account
- Simple files: Data in CSV format, excel files
- Partitioned files in optimized (Parquet) format
- REST API: Read-optimized data exposed from applications
- Master Data Management database
- Features for the machine learning models
- Visualizations and dashboard

Building and maintaining data products is a financial investment and takes up domain team's capacity. Therefore, the value and costs need to be evaluated upfront.

Data Product Canvas

To establish a structured process of data product design across an organization, we propose to start with a **Data Product Canvas**.



Describe everything you need to design a data product on a conceptual level.

Ingestion, storage, transport, wrangling, cleaning, transformations, enrichment, augmentation, analytics, SQL statements, or used data platform services.

Input ports define the format and protocol in which data can be read (db, file, API, visualizations)

Input ports define the format and protocol in which data can be read (db, file, API, visualizations)

Output ports define the format and protocol in which data can be exposed (db, file, API, visualizations)

Domain, data product owner, organizational unit, license, version and expiration date

Attributes, data types, constraints, and relationships to other elements

Description, logical model

Security rules applied to the data product usage e.g. public, org. internal, PII attributes

Requirements and metrics such as accuracy, completeness, integrity, or compliance to Data Governance policies.

Interval of change, freshness, usage statistics, availability, number of users, data versioning etc.

Thresholds for service level objectives to set up alerting

Who is the customer of the Data Product?

We believe that <....>
Will help achieving <...>
We know, we are getting
there based on <...>, <...>, <...>

The nature of the exposed data (source-aligned, aggregate, consumer-aligned)

Context-specific domain terminology (relevant for Data Product)

Data Products polysemes, which are used to create the current Data Product

A **Data Product Canvas** is a visual framework that guides your team through the data product specification. We suggest to fill out this canvas collaboratively. In total, the Data Product Canvas consists of ten building blocks:

1. Domain
2. Data Product Name
3. Consumer and Use Case
4. Output Port
5. Metadata
6. Input Ports
7. Data Product Design
8. Observability
9. Ubiquitous Language
10. Classification

In the following, we describe each of the building blocks that constitute the data product. The order of the building blocks represents the recommended sequence of the collaborative steps during the workshop.

Domain

Each data product should be implemented, evolved, and maintained by one domain team only. Therefore each data product belongs exactly to one domain. The following possible questions are relevant in this building block:

- Who is accountable for the data product?
- Who specifies its requirements?
- Who will answer questions about the data product?
- Who fixes it when it breaks?

Data Product Name

Each data product has a unique name to be identified and accessed within an organization. Typically, the product name should follow the common naming strategy.

Consumer and Use Case(s)

This building block describes the reason behind the existence of the data product. Data product design follows the "Product Thinking" philosophy. We always start with the consumer needs.

To identify the purpose of the data product, we describe analytical use cases and organizational objectives. Understanding the use cases is essential to specify the data required to implement the use cases. The consumer of the future data product might be either our own domain team or different domain teams.

Output ports

The output ports define the format and consumption protocol in which data can be exposed. For example, the output port can be a database table, file, API, or visualizations.

Metadata

In addition to the output port specification, the data product should define its metadata.

- The ownership part describes domain name, product owner, organizational unit, license, version and probable expiration date of the data product
- The data schema part describes attributes, data types, constraints, and relationships to other elements in the data unit (e.g functional dependencies)
- The semantics part provides description about the logical model of the data unit.
- The security block describes security rules applied to the data product usage e.g. public, organization, internal, and PII attributes

The discussion about required dataset attributes and their meaning contributes to the team's understanding of the data itself. Please use additional space to brainstorm and discuss on some example.

Input Ports

This building block describes the input data for the future data product. The input ports are receiving mechanism for data that will constitute the data product. The input ports define the format and protocol in which data can be read. We distinguish here between *operational* source systems and other *data products*, which might be either internal or coming from other domains.

Data Product Design

This is the core building block where we design the internals of the data product. This is the place to specify everything between the input and output ports. Please, describe everything you need to design a data product on a conceptual level. For instance, you might describe data ingestion, storage, transport, wrangling, cleaning, transformations, enrichment, augmentation, analytics, SQL statements, or data platform services.

Observability

In addition to metadata specification, the data product design implies information for observability.

- Quality Metrics outlines data quality requirements and metrics such as accuracy, completeness, integrity, as well as compliance to data governance policies.
- Operational Metrics might include interval of change, freshness, usage statistics, availability, number of users, data versioning etc.
- SLOs for data product enable the discipline of building trustworthiness in each data product. We specify here the thresholds for metrics to trigger alarms.

Please note, this is not an extensive list of metadata and observability points. Make sure that you define all the information required in your organization in order to establish the data mesh architecture.

Ubiquitous Language

Describe here a common language that is shared between everyone involved in the project. This is usually a context-specific domain terminology that is relevant for operational systems and data product.

Classification

In this block, we specify the nature of the exposed data, meaning, we classify our data product as either, source-aligned, aggregate, or consumer-aligned.

Data Product Canvas Templates

The proposed canvas is suitable for working collaboratively on data products design. The canvas might be easily used either plotted on paper or with the common online whiteboards like Miro or Mural. Once your organization has created a collection of data products using the proposed data product canvas, you can start connecting all data products and produce an actual mesh.

- Download Template as PDF (DIN A0)

Further Reading

- Product Thinking 101
- Data Mesh in practice: Product thinking and development
- How to identify Data Products? (by Agile Lab)

SUPPORTED BY
INNOQ