Published in Towards Data Science

Eric Broda Follow

Feb 11, 2022 · 8 min read · ✨ · Listen

Data Mesh Patterns: Enterprise Data Product Catalog

The Enterprise Data Product Catalog makes it easy to find, visualize, consume, and govern data in an Enterprise Data Mesh. Let's take a detailed look at how this pattern works and is used by an Enterprise Data Mesh

69

2

Photo by Denny Müller on Unsplash

Enterprise Data Product Catalog: A Foundational Data Mesh Pattern

This article discusses the Enterprise Data Product Catalog, the fourth in a series of articles on Foundational Data Mesh Patterns. I will be discussing the pattern summary, the problem domain and business context, how the pattern works, and candidate vendors that enable this pattern.

This article assumes that you have a high-level understanding of Data Mesh. If you need some background information on Data Mesh, there are a number of great articles available here (principles), here (architecture), and here (patterns).

A list of the full series of articles is provided at the end of this article.

Pattern Summary

An Enterprise Data Product Catalog is a "catalog-of-catalogs" that makes it easy to find, visualize, consume, and govern data (more specifically, meta-data) in any Data Product with the Enterprise Data Mesh. It is a one-stop-shop for finding, viewing, consuming, and governing data within the Enterprise Data Mesh.

Context and Business Problem

The modern enterprise is clogged with data. Data is stored in operational data sources, legacy platforms, data warehouses, data lakes, and now, so-called data lakehouses (and the list goes on). The challenge confronted by the enterprise, or more specifically those professionals whose responsibility is to safeguard, govern, and consume enterprise data, is that data is hard to find.

And once found, it is hard to understand. And once found and understood, it is difficult to determine if it is up-to-date or accurate: is it "book-of-record" quality, or is it several days old? Or has it been transformed, and if so, how, when, and by whom, or which application?

These are challenging problems to address today: Individual application-specific solutions are too narrowly focused and business group silo solutions produce limited silos of information. And even if they could be aggregated, how can so many unique solutions be bolted together to provide what users really need: a one-stop-shop for finding, viewing, governing, and consuming enterprise data.

A common yet problematic approach to addressing this problem is to use a "Master Data Management" (MDM) approach. Unfortunately, MDM is complicated, expensive, and rarely (if ever) works in large organizations.

Why? Mostly because of one simple observation: MDM (today) in most implementations tries to copy/centralize data into a new "master" even though an existing master already exists, typically, in an operational system. This need for centralization is contrary to Conway's Law, an almost fundamental truth in human behaviour, in that systems (and data) align to an organization's structure, or more specifically, its decision making and funding structure. In other words, incentives and human behaviour strongly favour keeping data in the various federated groups that own the data.

Rather, Data Mesh addresses this issue through "Data Products" and "Enterprise Data Product Catalogs" which allow data to stay where it already resides and making it available by APIs (this is one of the core Data Mesh principles). A Data Mesh approach requires no data movement, no data consolidation, fewer lengthy debates, and fewer organizational headaches. But most importantly Data Mesh provides a better understanding of where master data exists.

Solution

The Enterprise Data Product Catalog is the one-stop-shop for finding, viewing, governing, and consuming data within an Enterprise Data Mesh. There are several components that interact directly with or feed into an Enterprise Data Product Catalog:

- Data Product Catalogs
- Schema Registry
- Immutable Change/Audit Log

Solution Component: Data Product Catalog

A Data Product Catalog, first and foremost, is a repository of information ("meta-data") about the data contained in a Data Product in the Enterprise Data Mesh. But it also provides a way for users — which may include developers, data governance professionals, and data scientists — to find, analyze, and view information about data contained managed by the Data Product.

A Data Product Catalog contains data / meta-data from several sources including:

- Change Data Capture
- Schema Registry
- Immutable Change/Audit Log

A Data Product Catalog is consumed by several users:

- Data Scientists use it to find data in a Data Product and understand its structure and format to train AI/Machine Learning models.
- Business Users use it to find and consume data for decision making purposes.
- Developers use it to find and understand data structures and formats to build and integration applications.
- Governance Professionals use data lineage and consumption patterns to govern and audit data and support regulatory demands.

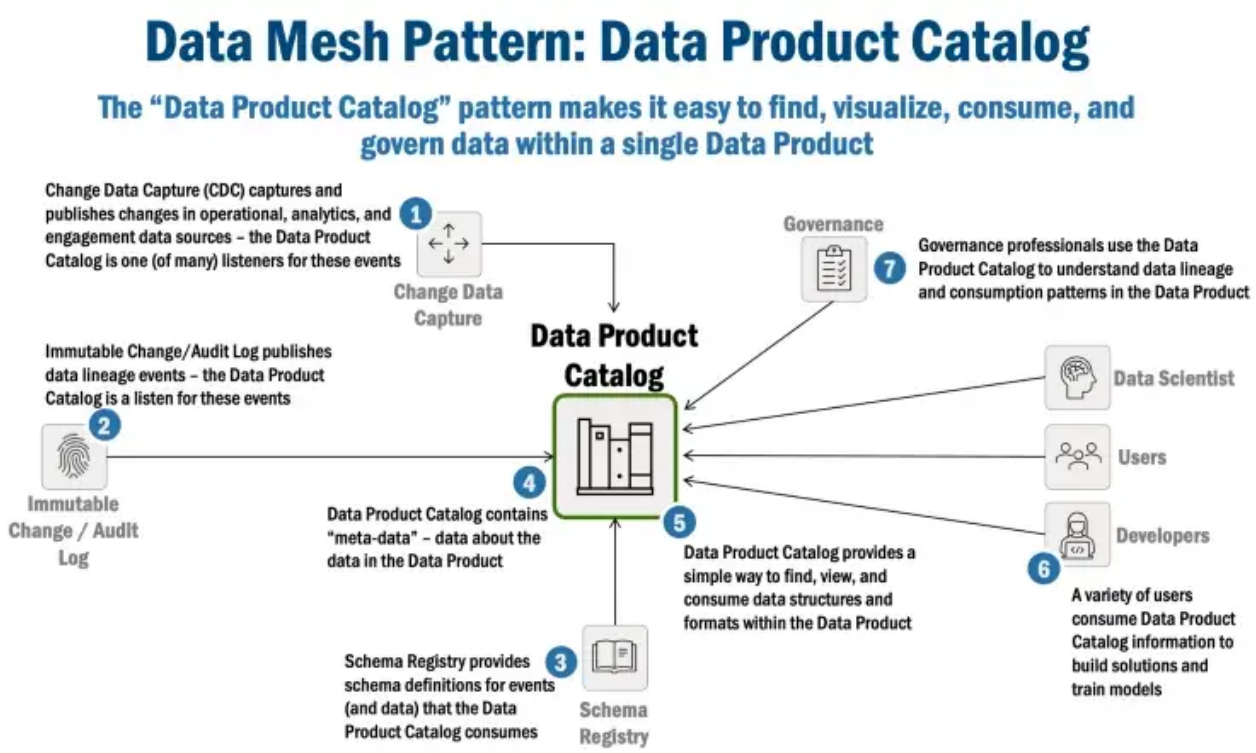

Figure 1, Data Mesh Pattern: Data Product Catalog Pattern

1. Change Data Capture (CDC) acquires and publishes events (using an Event Streaming Backbone) representing changes in operational, analytical and engagement data sources; The Data Product Catalog is a listener for these events.

2. Immutable Change/Audit Log acquires and publishes events representing data lineage changes; The Data Product Catalog is a listener for these events.

3. Schema Registry acquires and publishes events representing JSON (or AVRO) schemas that define structure, format, and content of data in the Data Product Catalog; The Data Product Catalog is a listener for these events.

4. The Data Product Catalog contains meta-data (data about data) about data managed by the Data Product.

5. The Data Product Catalog has a user interface that makes it easy to find, view, and consume data structures within the Data Product.

6. Many different types of users access the Data Product Catalog: Data Scientists use it to understand the structure of data used to train models; Business Users use it to understand data used to make crucial business decisions; Developers use information about data and event structures to build applications.

7. Governance professionals use the Data Product Catalog to understand data lineage and consumption patterns.

Solution Component: Schema Registry

A Data Product Catalog contains several types of meta-data. It contains references to schemas managed by a Schema Registry. These schemas define events and, if needed, data within the Data Product. These schemas are used to help users understand data structure, format, and component/element data types which makes it easy to understand and visualize data within the Data Product. Figure 2, below, shows how the Schema Registry works.

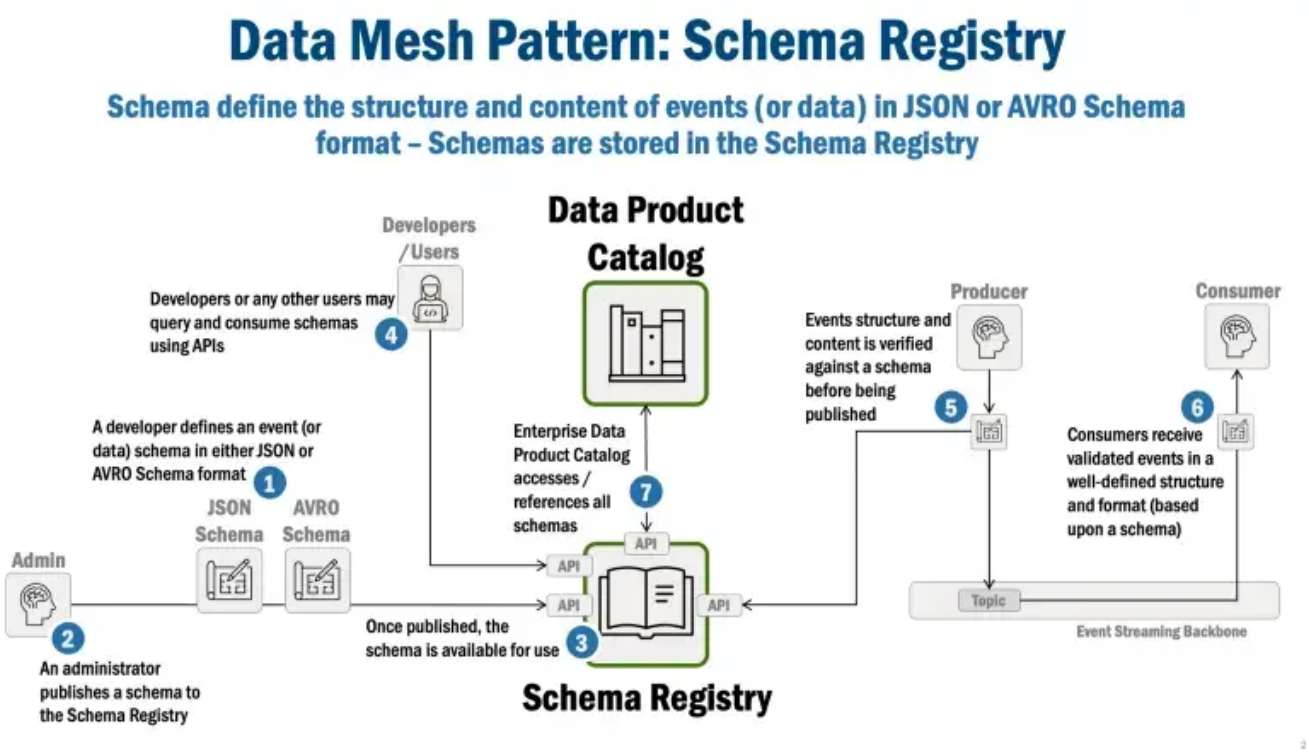

Figure 2, Data Mesh Pattern: Schema Registry

1. Developers define an event or data structure used by their applications in JSON (or AVRO) schema.

2. Administrators publish schemas to the Schema Registry

3. Once published the schema is available for use.

4. Developers, or any other user, may query and consume schemas from the Schema Registry using APIs.

5. Before publishing an event, a producer's event is verified against a schema in the Schema Registry; Only validated messages are published.

6. Consumers only receive validated and well-defined events.

7. Product Data Catalogs access event definitions in the Data Product Catalog.

Solution Component: Immutable Change/Audit Log

A Data Product Catalog contains references to data lineage events maintained within the Data Products Immutable Change/Audit Log. This allows users to find, understand, and how data changes as it is updated and transported across the Enterprise Data Mesh.

This has several uses: data scientists use the Data Product catalog to understand the lineage of training data to enable reproducibility, traceability, and verifiability of their models which is become a primary enterprise (and regulatory) concern. But it also makes the Data Product the primary "go-to" utility for data governance professionals. And for audit professionals, regulatory questions can be addressed much more easily when access to meta-data, and data history are easily searchable and visualized.

Enterprise Data Product Catalog: How it Works

The Enterprise Data Product Catalog is a "catalog of catalogs". It is a façade / proxy for accessing all local Data Product information. In most implementations, it does not physically maintain data product or schema registry data.

Rather, it has a simple registry of endpoints (APIs) for each local Data Product and Schema Registry that are used to "pull" relevant meta-data as needed. This approach avoids duplication of meta-data across the Enterprise Data Mesh while avoiding the inherent complexity in consolidating and keeping in sync a single data product catalog (ie. the problem that plagues "Master Data Management").

Figure 3 (below) illustrates how the Enterprise Data Product Catalog works.

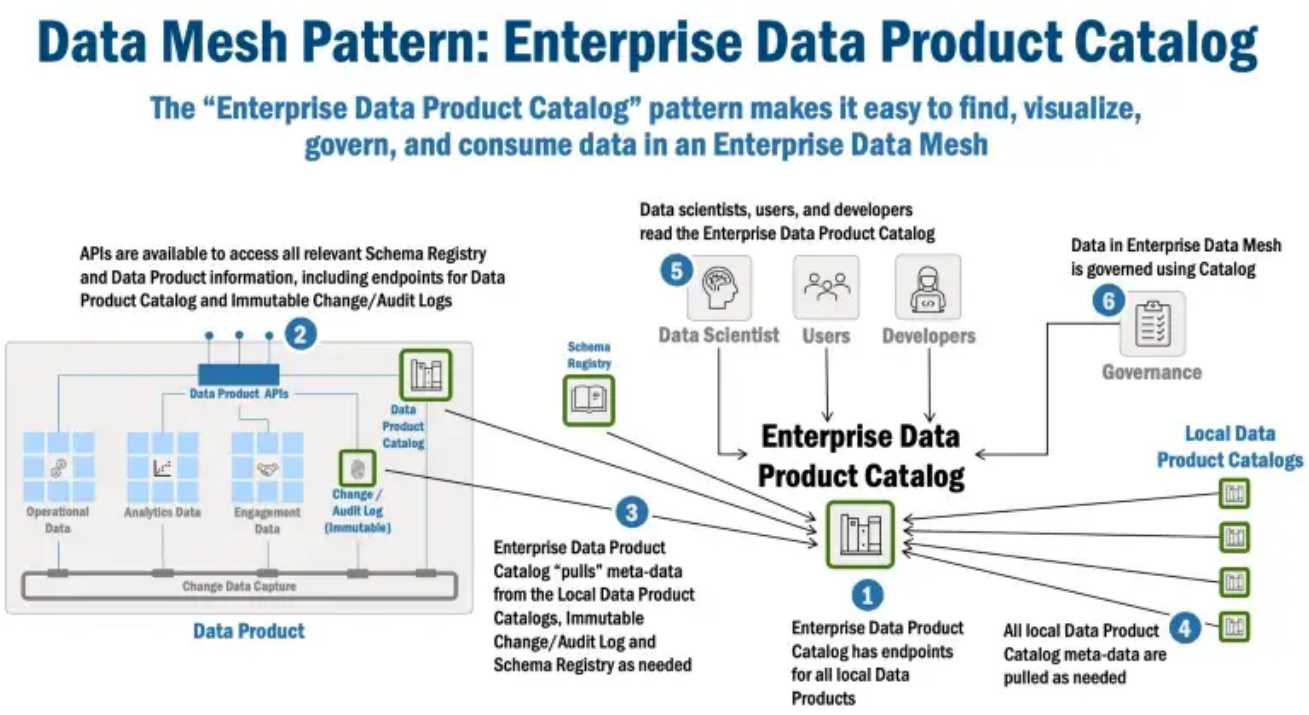

Figure 3, Data Mesh Pattern: Enterprise Data Product Catalog

1. Enterprise Data Product Catalog has registered API endpoints for all local data Products.
2. APIs are available in the Schema Registry as well as each Data Product providing easy access to all relevant data.
3. Enterprise Data Product Catalog calls a Data Product's APIs to "pull" data from Data Product Catalog and Immutable Change/Audit Log, and Schema Registry.
4. All local Data Products meta-data are pulled (accessed via APIs) as needed.
5. Data scientists use meta-data to improve AI/ML model training, business users use meta-data to improve decision making, and developers use data and event information to build applications.
6. Meta-data in the Enterprise Data Product Catalog is used to govern enterprise data and understand consumption patterns.

Vendor Landscape

The following is a list of products that implement the Event Streaming Backbone, each of which works will in an Enterprise Data Mesh.

- Schema Registry: Confluent's Schema Registry is really solid and it supports both JSON Schemas (what I mostly use) as well as AVRO Schemas; RedHat's

Service Registry appears to be plug-and-play compatible with Confluent's offering and it supports a wide array of schema types (although I have not used this product).

- Enterprise/Data Product Catalogs: I have had to build bespoke data catalogs for the most part, so I am not as familiar with specific vendor offerings although a few like Magda, Amundsen, and Atlan look promising.

*Full disclosure: I have **no financial interest** in recommending any of the above products — I am highlighting these products because I have some experience with them, and they have worked well for me.*

Other Articles in this Series

The full series of articles in this series on foundational Data Mesh patterns is list below.

- **Change Data Capture (CDC) pattern**, which tracks changes in a database and captures them as "events" (available here).
- **Event Streaming Backbone pattern**, used by CDC and other applications to publish and subscribe/receive events in an Enterprise Data Mesh (available here).
- **Immutable Change / Audit Log pattern**, which retains logs and tracks data lineage within the Enterprise Data Mesh for future audit and governance purposes (available here).
- **Enterprise Data Product Catalog pattern**, which is a catalog/repository that contains meta-data about Data Products in the Enterprise Data Mesh (this article).

Concluding Thoughts

The Enterprise Data Product Catalog is a foundational pattern that makes is simple to view, consume, and govern data managed by the Enterprise Data Mesh. The technical details — and the rational provided — should (hopefully) help you kickstart the design and implementation of an Event Streaming Backbone and bootstrap your Enterprise Data Mesh!

***

All images in this document except where otherwise noted have been created by Eric Broda (the author of this article). All icons used in the images are stock PowerPoint icons and are free from copyrights.

Data Mesh Data Science Artificial Intelligence Machine Learning Data Catalog