



Lore IO

Lore IO for Data Standardization Solution Paper

Purpose

This article discusses the challenges in mapping datasets from multiple business partners into a common data format. It describes the costs of the prevailing approach of hard-coding data standardization logic into the systems that house and move the data.

Next, the article proposes a new mapping strategy through data virtualization, and describes its benefits. It explains how the Lore IO platform is ideally suited to virtualize data mapping processes, detailing the platform's core features and the benefits they provide. Finally, the article offers two examples of data mapping virtualization solutions that Lore IO has enabled for the retail and the travel and hospitality industries.

Overview

Brands that sell through retailers, travel and hospitality aggregators, companies with independent subsidiaries, and other organizations that exchange data with many business partners, face a difficult challenge: how to ingest datasets that come in each formatted differently according to its provider's custom business logic, and have these datasets standardized so that they can be compared, aggregated, and otherwise analyzed consistently?

It's a common practice for data providers and aggregators to have their own data formats, and it's hard to conform to a common structure especially when both sides often change their formats. Under such conditions it's difficult for aggregators to reconcile the data and be able to analyze it consistently across providers.

The problem with hard-coding mapping logic

To address this multi-source data mapping challenge, data aggregators have relied heavily on hard-coded ETL operations and data pipelines to standardize the heterogeneous datasets they receive. Others have turned to spreadsheets and manual processes to make sense of the data. Either approach applies standardization logic directly where the data lives: the former does so after the data is on-boarded; the latter just before it's analyzed.

These approaches have worked in the past, but with the explosion of business data today they are no longer viable. Hard-coding mapping rules introduce several problems:

- **Maintenance overhead:** The more standardization code you have, the more expensive and time-consuming it becomes to maintain it. As your code base grows over time, data lineage becomes murkier, and your ability to reverse-engineer the mapping process and ensure that it's still accurate for all cases diminishes.
- **Increased risk of errors:** Related to the earlier point, as your standardization code base grows, it becomes more difficult to test the data and catch errors early before they reach business users. And when you do catch errors, it may take a while to identify the source of the problem and fix it. The problem increases markedly when it comes to capturing custom data elements from various endpoints.
- **Data availability limitations:** Often, data providers offer only datasets that are easy and cheap for them to produce. Aggregators who require full business visibility are constrained in their analyses with limited data. Moreover, partner data may arrive in various forms (log files, CSVs, JSONs, database dumps, etc.), further complicating the aggregator's ability to access and blend the datasets.
- **Technology lock-in:** New technologies continuously appear on the scene, improving business productivity and flexibility. But when you hard-code mapping logic directly into your data infrastructure components, you lower your ability to rip and replace these components, given the cost of having to build new mapping logic. Moving from expensive databases to a scalable and cost-effective data lake, for example, becomes very cumbersome and expensive.
- **Lack of reusability:** hard-coded mapping logic is not easily reused in subsequent pipelines or ETL processes. Rather, it must be essentially re-implemented, which is costly and error-prone. Brands lose the opportunity to reuse meta-data and combine it in different ways to increase efficiency.
- **Increased cost of doing business:** Perhaps the biggest impact of hard-coded data mapping is that it creates friction where the business attempts to advance its mission: A brand that sells primarily through a retail channel, for example, is in the business of selling more goods through more retailers. Hard-coded data standardization prolongs the onboarding process of new retailers, increasing time

to value, and causing the brand to lose revenue and market-share to more nimble competitors.

Virtualizing data mapping

Data mapping is here to stay as the world is not about to adopt a unified way of defining every business data element any time soon. The good news, however, is that mapping doesn't have to be painful. A modern strategy for handling data mapping is to virtualize the entire process.

In virtualization, the software algorithms that execute data transformations are decoupled from the very systems that house the data. Furthermore, these algorithms are not persisted in code; rather, their logic is maintained in human-readable rules that non-developers can maintain on their own using visual interfaces without relying on IT.

Virtualization abstracts away all the complex semantics of how the data is captured, standardized, and cobbled together. It offers aggregators the agility to onboard new partners quickly, enhance the rules that logically blend the new provider's data with existing data, and provide the business with faster and more accurate analytics.

Standardizing data with Lore IO

Lore IO is a big data management platform that helps companies prepare data across disparate sources without the need for engineering to build ETL and data pipelines. Customers unlock the full value of their data by empowering business users to work with datasets that are hard to understand, reconcile, and blend.

Lore IO is conceptually a virtual data mart factory. Its proprietary algorithms enable customers to instantly capture and validate business logic in support of a wide range of use cases -- one of which is data transformation.

When a new data provider is on-boarded, the Lore IO platform uses its proprietary Data Scanner to understand the source data, regardless of the format or the system it's in. The platform builds a universal and virtual data layer that is automatically enhanced with pointers to the new raw data elements and includes all the transformation logic that the business requires. These virtual data columns and their transformations allow the Lore IO platform to query the raw data at any time, eliminating data moves and copies, and ensuring that query results reflect the latest changes in the raw data.

Lore IO monitors the raw data. When schema changes are detected, the platform makes the necessary adjustments in the data layer to point to the raw data elements correctly.

With the virtual data columns added, business users define virtual rules to standardize and blend the data. The rules are virtual since they're not persisted in code. They are kept in human-readable form that business users maintain. It's only at query time that Lore IO automatically creates the necessary code that it executes to create tables and views.

There are three types of rules that business users maintain in Lore IO for data transformation:

- **Taxonomy rules:** These rules map the columns and values of the partner's data with the aggregator's. For instance, a partner can describe their transactions as having two columns: a settlement amount and a type, where the type can be one of three options.
- **Reshape rules:** These rules specify how to pull data elements together from the partner's side, and how to distribute them on the aggregator's side. For example, a retailer might provide all transaction data in a single file, but the aggregator needs to split it into three tables, one for transactions, another for retailer data, and yet another for consumers.
- **Semantic rules:** These rules articulate the meanings of data elements and how the business uses them to describe its domain. For example, what constitutes a successful transaction? And how should its final settled amount be computed after accounting for refunds? Each data provider has its own semantics that makes sense in the context of its operations, but one that the data aggregator must reconcile with all other providers' data definitions.

Lore IO customers define these rules declaratively using a visual tool. It has a rich set of transformation functions that make standardization easy. For instance, customers can map columns and translate values to a standard set. Customers can pull data together from multiple files including XML, CSV, JSON, EDI etc. Common problems such as a different order of columns, renamed columns, changes to the values or types of columns can be handled automatically. Customers can also use a SQL console to describe more complex logic.

In addition, customers can build data validations and reports to monitor and check that all the standardizations happened correctly.

Once the standardization rules have been defined, Lore IO waits for new data from each source to show up. As soon as a new file or record is added or changed, the Lore IO Data Scanner detects it, applies the relevant standardization rules -- by dynamically generating

relevant SQL code and executing it -- and exports to the data to a standard format. Lore IO also executes validations on the view to ensure data quality and reliability.

Depending on the mapping rules used, the SQL that is generated automatically can be formidable. Lore IO will often dynamically create a complex multi-step transformation pipeline every time new data is seen. Customers have full visibility into the entire data lineage, enabling them to ensure data quality and to make quick changes when problems are encountered.

When new data is detected, if Lore IO Data Scanner finds any changes to the underlying schema such as a different order of columns, changed column names or value sets, etc. Lore IO can attempt to automatically generate rules to handle any exceptions. Alerts are generated whenever an exception is found with details on whether it was handled by the system or manual intervention is needed.

With data views executing, Lore IO customers export valid, standardized datasets to their existing BI tools or other destinations of choice. All Lore IO activities are logged and queryable. In addition, customers can define specific alerts for their data.

Virtual data mapping in Retail

Electronic Data Interchange (EDI) files are a well-accepted mechanism to exchange data between suppliers and their retailer partners. Product Activity Data (EDI 852), Purchase Orders (EDI 850), advanced ship notices (EDI 856), and other types of EDI files help suppliers exchange data electronically and consistently. This helps all parties save time and operational costs, reduce inventory levels, and respond faster to supply and demand changes.

EDI challenges

EDI files, however, introduce their own challenges that suppliers must regularly overcome. In general, EDI files are not easily read by humans, and their format changes across industry verticals. In retail, while EDI files have a lot in common, retailers occasionally choose to modify EDI segments based on their unique business rules and guidelines.

Examples of changing business rules that suppliers must accommodate include different lengths of period for functional acknowledgements, whether to consolidate invoices for complex orders, and how to adopt global traded item numbers (GTINs). Suppliers must accept and adapt to these variations to partner effectively with retailers or risk facing EDI non-compliance charges.

Beyond the operational complexity that EDI format variances introduce, they also complicate reporting. Suppliers struggle to ingest and map EDI columns from different retailer files into a centralized data format. And they must invest significant resources to standardize the data.

Given the data transformation complications, it becomes cumbersome for suppliers to aggregate retailer data and analyze product, category, and channel performance holistically. Forecasting and planning cycles take longer and are more prone to errors.

Lore IO's data transformation for Retail

To help suppliers standardize their retail partner data, Lore IO has developed a custom solution on top of its Big Data Management Platform. The solution converts EDI files into workable and queryable files. Since retailers often introduce variations in the EDI files, the solution automatically detects schema changes and applies the appropriate transformations.

Once new data is scanned, the Lore IO solution leverages prebuilt taxonomy rules that map and group new EDI datasets into one common output.

The solution includes data validation and alerting mechanisms. For instance, suppliers can quickly identify problems where new retailer files are not received or when entries in new product EDI files lack important information, such as current inventory levels, quantities sold, universal product codes, or global location numbers.

Lore IO enables brands that sell through retailers to on-board new partners faster and at much lower cost. No longer do data engineers need to write and maintain code to handle the variations in new EDI files. Brands can run comprehensive analytics on all retailer data and rapidly identify changes in product demand and performance.

Lore IO's data standardization solution is available as self-serve or as a managed service. Lore IO can assist customers in implementing the solution and have the customers run it on their own. Alternatively, Lore IO can implement and operate the solution -- either on premise or in a managed cloud -- on a short-term or long-term basis, depending on customers' needs.

Virtual data mapping in Travel and Hospitality

Travel aggregators collect data from numerous parties, including airlines, car rental companies, hotel chains, and more. The data they ingest varies markedly from one source to another. Each hotel chain, for example, describes its properties and rooms in each property using different columns and values. To ensure that consumers can easily

search and browse through available inventory, travel aggregators must standardize the data they ingest.

Travel partner onboarding challenges

Given the large variance in data structure and format, travel aggregators take months to onboard a new travel partner. We've seen examples where an integrated team of 25 implementation engineers and 5 DBAs from both the aggregator and their partner would spend 4-6 months to onboard the partner.

Delays in partner onboarding can have significant impact on the business: contracts get delayed and new revenue is missed when the aggregator misses onboarding deadlines, hurting the confidence of the partner in the aggregator's ability to execute.

Lore IO's data mapping for Travel and Hospitality

To help travel aggregators standardize their partners' raw data, Lore IO has developed a custom solution on top of its Big Data Management Platform. The solution virtually maps travel partners' raw data into a standardized presentation form, runs validation rules and error reports automatically, and exports standardized data views to target systems.

Lore IO offers a highly visual interaction for travel aggregators to first capture their master set of requirements and rules for validating various fields. This master framework template is then leveraged while connecting with each partner's raw data.

Lore IO solution offers a library of pre-built rule templates that make it easy to stitch the partner's raw data with relevant translations to map it into the aggregator's standardized data. Leveraging the rules and validations captured in the master template upfront, this simplifies the job of automatically summarizing any errors or escapes and generating validation reports and standardized export templates for every on-boarded partner.

By using Lore IO, travel aggregators can typically cut down new partner onboarding from several months to 2 weeks. Lore IO enables aggregators to avail new partner data through its existing applications faster, as well as surface data elements that did not pass validation. Travel partners can now describe their data to the aggregator without relying on DBAs.

As mentioned in the retail section above, Lore IO's data standardization solution is available as self-serve or as a managed service. Lore IO can assist customers in implementing the solution and have the customers run it on their own. Alternatively, Lore IO can implement and operate the solution – on premise or in a managed cloud -- on a short-term or long-term basis, depending on customers' needs.

Summary

Standardizing business data from multiple partners is a critical and common task that is only to become more important and frequent as economic developments offer the opportunity to partner with more stakeholders, and as these data providers continue to shape their datasets according to their own business logic.

Given the impact that data standardization has on business agility and performance, brands that aggregate data from multiple sources should consider carefully the infrastructure and workflows they put in place, and their ability to onboard new partners.

For many years, brands have been hard-coding their standardization logic in code that resides in the systems that housed and moved data around. Such strong coupling meant that brands had to spend significant time creating, maintaining, and debugging standardization code that was spread around several locations, with limited ability to ensure its quality and reusability. With complex standardization logic, brands have struggled to onboard new partners quickly, causing them to miss onboarding milestones and new revenue opportunities.

Lore IO offers a unique approach to data transformation through virtualization. Its Big Data Management Platform decouples and abstracts away standardization code, enabling business users to define standardization rules using a visual interface that converts the logic to code at query time. With this type of virtualization, brands increase their business agility, and onboard new partners faster.

About Lore IO

Lore IO is a big data management platform that helps companies prepare data across disparate sources without the need for engineering to build ETL and data pipelines. Our customers unlock the full value of their data by empowering business users to work with datasets that are hard to understand, reconcile, and blend.

Lore IO constructs logical data marts using proprietary algorithms abstract all the complex semantics of how the data is captured and joined together, enabling customers to instantly capture and validate business logic in support of a wide range of use cases:

- Travel and hospitality aggregators, brands that sell through retailers, companies that oversee independent subsidiaries or divisions, and other organizations that exchange data with many business partners use Lore IO to standardize data that comes in formatted by custom business logic
- IT teams use Lore IO to migrate their enterprise data warehouses to big data lakes, as well as avail BI on their big data
- Finance and revenue teams use Lore IO to reconcile their data declaratively in lieu of maintaining convoluted custom apps in spreadsheets and other applications that are not natively built for this purpose

Lore IO takes an agile approach to partnering with new customers. We seek to explore strategic projects that will make a material impact on the business and then structure a partnership that demonstrates value immediately and scales from there

Headquartered in Sunnyvale, CA, Lore IO was founded by veterans of WalmartLabs, and battle-hardened in highly performant transaction environments. Our platform launched in 2016 and is now used by medium and large companies across countries and industries. We are well funded by top-tier investors. For more information, contact us at hi@getlore.io, or visit our website at data.standardization.getlore.io.