Enterprise Data Stores:
The Tradeoff between Agility and Control

Bottom Line Up Front: Large data-driven companies need multiple kinds of data stores to handle the volume and variety of data available to them. Master data management systems, enterprise data warehouses, and cloud-based big data platforms each have their own performance characteristics and each play a key role in data-driven decision-making. Large data-driven companies need all three.

INTRODUCTION

Today, companies are barraged with more data than ever—all with varying business value, velocity or volume. Some data, such as your customer list, have a very high value, but content and schema that change relatively slowly. Other data, such as anonymous website visitors' clickstream logs, are lower value, but have much higher velocity and volume. Multiple technologies are needed to cost-effectively handle such a wide variety of data, since one size does not fit all. This brief compares three data store technologies: master data management (MDM), enterprise data warehouses (EDW), and cloud-based big data platforms.

MASTER DATA MANAGEMENT

MDM creates a single source of truth for non-event meta data that is critical to many departments. Two common examples of such data are a customer list and a product list. Because so many departments depend on this data, it is highly beneficial to get these lists right, once, in a central MDM system. Key customer-facing functions such as multi-channel customer support and direct marketing require that multiple departments use the same lists in order for processes to run smoothly. This leverage, and cross-department dependencies, justify the expensive ETL and data cleaning of a MDM.

MDM's strength is its centralized control, which restricts change to critical, widely used data. For example, a change to the schema for customer billing addresses would have ripple effects throughout the company. These types of changes are worth controlling centrally with MDM.

ENTERPRISE DATA WAREHOUSE

EDWs, in general, provide a single source of the truth for transactional and event data used by multiple departments. Common examples include supply chain data, sales, returns, and direct marketing. EDWs integrate data so that it's clear when supply chain and returns are referring to the same product, or when sales and marketing are referring to the same customer. Data is "democratized" when available and easily usable by many departments, and by analysts of different skill levels, via reporting tools such as SQL, SAS, OLAP, and business intelligence suites.

EDW's strength is that it integrates events from different departments to create a single, easily accessible view of critical transactions for reporting and analytics. It "connects the dots" between supply chain, sales, marketing, and other departments within a company.

CLOUD-BASED BIG DATA PLATFORM

Cloud-based big data platforms integrate a wide variety of high-volume data, provide APIs for real-time decision-making, and conventional business intelligence interfaces for analytics. These new platforms drastically lower storage, ETL, and computation costs. Key characteristics include:

- Very low storage costs, enabling management of high-volume data such as clickstream, in-store wi-fi telemetry, and online display ad impressions, which may be too expensive to store in an EDW

- Semi-automated data integration so that new data feeds can be added quickly

- Very low computation costs, enabling ad hoc queries of big data, and frequent updates of hundreds of customer models

- Real-time data integration and API response, supporting a wide variety of applications and use-cases of real-time decision-making

- Distributed and replicated in multiple data centers, enabling high availability and real-time APIs

- Easy integration with suppliers and partners via web technologies and emerging "big data" transfer technologies

Gartner recently summarized the demand for these platforms as follows:

"Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.[1]

---

[1] Svetlana Sicular, "Gartner's big data definition consists of three parts, not to be confused with three 'V's,"

CONCLUSION

Each of these technologies has a role in data-driven companies: MDMs control the vital few data feeds that are critical to many departments, EDWs integrate transactions that inform multiple departments, and cloud-based big data platforms integrate a wide variety of high-volume data. The low cost of big data platforms has another important benefit: innovation.

Innovation is enabled by agile, uncontrolled processes that foster experimentation with new, disparate data sources. MDM is too controlled to support a high pace of innovation, while EDWs are too integrated to enable experimentation with new data sources. Cloud-based big data platforms present an ideal platform for innovation with big data.

Consider an analogy. On a recent flight, I sat next to a pilot, and we got talking about the new onboard Wi-Fi available to passengers. I showed him how the MyRadar app on my smart phone allowed me to get real-time weather radar for our destination airport. He said he wished he had real-time weather radar in the cockpit but that they aren't allowed to use onboard Wi-Fi in the cockpit for fear that it might interfere with equipment used to fly the plane. The cockpit, like MDM and EDWs, is restrictive to change because side effects can be costly. Cockpits have a slow pace of innovation. Cloud-based big data platforms, in this analogy, are like smart phone apps: build lots of use-cases, experiment and see what works.

The tradeoff between agile innovation and central control impacts diverse fields including political science, economics, organizational development, and computer science. While some processes are better suited for central control, others thrive and improve with agile innovation. MDM and EDWs provide control but cloud-based big data platforms provide agility.
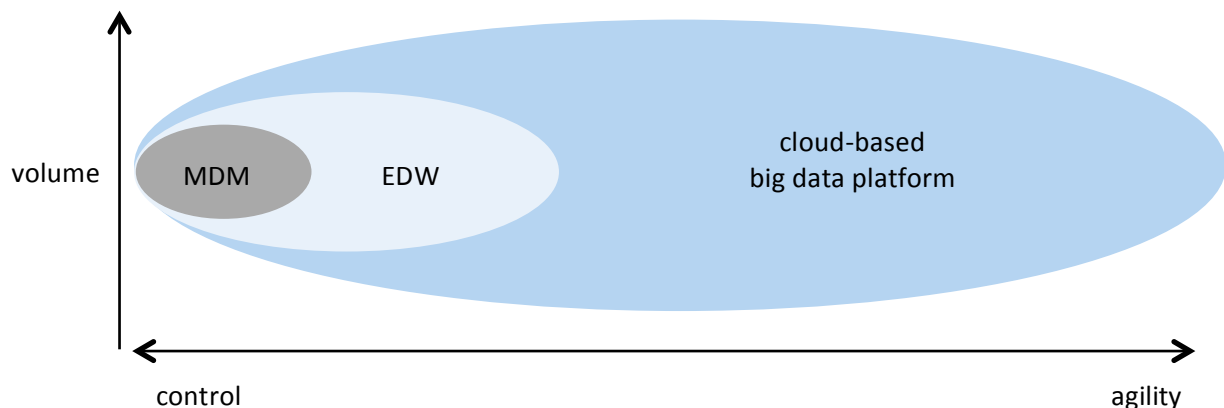
Forbes.com, March 29, 2013. See also Douglas Laney, "The importance of 'big data': a definition," Gartner, June 21, 2012; and Doug Laney "3D data management: controlling data volume, velocity and variety," META Group, February 6, 2001.

Day to day, data *records* tend to flow from controlled to more agile stores so that, for example, ad hoc analytics in a cloud-based big data platform can utilize clean product feeds generated by MDM. Over time, however, data *types* tend to flow from agile to controlled stores as the business value of new data are proven and the data are used by more departments.
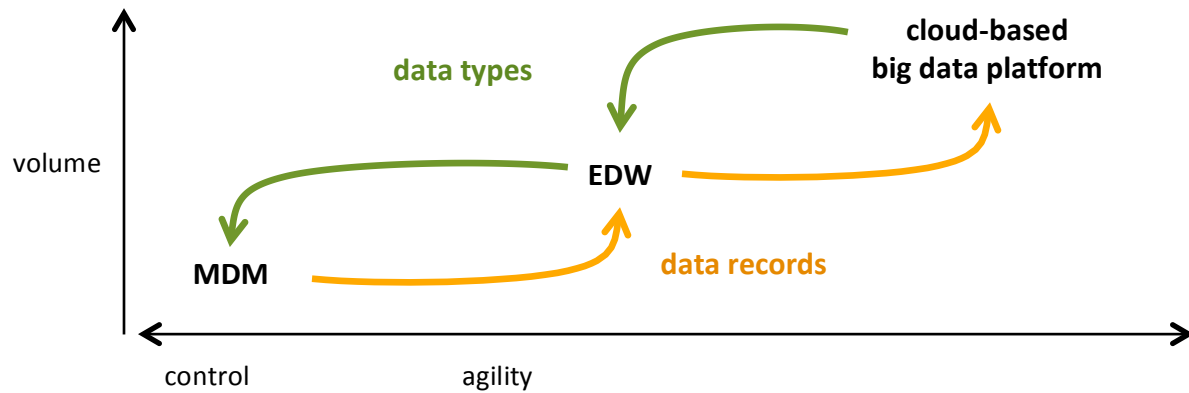
Of course one size does not fit all and all three technologies are needed by large data-driven companies.

IMPLICATIONS FOR RETAILERS AND BRANDS

Data is a new asset class.[2] However, like stocks and bonds, each data type has its own volume, volatility, and value characteristics. Technology has made vast volumes and varieties of data available to retailers and brands, far more than they could ever cost-effectively use. Retailers and brands should begin characterizing their data types, managing them as a portfolio, and considering which data store technology is appropriate for each type. Advanced companies will create data development pipelines, pipeline metrics, and criteria for key pipeline steps such as evaluating a new data type, moving a data type from a big data platform to an EDW, and retiring a data type. Here are some examples of emerging data types being evaluated by retailers and brands:

Retailers	Brands
• display ad impressions • combined online and in-store sales • customer social media profiles and activity • real-time inventory levels • store sales associate hours and profiles • in-store wi-fi traffic • in-store customer tracking	• display ad and promotion impressions • combined online and in-store sales • customer social media profiles and activity • real-time in-store inventory levels • in-store customer dwell time

---

[2] For example, see "Personal Data: The Emergence of a New Asset Class," World Economic Forum, 2011