DataPorts

Solving End-to-End Value Chain Content Integration

The Consumer Goods FORUM

End-to-End Value Chain

intel

Index

Introduction

Product content and content quality are key to commerce, and as supply chains accelerate so the impact of errors escalate. When the wrong product is shipped to a consumer because of errors in an online catalogue it costs the consumer, retailer, wholesaler, and manufacturer in satisfaction, shipping, restocking, lost sales, and reputation. When content errors occur upstream of the consumer, the impact worsens as multiple parties see losses on entire shipments of products. Bad demand and supply planning drives loss at scale. Content errors have the power to break not just individual transactions but longer-term trust and relationships, anywhere in the value chain.

The industry is asking for more efficient, accurate ways to move product content through the value chain - and the DataPorts project has emerged to address this need.

The fundamental concept of DataPorts is “lean content”. Just as we strive for “lean” in our physical value chains, we want the same for digital value chains. While many have already been addressing digital transformation within the enterprise, we must also do the same for digital transformation between enterprises.

This paper is a summary of the work done to date with the Consumer Goods Forum to answer a single initial question:

> *Is there a durable way in which we can design and build an open-source technical framework for peer-to-peer content integration for value chain partners, in full support of current and future industry and open standards?*

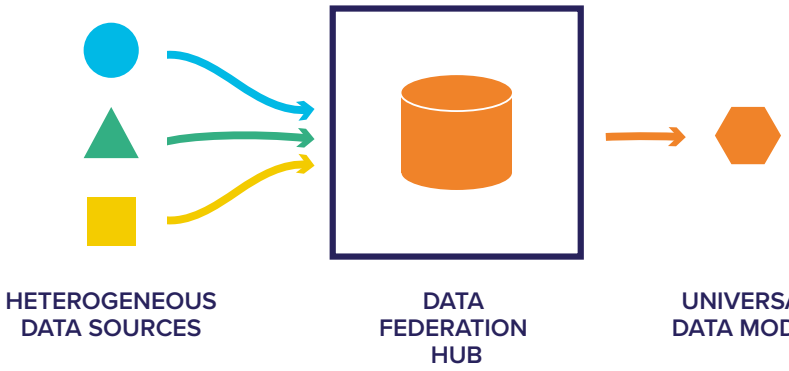
1. Concept

Today, when a retailer needs to assemble a complete view of information about a product for an online product catalogue, that retailer will typically go to multiple web-based sources, search for information, transform the information obtained into a format fit for purpose, and assemble the information elements for publication to the catalogue.

Today that content integration process is difficult. It is difficult because the steps of the process and the technology required for search and retrieval of information vary sufficiently from source to source, and even from item to item such that it is difficult to fully automate. The content integration problem blocks operational efficiency, introduces product information errors, and slows down the speed to market - resulting in increased costs and lost sales.

One of the approaches traditionally proposed for solving this content integration problem is “data federation”. Data federation works by standardising on a common federated data model and mapping all data sources to that standardised model. Content might be mapped either real-time in response to requests, or it might be stored using the common data model in an intermediate data store ready for consumption.

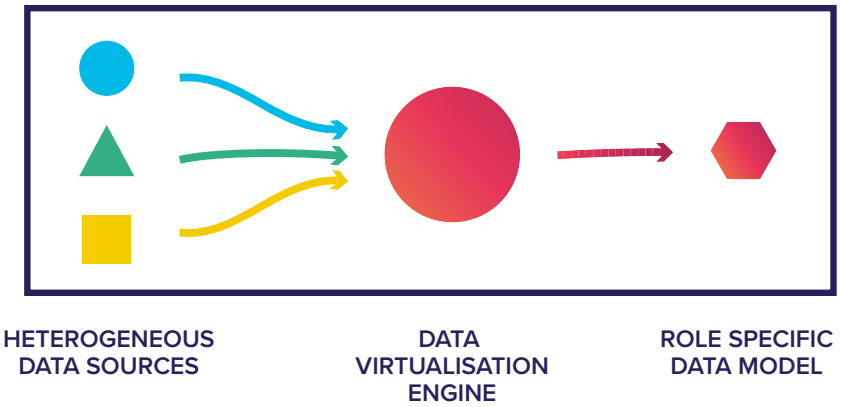
The challenge though is that agreeing on a standard model is difficult and despite big efforts and successes in standardisation across the value chain there always seem to be exceptions and in most cases there remains data which does not fit the model. We’ve been asked to find a more general way to share information such that partners can choose when and where to apply standards, adapting dynamically to changing needs.



Data Federation

Refers to the idea of creating an all encompassing data model which covers the data transfer needs for all participants and for all purposes. This can work for some situations, but in general it is a challenge, either because the resulting model is unreasonably large and complex, or because it's not possible to come to a stable agreement on a model.

With DataPorts we aim to simplify and automate content integration tasks between value chain partners by virtualising value chain content such that it can be more simply discovered and utilised by value chain participants. Rather than hoping for a single common federated data model to be shared by all supply chain participants, DataPorts rely on role-specific data models connected peer-to-peer allowing for the consistent sharing of data between pairs or groups of supply chain partners. We do this through an integration approach known as Data Virtualisation.



Data Virtualisation

A lean approach to a data integration in which data sources are decoupled or "virtualised" so that we can treat them alike, regardless of their location and implementation. This dramatically simplifies content integration tasks - especially across peer-to-peer networks of data sources.

2. Approach

DataPorts solve the content integration challenge in the end-to-end value chain by circumventing the difficulty faced in seeking to create a federated data model. Instead, DataPorts work on the principle that we can use **role-specific data models** across **peer-to-peer** communication channels to convey information between individual participants, on-demand, and in a format ready for use.

DataPorts are specialised web servers which make peer-to-peer content integration transparent. No matter where data sources actually reside DataPorts make data sources accessible together, through a single, common programming model; they neutralise the differences in location, programming interfaces and formats, to allow content integration to be automated efficiently.

We see three basic patterns of DataPort deployment required to meet different needs: Outbound, Inbound, and Peer-to-Peer. In general though it is the Peer-to-Peer pattern which is most interesting to this community - providing the secure mechanism for upstream and downstream content integration across the supply chain.

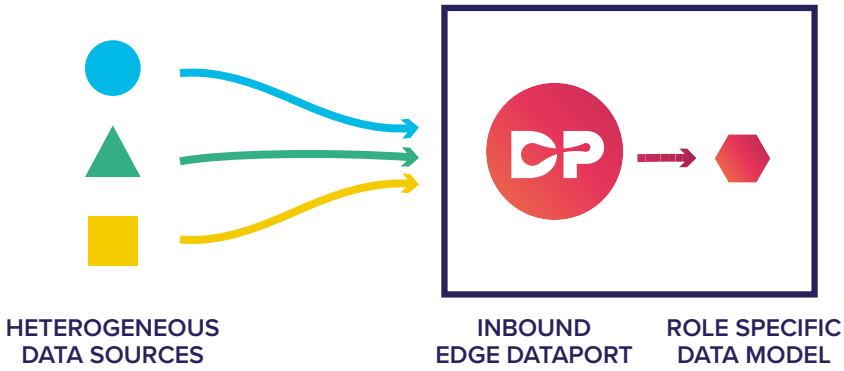
Outbound DataPorts serve internal data to other value chain participants through their published, role-specific data models matched to value chain participant needs.

HETEROGENEOUS DATA SOURCES | OUTBOUND EDGE DATAPORT | ROLE SPECIFIC DATA MODEL

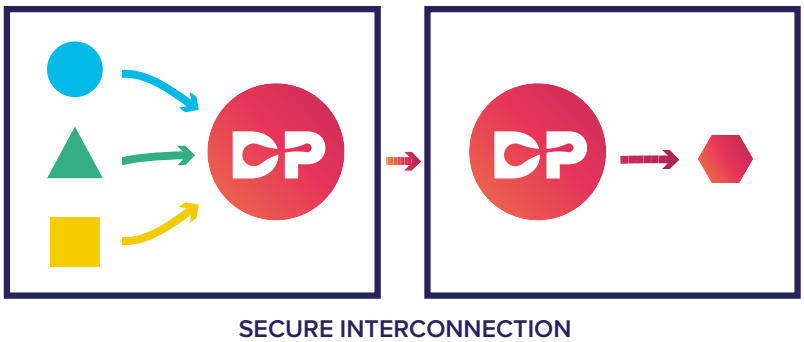
DataPorts are **data virtualisation** servers which share role-specific data schemas through which applications may run composite queries against decoupled heterogeneous data sources. They work by virtualising the participant data sources, optimising queries, transforming query results inline, and creating an aggregate query response.

Role-specific Data Model
Refers to the idea of creating a limited data model which fits the needs of communicating just the information needed to support a specific task or tasks.

Inbound DataPorts serve external data from value chain partners for consumption internally, through role-specific data models matched to internal needs.

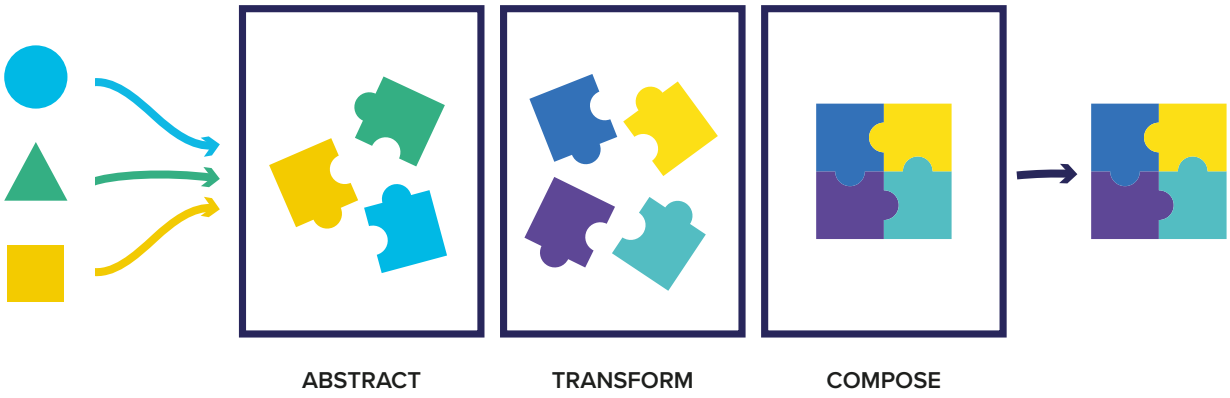


DataPorts connect **peer-to-peer**, via secure network connections, to allow complex interactions, provide data on-demand as well as data subscriptions for real-time data updates.



3. The Anatomy of a DataPort

DataPorts are comprised of three processing layers: Abstract, Transform, Compose - which together enable the rich capabilities which we need to meet the needs of content integration across the value chain: a single query interface, optimized across diverse and distributed data sources, delivering content aligned to requirements in terms of scope, units of measurement, file formats, consistency, and standardisation.



Abstract

Access any data source efficiently through a common interface, independent of the source implementation details. It is the abstraction layer which virtualises sources ensuring interoperability, inside and outside of the enterprise, from the simplest data sources all the way to full featured partner data hubs.

Transform

Perform operations on data retrieved from the different sources, including operations such as unit conversion, image format conversion, and deriving or inferring new information through advanced operations such as machine learning based labeling or the extraction of new insights from the data.

Compose

Combine the query result components from each source into a single response respecting the relationship between components according to the role-specific data model.

4. Implementation

DataPorts based on data virtualisation can be implemented using a variety of technologies, but for practicality we need to decide on specific technologies to make implementation and interoperability efficient. One of the most important areas of concern around interoperability is how to express the semantics of queries and data between DataPorts. The two major contenders we considered were: i) GraphQL with GraphQL schemas; ii) SPARQL with OWL/RDF.

We have chosen GraphQL for simplicity and completeness in terms of expressing queries and in handling abstraction, transformation, and composition. While OWL/RDF might win for being part of W3C standards and for its more complete semantic description of data, it is more complex than GraphQL and it does not offer a built in solution for abstraction, transformation and composition.

By using GraphQL as the engine for data virtualisation we have been able to quickly build working DataPorts using well supported, off-the-shelf open source libraries. Complex DataPort queries written intuitively in GraphQL and spanning multiple backend data sources are performance optimised with results returned fit for purpose according to the chosen data model.

GraphQL or Graph Query Language is an open source project contributed initially by Facebook® and the focus of attention by a growing community of contributors and value added platform developers. It solves the challenge of data integration through data virtualisation using a simple schema definition language, a simple query language and an architecture which optimises execution of queries across simple “resolvers” or dedicated functions which access individual component data sources.

5. Getting Started

The best way to get started with work on DataPorts is to start programming with GraphQL today. There are sufficient open-source libraries to build out all of the concepts expressed in this paper, with a number of extensions and commercial offerings which can help to accelerate development. At this time we are assembling a small group of interested parties to work together on publishing minimal reference DataPort implementation guidelines which anyone can work with. In the meantime, building up familiarity with the core concepts of data virtualisation and schema based integration will provide a good foundation.

Start learning about GraphQL
https://graphql.org

As a group we are working actively towards a common way of registering DataPorts so that businesses can find them, and subsequently query them, using both standard web browsers and using machine-to-machine requests for information. For those interested in the detail, we will publish a prototype DataPort registry at a URL which will be shared online alongside this paper on www.cgf.com.

6. Scale

With DataPorts we are concerned about ALL scales of participants, from the smallest, lowest “tech” network partners to the largest corporations in the value chain. We need to address the needs of all, providing a minimal effort and cost on-ramping solution for the smallest of suppliers as well as full scale implementations of DataPorts for participants wishing to connect with tens of thousands of suppliers. The approach we have taken has the capability to address all scales.

Our first DataPort prototype was built to automatically deploy simple tabular data to a minimal DataPort running on a cloud service for fractions of cents per transaction - equating to a small farmer with a few products uploading product information via a smartphone or basic PC, at the lowest possible cost. At the same time, the core technology and architecture we have chosen was built for data virtualisation across massive global data sources at a scale way beyond that which we anticipate for our use cases.

7. Implications

One of the great benefits of the approach outlined so far is automation of the content aggregation process downstream of content sources. This includes automation of simple tasks, such as the conversion of units of measurement, or image format and resolution conversion - but also, more complex computational tasks such as mapping between different standards. We may need to convert between packaging sizes, or decide packaging requirements for a specific quantity. All of these types of operations can be performed in-line in the “transformation” layer of our DataPorts design - but more than that, transformations can generate new attributes which can be used in searches and content delivery.

Simple conversions and mappings are just the beginning. By adding machine learning (ML) to the “transformation” layer of our DataPorts design we can further enrich content selection and thereby the content aggregation process. We might use ML to perform content labeling, such as the identification of image attributes (mood, gender, season, colors, packaging, etc) - so that we can automate the selection of images to meet specific criteria, and thereby ensure that when we need images of a specific product including “women” with “children” and “wool hats” in “winter”, then we can get precisely that. With recent advances in ML we might use the transformation layer to automatically generate an image which fits our query requirements, by synthesizing an image based on the raw content of multiple images or even fully synthesizing a photorealistic image from the given attributes.

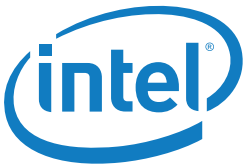
Beyond solving the original content integration problem our approach to DataPorts, with a focus on interoperability through data virtualisation, lends itself to a broad range of additional use cases based on supporting bi-directional, peer-to-peer communication between value chain partners and systems, at all scales. DataPorts can serve as a mechanism to accelerate modernization of data communication for the entire value chain.

8. Conclusions

We started this work with the simple ambition of finding a better way to address the needs of content integration across the value chain, expressed through one simple question: “Is there a durable way in which we can design and build an open-source technical framework for peer-to-peer content integration for value chain partners, in full support of current and future industry and open standards?”

We believe that the answer is a clear yes - supported by the concept of DataPorts as simple connection points in a network of peer-to-peer relationships between value chain participants. Through experimentation and proofs-of-technology we have established that we can build DataPorts using foundational, open-source technologies - with the option of building on and accelerating implementation using commercial extensions to those technologies.

While we chose what we consider to be the simplest implementation path for the purpose of experimentation, there exist both open-source and commercial alternatives - and while there are advantages to choosing a single common implementation technology, we have designed with interoperability in mind, so that by the very nature of DataPorts, different protocols and mechanisms can work together efficiently to provide the foundation for a “lean content” solution for the industry.



About The Consumer Goods Forum

The Consumer Goods Forum ("CGF") is a global, parity-based industry network that is driven by its members to encourage the global adoption of practices and standards that serves the consumer goods industry worldwide. It brings together the CEOs and senior management of some 400 retailers, manufacturers, service providers, and other stakeholders across 70 countries, and it reflects the diversity of the industry in geography, size, product category and format. Its member companies have combined sales of EUR 3.5 trillion and directly employ nearly 10 million people, with a further 90 million related jobs estimated along the value chain. It is governed by its Board of Directors, which comprises more than 50 manufacturer and retailer CEOs.

For more information, please visit:
www.theconsumergoodsforum.com

Contact Ruediger Hagedorn

Director, End-to-End Value Chain

The Consumer Goods Forum

r.hagedorn@theconsumergoodsforum.com

About Intel

Intel (NASDAQ: INTC), a leader in the semiconductor industry, is shaping the data-centric future with computing and communications technology that is the foundation of the world's innovations. The company's engineering expertise is helping address the world's greatest challenges as well as helping secure, power and connect billions of devices and the infrastructure of the smart, connected world – from the cloud to the network to the edge and everything in between. Find more information about Intel at newsroom.intel.com and intel.com.

Contact Chris Hunt

Solutions Architecture for
Retail and Consumer Products

Intel Corporation

chris.hunt@intel.com