



Published in Towards Data Science



Xavier Gumara Rigol

Follow

Jul 8, 2021 · 5 min read · Listen



Save



Data as a product vs data products. What are the differences?

Understand with examples the similarities and differences between a “data product” and “data as a product”



Green rope meshwork by [Clint Adair](#) on [Unsplash](#)

Since the publication of the [data mesh introductory article](#) by Zhamak Dehghani, there has been a lot of discussion around the definition of what is a “data product” in and outside of the data mesh context.

By clarifying a couple of definitions in this article, we hope the concepts of “data product” and “data as a product” become clearer to anyone entering the data and data mesh worlds.

Data products

Let's start with the generic definition of “data product”

Open in app ↗

Sign up

Sign In



Jujitsu: The Art of Turning Data into Product, 2012).

This means that any digital product or feature can be considered a “data product” if it uses data to facilitate a goal. For example, the home page of a digital newspaper can be a data product if the news items featured in the home page I see are dynamically selected based on my previous navigation data.

In 2018, [Simon O'Regan](#) published an article called [Designing Data Products](#) that lists very clear examples of data products and groups them by type: raw data, derived data, algorithms, decision support and automated decision-making.

Here is a list of example data products including the category they belong to and the interfaces used to access it:

- **A company dashboard to visualise the main KPIs of your business.** This data product is of the type decision support system and the interface to access it is a visualisation.
- **A data warehouse.** This data product is a mix of raw, derived data and decision support system. The interface to access it are probably SQL queries.
- **A list of recommended restaurants nearby.** Since the list is curated specifically for you, this data product is an automated decision-making one. The interface to access it is an app or website.
- **A “faster route now available” notification on Google Maps** is a decision support data product (as you are the one making the decision) and its interface

is a web/app.

- A **self-driving car** is a data product too. Since it drives automatically, it is of the type automated decision-making. Its interface is, well, the car itself.

Data as a product

One of the principles of the data mesh paradigm is to consider data as a product. Sometimes this principle has been abbreviated to “data products”, hence the confusion.

“Data product” is a generic concept (👏 173 | 💬 4 above) and “data as a product” is a subset of all possible data products. More specifically, if we use Simon’s categories, “data as a product” belongs to the raw or derived data type of “data product”.

If we dive in the data mesh world, this quote from Zhamak Dehghani’s original article is key to understand the definition of data as a product: *“Domain data teams must apply product thinking [...] to the datasets that they provide; considering their data assets as their products and the rest of the organization’s data scientists, ML and data engineers as their customers.”*

In summary, “data as a product” is the result of applying product thinking into datasets, making sure they have a series of capabilities including discoverability, security, explorability, understandability, trustworthiness, etc.

An example of data as a product

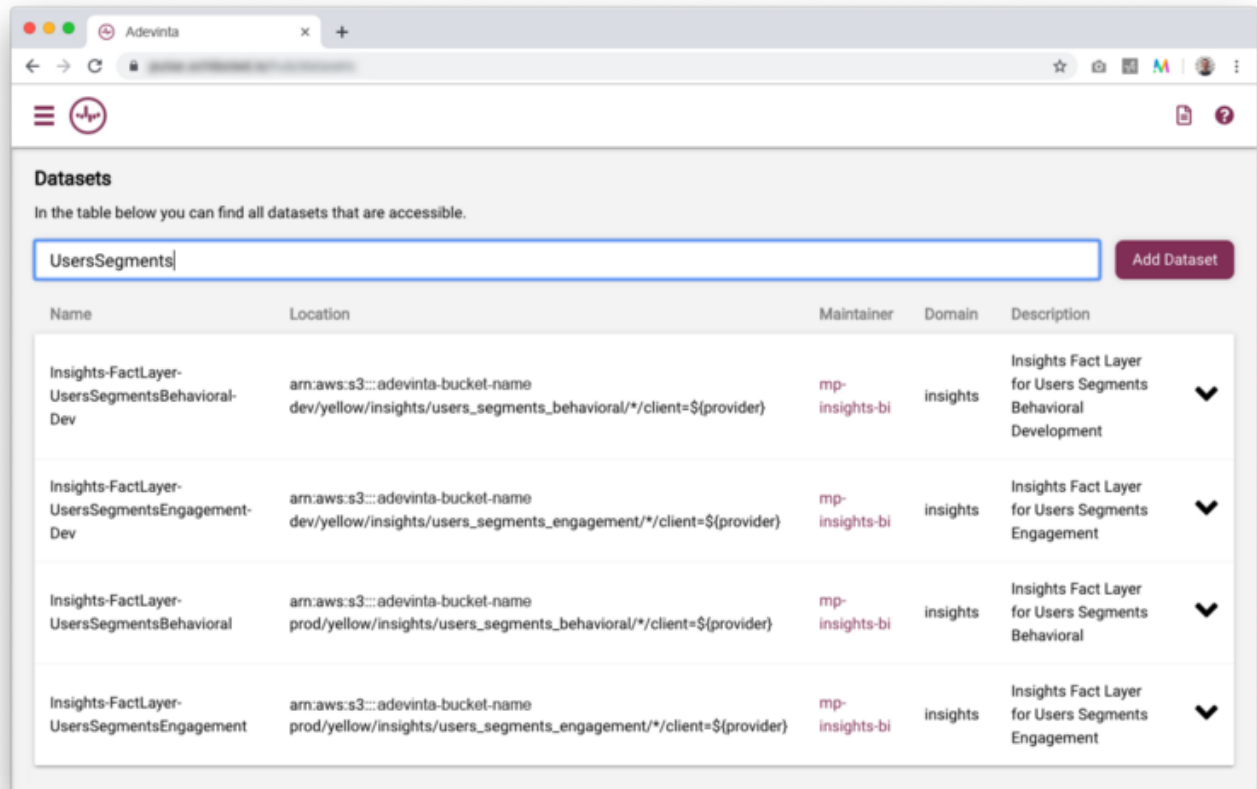
So, what does data as a product look like? A data as a product contains the code, its data and metadata, and the necessary infrastructure to run it. Furthermore, it needs to fulfil the capabilities described before.

In a talk we gave at the Data Council conference in Barcelona in 2019 titled “A Federated Information Infrastructure That Works” (video, transcript as blog post) we put an example of a dataset as a product as used in Adevinta by listing its qualities:

Discoverable

In order for data as a product to be discoverable, a search engine is needed and users must be able to register datasets in this engine and request access to them (this will increase security, another capability explained below).

The first iteration for this capability could be just a list of datasets in your de facto internal intranet and you can iterate and build incrementally from that. Remember that processes and culture are more important than deploying the ultimate data catalogue tool too early (which can be too complex for employees to use).



Datasets

In the table below you can find all datasets that are accessible.

Search: Add Dataset

Name	Location	Maintainer	Domain	Description
Insights-FactLayer-UsersSegmentsBehavioral-Dev	arn:aws:s3:::adevinta-bucket-name-dev/yellow/insights/users_segments_behavioral/*/client=\${provider}	mp-insights-bi	insights	Insights Fact Layer for Users Segments Behavioral Development
Insights-FactLayer-UsersSegmentsEngagement-Dev	arn:aws:s3:::adevinta-bucket-name-dev/yellow/insights/users_segments_engagement/*/client=\${provider}	mp-insights-bi	insights	Insights Fact Layer for Users Segments Engagement
Insights-FactLayer-UsersSegmentsBehavioral	arn:aws:s3:::adevinta-bucket-name-prod/yellow/insights/users_segments_behavioral/*/client=\${provider}	mp-insights-bi	insights	Insights Fact Layer for Users Segments Behavioral
Insights-FactLayer-UsersSegmentsEngagement	arn:aws:s3:::adevinta-bucket-name-prod/yellow/insights/users_segments_engagement/*/client=\${provider}	mp-insights-bi	insights	Insights Fact Layer for Users Segments Engagement

An example of Adevinta's custom build data catalogue that makes datasets discoverable

Addressable

Having addressable datasets makes your teams more productive. On one side, Data Analysts and Data Scientists are autonomous in finding and using the data they need. On the other side, Data Engineers have far less interruptions from people asking where they can find data about X.

Data Location

The master location of events is S3, accessed through [Athena](#) or [Jupyter](#). See the data model of the entity in the section below: [Events Behavioral schema](#).

1.- **Dataset** name to [request access](#): `Insights-FactLayer-EventsBehavioral`

2.- **Athena** (SQLaaS): `XXXXXX_databox.insights_events_behavioral_fact_layer_365d` where XXXXXX is your client_id.

3.- **S3 path** `s3://adevinta-bucket-name`

```
prod/yellow/insights/events/source=pulse/version=4/year=$year/month=$month/day=$day/
gen=0/client=$client/
```

Metadata for “data as a product” that makes the dataset addressable

Self-describing and interoperable

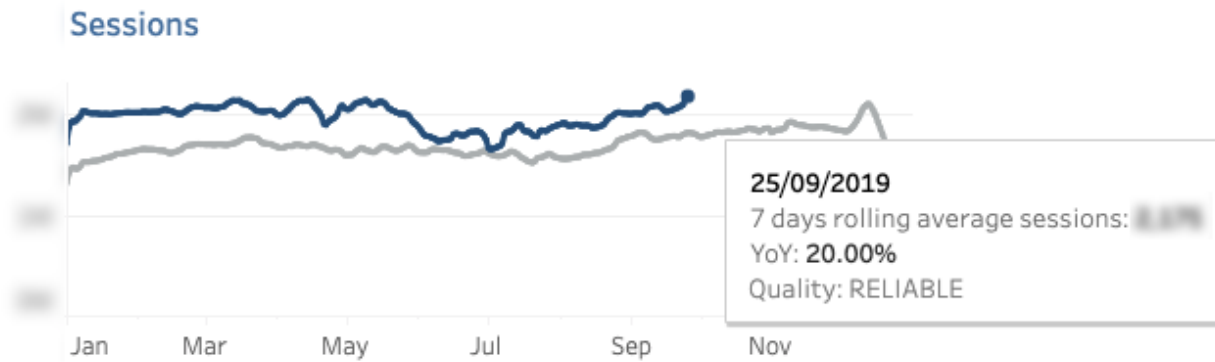
As we commented in the blog post where we explained [Adevinta’s data mesh journey](#), datasets need to contain metadata that make them understandable and follow the same naming conventions (which will make the datasets interoperable). We found these pieces of metadata to be super useful to our Data Analysts:

- Data location (as seen above)
- Data provenance and data mapping
- Sample data
- Execution time and freshness
- Input preconditions
- Example notebook or SQL queries using the data set

Trustworthy and secure

Checking data quality regularly and automatically is a must to fulfil the trustworthy characteristic of data as a product. And owners of the datasets need to react accordingly to the results of these checks.

Quality checks must be done at pipeline input and output and it doesn’t hurt to provide contextual data quality information to consumers of the data; like for example in Tableau dashboards.



Contextual data quality shown in a Tableau dashboard

Finally, registered data sets should not be automatically available to everyone. Employees need to request access to each one of them and data controllers need to grant or deny access individually.

When requesting access, it is mandatory to specify until when the access is needed and for what purpose.

Further reading

Understanding data as a product is fundamental to succeed in the implementation of a data mesh in your organisation. You can expand the knowledge around this topic by reading the following articles cited in this post:

- [Designing Data Products](#) by Simon O'Regan
- [How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh](#) by Zhamak Dehghani
- [Data Mesh Principles and Logical Architecture](#) by Zhamak Dehghani
- [Building a data mesh to support an ecosystem of data products at Adevinta](#) by Sandra Real and Xavier Gumara Rigol
- [A Federated Information Infrastructure that Works](#) by Xavier Gumara Rigol

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. [Take a look.](#)

By signing up, you will create a Medium account if you don't already have one. Review our [Privacy Policy](#) for more information about our privacy practices.



Get this newsletter

[About](#) [Help](#) [Terms](#) [Privacy](#)

Get the Medium app

