The Rise of the Data Marketplace
Data as a Service

By Dave Wells
March 2017

Research Sponsored by Datawatch
About the Author

Dave Wells is an advisory consultant, educator, and research analyst dedicated to building meaningful connections throughout the path from data to business impact. He works at the intersection of information and business, driving value through analytics, business intelligence, and innovation. More than 40 years of information management experience combined with more than 10 years of business management create his unique perspective about the connections among business, information, data, and technology.

Knowledge sharing and skills building are Dave’s passions, carried out through consulting, speaking, teaching, research, and writing. He is a continuous learner—fascinated with understanding how we think—and a student and practitioner of systems thinking, critical thinking, design thinking, divergent thinking, and innovation.

About Eckerson Group

Eckerson Group is a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions. Through its reports and advisory services, the firm helps companies maximize their investment in data and analytics. Its researchers and consultants each have more than 20 years of experience in the field and are uniquely qualified to help business and technical leaders succeed with business intelligence, analytics, data management, data governance, performance management, and data science.

Report Sponsors

This report was made possible by, Alation, Datawatch and Podium Data.
Executive Summary

Remember the data warehousing promises of the past? Single source of the truth, reliable data when and where needed, and so on. These promises were rarely realized, and even when they were, they were never quite enough. Why the gap between expectations and reality? Much of the problem stems from the way we think about data. Enterprise data warehousing assumes that data is a thing to be stored, processed, and managed by technology organizations.

We must learn to think about data in an entirely different way—data as a service that is accessed, processed, analyzed, and reported by everyone who has a need for information. With data services presented through a “storefront” interface where data users can shop to find best-fit data to meet their needs, we can make the shift from the enterprise data warehouse (EDW) to the enterprise data marketplace (EDM).

Imagine the power and impact of a curated marketplace that serves a community of analysts “shopping” for data. Finding the right data quickly is essential in the age of self-service analytics. With ever-increasing numbers of people who analyze data, many of today’s business and data analysts depend on tribal knowledge and word of mouth to find the data they need. They don’t have visibility into the data sets that exist, their contents, or the quality and usefulness of each. As a result, they spend too much time finding data, evaluating data, and recreating data sets that already exist.

EDM can change all of this, empowering data seekers to find data sets, evaluate their fit for purpose, read reviews from others who have used the data, get immediate access, and prepare data for use. Recent technology innovations provide advanced capabilities for data cataloging, data preparation, data curation, and collaboration. These tools place the EDM within the reach of data-dependent organizations everywhere.
Rethinking Data

Data the Old, Hard Way

Do you recall the great expectations for enterprise data warehousing in the recent past? For much of two decades we pursued data warehousing as the solution to problems with data integration, quality, availability, and accessibility. The enterprise data warehouse (EDW) was designed and expected to provide reliable, right-time data whenever and wherever it was needed—a single source of truth for all of an organization’s information workers.

Most of today’s data warehouses are business critical, and some are mission critical. Most data warehouses deliver some business value, but only rarely have the promises of the EDW become reality. Even where those promises are realized, they’re usually insufficient to meet today’s data and information needs.

What happened that makes data warehousing insufficient to meet today’s data and information needs? Why, you may ask, did data warehousing fail? But fail is not the right word. In fact, data warehousing succeeded and still succeeds today to meet turn-of-the-century needs. The need to answer routine and recurring business questions still exists, and data warehousing is here to stay. But the world of data has changed, and the needs of business have changed. Data warehousing is part of the solution, but is not the entire solution.

EDW works best when data and information needs are stable, but in a dynamic and rapidly changing business world, data warehouse engineering can’t move quickly enough. EDW simply cannot keep pace with the rapid growth in variety and volume of new data sources.

Data in a New and Different Way

Much of the problem stems from the way we think about data. EDW treats data as a thing to be stored, processed, and managed as a technical asset. We need new and innovative ways to manage data as inventory and make it available and accessible to those who need it. We must think about data in an entirely different way: data as a service and a knowledge asset that is accessed, processed, analyzed, and reported by everyone who has a need for information.

Data Storefronts. Steve Jobs once said, “Let’s go invent tomorrow instead of worrying about what happened yesterday.” For those who depend on data, let’s invent a tomorrow where data sets are easy to find, understand, evaluate, access, and trust. Data as a service is a fundamental part of tomorrow’s data management practices. With data services presented through a “storefront” interface where data users can shop to find best-fit data to meet their needs, the enterprise data marketplace ( EDM ) replaces the EDW as the primary business-facing data resource.
EDM changes the game. It gives data seekers the services needed to find data sets, evaluate their fit for purpose, read reviews of others who have used the data, immediately access data, and prepare it for use. They can acquire data from all curated data sources—data warehouse, data lake, MDM repository, operational data store, ERP systems, legacy databases, commercial data providers, and even analysis data sets that are created by other analysts. (ERP and legacy data are notably valuable because they are difficult to access and frequently isolated from analytics projects.)

Recent technology innovations provide advanced capabilities for data sourcing, data cataloging, data preparation, data curation, and collaboration. These tools put the EDM within reach of data-dependent organizations everywhere. Data services are presented to users in a business-friendly context, and the digital marketplace is a familiar concept for everyone.

Finding the right data quickly is essential in the age of self-service analytics.

Finding the right data quickly is essential in the age of self-service analytics. With ever-increasing numbers of people who analyze data, many of today’s business and data analysts depend on tribal knowledge and word of mouth to find the data they need. They don’t have visibility into the data sets that exist, the contents of those data sets, and the quality and usefulness of each. As a result, they spend too much time finding data, evaluating data, and redundantly creating data sets that already exist.

Imagine the power and impact of a curated marketplace that serves a community of analysts “shopping” for data.

The Enterprise Data Marketplace (EDM)

A Place for Data Shoppers

Internal versus External Data Markets. The enterprise data marketplace is built for people seeking data sets to meet their analysis and reporting needs—data shoppers. The term data shopper describes the kind of user experience that data seekers should experience—a true digital marketplace experience—and is not intended to imply exchange of money. In the self-service world, we think first of people seeking data, but it is important to note that downstream applications and analytic algorithms may also be data shoppers.

It should be noted that several commercial and open data markets already exist, and they are useful for acquiring generally available external data about business, society, the economy, health, the weather, and so on. Commercial and open data can enrich and add context, but data shoppers rely primarily on enterprise data that is generated internally.

The EDM is a company store for data, one that is accessible only to employees and invited guests.
The EDM is distinctly different from commercial and open markets specifically because it is a marketplace for enterprise data. Think of it as the company store for data—a store that is accessible only to employees and invited guests. EDM is to data markets what a corporate intranet is to the broader Internet.

**The Business Case for EDM**

Creating a faster, easier way to find and access data makes a compelling business case for EDM, but it goes deeper. Access, speed, usability, and cost savings are all possible with EDM. The ability to find and access data quickly is directly tied to speed of analysis and insight. Curation and extensive metadata enhance usability and help users find the best-fit data for each analytic use case. The cost of EDM is lower than the high (and often unreported) cost of staffing and projects for IT-centric data delivery. Accounting for cost of delays, failures to deliver, and lost opportunities makes the EDM case even more persuasive.

**Marketplace Characteristics**

Commercial marketplaces exist in many different forms—at the highest level, brick and mortar versus digital—and there are substantial differences among digital markets. Each has different characteristics related to product selection, inventory management, and shopper experience. Consider, for example, the differences between shopping at Amazon, Etsy, eBay, and Uber. From search and browse to the delivery of products and services, each interacts with shoppers differently. Data is especially well suited to a digital marketplace because data itself is digital. The advantages of brick-and-mortar shopping – physical touch and instant access – simply don’t matter with digital products.

The EDM is a digital marketplace that is geared to data. Four characteristics of digital marketplaces are especially important in the data world:

- **Categorization** organizes the marketplace to simplify browsing. For example, a shopper seeking budget data doesn’t need to browse through unrelated data sets about customers, employees, or other data subjects. Categories complement tagging and smart search algorithms, offering a variety of ways to find data sets.

- **Curation** is active management of the data sets that are available in the EDM. Curation selects and qualifies data sets, describes each data set, and collects and manages metadata about the collection and each individual data set.

- **Cataloging** exposes data sets for data shoppers, including descriptions and metadata. The catalog is a view into the inventory of curated data sets. Rich metadata and powerful search are important catalog features.
- **Crowdsourcing** is the equivalent of a social network for data. Data shoppers actively participate in cataloging, curating, and categorizing data. This virtuous cycle (a chain of events that reinforces outcomes through a feedback loop) continuously improves the quality and value of data in the marketplace.

**EDM Components Framework**

Like a commercial marketplace, an EDM is made up of supplier communities, shopper communities, products and categories, and the marketplace activities that connect shoppers with products. Supplier communities are the source of products. Shopper communities are the consumers of products. Product categories describe and manage the product selection. Marketplace activities connect shoppers with products and carry out the commercial functions of the marketplace. A data marketplace has similar components as shown below.

<table>
<thead>
<tr>
<th>Commercial Marketplace</th>
<th>Data Marketplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplier communities</td>
<td>Data providers</td>
</tr>
<tr>
<td>Shopper communities</td>
<td>Data shoppers</td>
</tr>
<tr>
<td>Product categories</td>
<td>Curated data sets and processes</td>
</tr>
<tr>
<td>Marketplace activities</td>
<td>Data services</td>
</tr>
</tbody>
</table>

For the EDM, data providers (both internal and external) are the supplier community. Data shoppers—people, applications, and algorithms—form the shopper communities. Curated data sets and processes are the products of a data marketplace. Data sets are the primary product. Reusable data preparation processes enrich the product selection, essentially packaging the data for reuse. Data services are the means to connect data shoppers with the catalog of data sets. Data services include acquiring, preparing, cataloging, curating, tracking, and governing data. (See Figure 1.)
**EDM Components in the Virtuous Cycle**

The real power of the data marketplace becomes clear when it creates a virtuous cycle—going beyond simple data shopping to data consumption that drives a feedback loop. (See Figure 2.) In this loop, consumers prepare data for analysis (additional preparation steps beyond those provided through data services) and analyze the data. These activities produce new data sets and new metadata, continuously enriching the data services and delivering new value for data shoppers.

**Figure 2. EDM in the Virtuous Cycle**

---

**Data Providers**

The supplier community includes three kinds of data providers.

**Internal data providers** are most prominent among the data suppliers. These are the people and systems within the company who supply enterprise, departmental, and local data for analysis and reporting. The internal data supply may include:

- Transaction systems and databases
- Data warehousing systems and databases
- Data lake systems and databases
- Master data systems and databases
- Departmental, local, and individual databases
- A variety of spreadsheets
- Data sets produced by data science and analytics projects
Open data providers are organizations that provide access to free external data. Open data can be used to complement, enrich, and contextualize internal data. Common open data categories include government and political data, geospatial data, social data, weather data, sports data, university research data, and news. There are many open data providers—perhaps 50 or more—ranging from data aggregators to topical specialists. Some of the more widely known include data.gov, Infochimps, Google public data, Weather Underground, ArcGIS open data, and OpenSecrets.

Commercial data providers provide fee-based and subscription-based external data. As with open data, commercial data is used to complement, enrich, and contextualize other data. Commercial data may be more targeted and tailored than open data, improving the ability to blend external and internal data. A few examples of the many commercial data providers include Factual, Nielsen Data sets, Big Data Exchange (BDEX), and LexisNexis.

Curated Data sets
Data sets are the core of the data marketplace; they are the product that data shoppers are seeking. Curation is the process by which data sets become ready for consumption in the marketplace. Categorizing data sets is an essential part of curating for the marketplace. You can read more about data curation in our report on Big Data Management Software for the Data-Driven Enterprise.

Categories
Data sets need to be organized by categories that are meaningful to data shoppers. At a very high level, data can be categorized as raw data (unmodified data in the form in which it is received from providers), integrated data (data from multiple sources that is reconciled to resolve disparity and inconsistencies), and aggregated data (data expressed in summarized forms, often to inform about groups instead of individuals). Some data sets are both integrated and aggregated. Data shoppers, especially those seeking data for a particular analytic use case, may have specific requirements for raw data, or for data with specific integration or aggregation characteristics.

Subject-oriented categories are especially important and are usually among the first search criteria. Shoppers typically seek data about known subjects of analysis and reporting—customers, products, competitors, and so on.

Metric categories improve the precision and specificity of search for performance reporting and for analytics. They describe quantitative and directional characteristics such as value, quality, loyalty, and propensity, and also include categories for dimensions of analysis—geography, demography, and so on. Such categories relate to the ways that analysts think about data.

There are many useful ways to categorize the same data: by quality, subject area, time frame, line of business, and more. Categories are dynamic and ever evolving, so crowdsourcing of categories is important. After all, who better to identify useful categories than data shoppers?
Again, categories are instrumental in finding the right data. A rich marketplace uses many different categories and categorizes in orthogonal ways. Common categories can be defined by a curator-managed taxonomy that provides a hierarchy for browsing and filtering. They are complemented with tags that are used for searching. To support selective browsing and highly granular filtering, data should be categorized from at least four perspectives—content (raw, aggregated, integrated), subjects, metrics, and dimensions. (See Figure 3.) Assigning data sets to multiple fine-grained categories may seem on the surface like an overwhelming task. Yet, it becomes manageable—even fast and easy—with intelligent tools that use algorithms to infer categories.

*Categories enable browsing. Tags support searching. Together they support individual shopper preferences for seeking data sets.*

**Figure 3. Categorizing Data sets**

*Tags.* Categories are based on a formal taxonomy; tags complement the categories to help shoppers find data sets quickly. Tags can be discovered automatically by data cataloging tools, or they can be interactively crowdsourced as analysts work with data sets.

Categories enable browsing of data set inventory. Tags enable searching of the inventory. Together they support individual shopper preferences for seeking data sets.

**Smart search.** Search algorithms that infer facets of data sets create a particularly rich user experience of data set discovery. Combining inferred tagging with curated tags and categorization provides robust data seeking functionality.

**Curated Processes**

Data preparation processes may also be curated and become part of the overall inventory. Making data preparation processes available to be reused accelerates the path from data to analysis and improves the quality and consistency of data sets used in analytics. Both data preparation procedures (individual steps of data manipulation) and data preparation workflows (linking and sequencing procedures) are valuable as curated processes.
**Backroom Services**

Delivering data as a service requires both a storefront and backroom services. Let’s begin with a look at the backroom, where many interesting and important things happen. Unlike a physical marketplace, a data marketplace can offer backroom services directly to shoppers, as long as the steps are automated and well-governed.

*Data acquisition* connects to data sets and obtains data for exploration, preparation, or publishing. Acquisition demands the ability to connect to many different kinds of data sets—SQL, NoSQL, flat files, JSON, XML, geospatial, etc.—and to deliver data to exploration, preparation, and publishing processes. Known data quality is a marketplace imperative, so data profiling and verification are important acquisition steps.

*Data preparation* is the work of combining data from multiple data sets and making that data ready for analysis and reporting. Data preparation is performed to improve and enrich data, to format data, and to blend data from multiple sources. Self-service data preparation is an integral part of data-as-a-service.

*Data curation and cataloging* manages the inventory of data sets and processes by collecting and maintaining the metadata that is the critical understructure of the EDM. Data curation is responsible to oversee a collection of data assets and make it available to and findable by data shoppers. Curation also involves tagging data that is sensitive to privacy, security, compliance, and personally identifiable information (PII) constraints.

*Tracking* gathers metadata about data lineage and data usage, which data shoppers use to judge whether data sets are trustworthy and useful. Lineage documents the sources and steps by which data is created (parent lineage) and the downstream uses and derivatives of a data set (child lineage). Usage tracking collects metadata about how frequently data sets are used, as well as who use them. Usage metadata is valuable to inform recommendations and suggestions – an intelligent marketplace that learns from usage.

**The Storefront**

The storefront comprises shopper-facing services that shape the user interface and the shopper experience. These services expose the catalog of data sets to the shopper within the limits of governance controls. Ideally, a shopping cart concept enables shoppers to select data sets from the catalog and to explore and view them. The ability to add and remove items from the cart supports the real-life processes of iterative discovery and exploration.
Data shoppers interact with data services in many ways. (See Figure 4.) They must be able to:

- Browse the data catalog navigating categories and hierarchies
- Search the data catalog using keywords, business terms, and natural language
- Transform data in a variety of ways needed for data preparation
- Explore data to understand content and judge quality
- Access data needed for analysis and reporting
- Transform data to improve, enrich, format, and blend
- Publish data to consumers, both people and systems
- Review and rate data to inform other shoppers about their experiences

As the EDM matures, it will support context-aware recommendations and suggestions based on historical shopping patterns and behaviors. For example: “People who looked at this data set also viewed these data sets,” or “People who acquired this data set also used these related processes.”

**Data Governance**

EDM and data shopping do not diminish the need for data governance. The catalog view is ideally constrained by shopper permissions, allowing each shopper to see only the data sets authorized for their identity and role. Redundant security administration is avoided when the marketplace is built for graceful interoperability with Active Directory, LDAP, and other common authentication and authorization technologies.

Privacy-sensitive data, compliance-sensitive data, and data sets containing PII must be tagged (by curating activities) to label sensitive data, restrict access, and mask and anonymize data. Usage-tracking metadata is a valuable source of information for governance audits.

**Collaboration and Crowdsourcing**

Because it serves a community of shoppers, the EDM has human and cultural dimensions as well as data and technical components. Crowdsourcing is a powerful way to enrich the catalog and to formalize and share tribal knowledge.
Good cataloging tools automate much of the metadata collection, but they are best at discovering technical metadata. Business metadata is typically acquired by tapping into the knowledge of individuals. Crowdsourcing of business metadata is an effective way to capture that knowledge, especially when it is woven into the day-to-day processes of using the catalog. Some data cataloging tools directly support user entry of business metadata into the catalog. Advanced catalogs can parse business metadata from collaboration and social interactions around data that occur during team projects. Manual collection and entry of business metadata is a last step to complete the metadata that isn’t easily automated or crowdsourced.

User reviews and ratings are also valuable. Capturing the impressions, experiences, and opinions of those who have worked with a data set helps to inform others who are interested in it. The review and five-star rating systems that work so well in online stores fit neatly into the EDM. Even data can be crowdsourced, growing the inventory by curating user-created data sets produced through data preparation and data analysis activities.

Social interactions, such as “following” other shoppers, may also be built into the marketplace. This creates an effective way for the community to share, inform, and learn. Following experienced and talented analysts can improve the speed and quality of analysis and accelerate the process of bringing new analysts on board.

Gamification as a marketplace feature is an especially powerful way to fully engage data shoppers as a community. Recognizing marketplace contributors and rewarding valuable contributions helps to drive a dynamic marketplace and an enthusiastic shopper community. Contributors may offer business metadata, helpful reviews, tips for data preparation and blending, and even new data sets vetted by a curator. Rewards can range from company-wide recognition to gift cards and financial rewards.
EDM Technologies

The EDM scope of functions is extensive, ranging from data ingestion to data set publishing—with many processes in the middle. The scope of technology is equally broad. It must span the continuum from data providers to data shoppers. To connect shoppers with providers, it must support a curated and cataloged inventory of data and processes, and a robust set of data services.

The EDM can be built by blending best-of-breed tools, building upon existing technologies, or selecting one tool that spans the needed functions and capabilities. Existing technical infrastructure is a key consideration. The EDM must be capable of working with existing security and data quality technologies, synchronizing metadata, and integrating with deployed data preparation and data analysis tools. The EDM needs pipeline tools for managed data flow, data cataloging to manage the inventory and the metadata, and data preparation to improve, enrich, format, and blend data. (See Figure 5.) Leading data lake management tools package many of these functions in a single tool set that operates gracefully with Hadoop.

Figure 5. EDM Technologies

Interoperability is a must when selecting EDM technology. Whether you choose a single data lake management tool or combine best-of-breed solutions, data supply and data consumption must not be constrained by technology. The EDM must work with on-premises data sets and cloud-hosted data, and work with Hadoop while minimizing its complexities. It must be able to source multiple databases, both relational and NoSQL, and it must be able to publish to a variety of consuming technologies for reporting, visualization, dashboards, and analytics.
How Do You Get the Magic of EDM?

The enterprise data marketplace is exciting, innovative, and an engaging way to connect people with data. It is the next step in the maturation of self-service data organizations. Buying the technology is only the beginning. You must add inventory, curation, and services, and you must engage the community of data shoppers.

Much of EDM success is cultural—rethinking how we work with data. IT organizations must evolve to modern data architecture. Replace the linear architecture of EDW with the virtuous cycle architecture of EDM. Line-of-business organizations must stretch their data skills. Go beyond reporting and analyzing data; embrace the feedback loop by becoming conscious and conscientious providers of data and metadata.

EDM is a relatively new concept, but it is the way of the future, and you can’t start building too soon. Cataloging your data is a good beginning. If you have self-service analytics tools, you need a data catalog. If you’ve built a data lake, you need a data catalog. With cataloging under way, the next step is curating. Recognize “data curator” as a role on par with data steward. Identify and train data curators to build the inventory of curated data sets. Then tap into the collective knowledge of your analytics community to crowdsource business metadata and formalize tribal knowledge. Harness the power of three Cs of EDM—cataloging, curating, and crowdsourcing—and you’ll be well on the way to your own enterprise data marketplace.

Need help with your business analytics or data management and governance strategy? Want to learn about the latest business analytics and big data tools and trends? Check out Eckerson Group research and consulting services.
About Datawatch

Datawatch has been at the forefront of innovation in the self-service data-preparation industry, for over 20 years. We were the first to access dark data locked in PDF documents and semi-structured data with Monarch. We pioneered point-and-click access to critical business data; drag-and-drop access to web data; over 100 predefined data preparation functions and direct exports of analytics-ready data to any data visualization tool or BI platform with Monarch Complete. And, led the market with automation and governance integrated with data preparation in Monarch Server.

Data socialization is the next evolution in self-service data preparation. Datawatch Monarch Swarm, the industry’s first data socialization platform, combines data preparation, cataloging, stewardship, automation and governance with attributes common to social media platforms - user ratings, recommendations, discussions, comments and popularity - to make better decisions about which data to use. Business users can search for, share and reuse prepared, managed data to achieve true enterprise collaboration and agility. Learn more at www.datawatch.com