syncsort

Enterprise Data Marketplaces

Enabling a Data-Driven Culture

Introduction

We all know there is a data explosion, but did you realize we now create 2.5 quintillion bytes of data every day, with 90 percent of the world's data created in the last two years?[1] Its staggering. Every search, click, view and stream, from sources like Google, YouTube, Twitter and Netflix – not to mention data from IoT devices – continue to drive that number up.

"Standard" data is also growing. This includes personal, health, financial and operational data generated by a global population of 7.6 billion that increases by 83 million each year.[2]

To outsmart the competition and even change the rules of the game, companies are relying on their ability to use analytics to unlock the insights buried deep within this data. Data analytics is the new disruptor.

So, what happens when you have supply and demand? You have a market. In the past few years third-party data marketplaces, often provided as Data as a Service, have taken off. But most organizations already own the data most relevant to their business – data pertaining to their own customers, transactions, products, etc. That's why the most successful organizations are applying the concepts of external data markets to create enterprise data marketplaces, where users can easily find and access data from across the company that is clean, trustworthy and auditable. Let's take a closer look.

The Rise of the Enterprise Data Marketplace

Businesses have always relied on insights from data to make better decisions. What's changed is the volume, velocity and variety of data, and how we use it. Every part of the business demands sophisticated data analysis not only to report on the past – but also to predict the future.

To store, process and manage it all, we built data warehouses, data lakes and data marts. Each has tradeoffs. Data has outgrown the warehouse, is polluted and chaotic in a lake, and inconsistent across marts. What's more, each approach assumes data is owned by IT, not something to be accessed by anyone who needs it. Each department often needs to be able to choose from the company's many datasets, combined in different ways. Enter the enterprise data marketplace.

Enterprise data marketplaces overcome the limitations of previous solutions and give you the best of each in one central repository: the volume and variety of the data lake, the veracity and auditability of the data warehouse and the velocity and specificity for purpose of the data mart. Analytics teams and business users can shop and find the data they need, prepared and combined for ever-expanding applications.

Curated enterprise data marketplaces help organizations be truly data-driven.

What an Enterprise Data Marketplace Can Do for Your Business

The retention team needs 10 years of customer data for churn prevention and customer lifetime value calculations. Risk management needs three years of accident data to set insurance rates. Order fulfillment needs current address geo-location data to make sure shipments arrive when promised. E-commerce needs a year of customer purchase data to train the machine learning algorithms that drive a recommendation engine. Research needs data from yesterday to track the effects of a new drug. Security needs information from three seconds ago to see if a transaction is fraudulent before it completes.

Business value today often depends on the ability to know more and deliver more, sooner. Enterprise data marketplaces elevate data reliability, flexibility and availability to new levels to help you maximize value.

Reliability: Enterprise data marketplaces provide a centralized location for all data and ensure that as it comes through the pipeline it is cleaned, standardized, verified and ready for various types of analytics.
With their Amazon-style data marketplace, Guardian Life Insurance reduced time-to-market for analytics projects with data that is centralized, standardized and reusable.

Flexibility: Different people need different subsets of data from different sources – from legacy to streaming and everything in between. Enterprise data marketplaces pull data from different places and allow you to pick and choose the data you need depending on what you want to accomplish.
Progressive Insurance has over 50 data sources in their data marketplace and as each department's needs change, or another department or team asks, they add more.

Availability: Enterprise data marketplaces empower analytics teams to create new data schemas and queries on their own, trusting that the data is clean and fresh, or they can request data combined in a certain way by IT.
Analysts at Symphony Health no longer wait for requests for specific data schemas, or data sub-sets, to work their way through the IT team's queue. Since building their data marketplace, they can access data within minutes to get answers faster.

5 Potential Roadblocks in Your Path to an Enterprise Data Marketplace

With more data widely available across your organization, many teams can benefit. But there are challenges along the way that can hamper your efforts to gather, transform and maintain that data so that it is ready for self-service access. In the spirit of "forewarned is forearmed" here are five potential roadblocks in your path to creating an enterprise data marketplace.

1) **Scattered and difficult to access datasets.** The data you need may be trapped in mainframes, located in databases or your data lake, stored in the cloud, streaming in from systems like Point-of-Sale terminals, Automatic Teller Machines or social media, or all of the above. Not only are the data formats and files incompatible with each other, but also with the target system that will host your enterprise data marketplace, making it difficult to gather and prepare the data for analysis.

2) **Cleaning massive volumes of data.** We're all familiar with the phrase, "garbage in, garbage out." Combining data aggravates problems inherent in almost all datasets. Data records with incompatible layouts, data in the wrong field, incomplete data and spelling errors are just some of the inconsistencies that need to be addressed to ensure you're getting a complete and accurate picture. Data cleansing and preparation routines that have been part of trusted data architectures for decades must be reproduced at modern scale, but most data cleansing tools are not designed to handle massive volumes of data.

3) **Combining and deduplicating data.** One dataset by itself is useful, but where analytics is concerned, accuracy often comes from looking at multiple datasets, which means you need to deal with matching records from different datasets that pertain to the same person, company, product, etc. With billions of records, how do you identify the same entity? You need sophisticated multi-field matching algorithms and a lot of compute power to compare everything with everything and quickly reveal matches across massive datasets.

4) **Keeping the data fresh.** Tracking and detecting changes in traditional datasets that may have thousands of changes a day or more requires high speed change data capture, and a robust distributed pipeline. When you're dealing with streaming data which changes continuously, a data capture capability that can keep up is essential. Regardless of the source and type of data, you need the ability to cleanse, transform and update combined datasets in the enterprise data marketplace in as close to real time as possible. Different use cases have different data refresh requirements. Less than a second may be required in some cases, while hourly is more than adequate for others.

5) **Tracking data lineage.** Data changes must be auditable to ensure the integrity of decisions and predictions you make based on the data, and to comply with regulatory requirements. You need to know where the data came from and have confidence that no changes have been made that affect accuracy. When training machine learning models, tracking what changes were made to data to make it fit for training is essential so that those exact same steps can be reproduced in production.

A Way Forward

Gartner finds that by the end of 2019, 90 percent of large organizations will have hired a Chief Data Officer.[3] Data-driven cultures must begin at the top and permeate the organization so that everyone is empowered to do their jobs more effectively with better data and insights. However, this doesn't mean you need an army of programmers. In fact, quite the opposite. Despite the potential roadblocks in your path, with the right technologies you can create an enterprise data marketplace without coding everything. Much of this work has already been done by tool vendors, and it is a huge waste of resources to continually reinvent those wheels.

Using point and click tools, projects that required weeks for a team of developers to accomplish, can be completed in a few minutes or hours by one person. Creating a data pipeline that brings in and prepares data for analysis should not take weeks or months. The company needs to get the ROI from that data as soon as possible. And keeping data fresh by regularly syncing with the source should be a standard part of the process. Moving data is never a one-time project.

Data scientists can spend less time cleaning data and more time on the higher value tasks you hired them to do, including creating and refining analytics models to improve accuracy. Analytics teams are empowered to create new data schemas and queries on their own to answer more questions sooner. And business users can have access to the insights they need for faster decision making.

Conclusion

Organizations are embracing enterprise data marketplaces to drive their strategies forward and are realizing business value.

Symphony Health turns large volumes of anonymized health information into actionable insights that their clients use to enhance the patient journey. By creating an enterprise data marketplace, the process is faster, less expensive and available to more users.

Progressive pioneered reduced rates for low-risk drivers, comparison shopping on the web and Name Your Price® car insurance shopping. With data pipelines that supply their enterprise data marketplace with an ever-expanding number of datasets, scrubbed, de-duplicated and ready to use, they continue their tradition of analytics-based innovation.

Guardian delivers on its commitment to bring peace of mind to insurance with an enterprise data marketplace that makes up-to-the-minute current data assets available to the entire firm. This data provides fast and accurate funding options to meet each customer's unique needs, leading to award winning, customer-focused service.

There's no end in sight to the types of data and analyses businesses will need to thrive. Thankfully, with an enterprise data marketplace where users can easily find and access data that is clean, trustworthy and auditable, your organization can be prepared.

Sources:

1. https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN&
2. https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html
3. https://www.gartner.com/smarterwithgartner/keys-to-success-for-chief-data-officers/

About Syncsort

Syncsort is the global leader in Big Iron to Big Data software. We organize data everywhere to keep the world working – the same data that powers machine learning, AI and predictive analytics. We use our decades of experience so that more than 7,000 customers, including 84 of the Fortune 100, can quickly extract value from their critical data anytime, anywhere. Our products provide a simple way to optimize, assure, integrate and advance data, helping to solve for the present and prepare for the future. Learn more at syncsort.com.

www.syncsort.com