How data curation creates business value

ROZETTA
TECHNOLOGY

Overview

- Data is often an organization's most valuable asset, and there needs to be a dedicated commitment to maximizing the value it can generate. Organizations are increasingly demanding more from their data, supporting data-driven decisions and creating value through innovating new products and services.

- Data curation is the capability to ensure data is the best available, discoverable and capable of enhancement through applying data engineering, data science, and advanced analytics.

- Data curation is a critical skill set for organizations to achieve value maximization from their data assets, and it remains a hidden and distributed discipline within most organizations. If Data Curation is formalized and centralized then enormous efficiencies and value can be shared across all organizations although particularly large scale complex organizations.
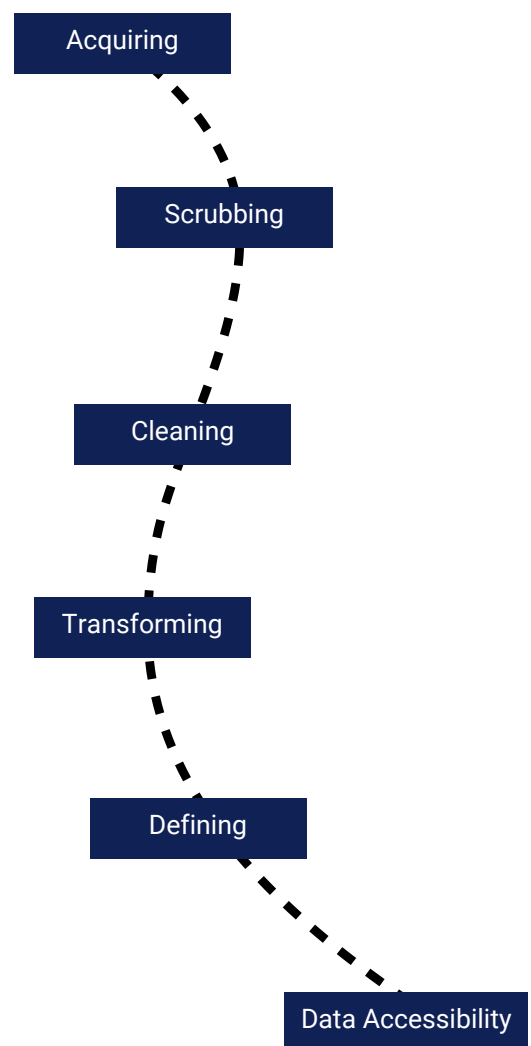
Data Curation

At a high level, data curation involves collecting, organizing, normalizing, preserving, enhancing, and maintaining data for current and future use. It is an integral part of data management. The technical skills required include acquiring, scrubbing, cleaning, transforming, organizing, defining and making data accessible.

Organizations expect data curation to include the users' needs, to know and communicate knowledge of the data, and to make it available for analysis, risk management, compliance oversight, workflow integration, reporting and decision-driving MIS platforms.

Data curation is not just an IT task; it requires organizational involvement to ensure the outcomes align with business strategies and goals. Much of the high-value data generated by organizations is derived data assets generated by user groups across the enterprise and data curation is responsible for supporting these data assets.

While all of this is reasonable, curation standards need to set enduring expectations, that is, defining what data curation delivers to data users within an organization.

Acquiring

Scrubbing

Cleaning

Transforming

Defining

Data Accessibility

Technical skills required as part of data management

Data curation produces valuable Data Products

Derived data products combine data from multiple sources and are often enhanced through analytics, to create unique data assets for an organization. Data products are an asset class developed by team collaboration, business subject matter experts and analysts solving significant issues for management. Data products improve data reliability and trustworthiness and present opportunities for cross-functional sharing. The benefits of actively creating and curating data products include -

Capture company IP
Essentially data products capture company intellectual property that can benefit the enterprise by enabling digitization of key processes, scalability and maintenance of unique competitive advantages.

Recurring Data Demand
Designed and developed to solve specific iterative analytical and workflow processes. Data products often create unique assets that efficiently present solutions to decision-makers. Key processes can be improved through test and learn strategies as the data products provide a consistent measure of performance.

Enhance Efficiency
Reduce time spent liaising between departments and teams to repeatedly obtain and integrate data sets.

Ownership and Commitment
Data products allow analysts, data scientists, and data engineers to own the outcome, building a sense of ownership and commitment to their product.

Create Testing and Development Data
Data science, particularly machine learning and artificial intelligence applications, requires defined and high-quality data sets for development and discovery. Machine learning is crucial to many procedures in finance, and well-trained algorithms rely on a consistent data set for growth. Data products include updated data sets and analytics ready for data scientists to test and validate models used throughout an organization.

Leverage Data Science
Data assets are an opportunity to capture the value of data science. Data assets often contain derived data generated by data scientists. The contribution of data science will include improving the overall data quality, using analytics to create new fields of data, and innovating products and services.

Improved Data Security and Compliance
Data products integrating data from many sources will only expose data to users as needed. Data products are useful in managing the risk of exposing data to unauthorized roles.

Case Study

Capital markets have long been a data-intensive industry and reliant on data products to build and maintain competitive advantage. High-quality data products contribute to many aspects of capital markets, from trading decisions to managing a portfolio of listed entities. RoZetta Technology has delivered to capital markets participants a seamless pipeline of data from acquisition to consumption -

Entity Master

Fusing data science and engineering, RoZetta creates an Entity Master, a record of every entity an organization receives or holds data on. This data product can tag entities from unstructured data and retain data on market participants who may not be listed entities.

Security Master

Used to store listed entity information. Information captured and curated includes entity identifiers, reference data, and trading instrument identifiers. A Security Master is essential for integrating information from various data sources to entities traded on an exchange. RoZetta's Security Master builds and maintains multiple entity identifiers to ensure maximum matching with the lowest risk of error.

Time Bar Curated Data

This data product is capital markets analysts' most commonly used source file. It records the market activity for all equities traded on a venue in one-minute intervals.   The field structure can be changed or calculated fields added to customize time bars for clients. Analysts sometimes require variations, like one-minutes or five-minutes time bars, and the data product is easily modified to produce customized time bars.

Business expectations

The outcomes of data curation must be observable and measurable. While there are quite a few measures around data management, and hence curation, that focus on how much data is managed, more meaningful measures should be related to productivity, data quality, trust, and value creation. Some of the performance measures should include -

Decision Velocity

Is there a gap between the need to make a decision and the data being turned into information to make the decision? In moving to a culture of data-driven decisions, timeliness is essential.

As commercial realities change more frequently, competitors are active, and operational decisions become more critical, there cannot be instances where management waits for the data to decide. Data curation plays a crucial role in the process of data to information to decisions in the time windows the organization needs.

Efficiency and effectiveness

End users often become the experts in developing workarounds and data fixes for their tasks. Research reports highlight that fixing data issues or wrangling takes more than half an analyst's time.

Introducing platforms and processes where these fixes are socialized across teams and organizations will significantly impact the efficiency and effectiveness of end users. Improving productivity will lead to lower margins, new product development, and faster decision-making. High-performing data curation creates value. Data curation is not solely a technology responsibility.

Data Quality

Improving the overall data quality can only be achieved if measured. There must be rules around normalization and establishing the quality of both data files, created or licensed, and data vendors. Data vendors need to be accountable for data quality remediation. Data curation has a role in ensuring the continuous improvement of data across an organization.

Users often improve data quality as they wrangle the data into shape for use in their analysis, reporting, or workflow management. Data Science is increasingly involved in identifying and fixing data quality issues, particularly where data comes from an external supplier and every file needs validating before ingestion.

Trusted Data

Aspects like accuracy, completeness, reliability, and timely supply build trust in the data. Trust is about the confidence to use the data for decision-making and analysis.

A key outcome of good data curation is validated by the quality of decisions and analysis by data-dependent groups across the organization.

> **Data curation is essential in managing all data sets used by an organization in running the business and plays a vital role in protecting data assets created by the organization, as these represent a competitive advantage.**
>
> Scott Matthews, Chief Data Scientist at RoZetta Technology

About DataHex Data Library

A proven platform to harness the value of your data assets. Drive your organization's insight and innovation through seamless discovery, access and management of your data product.

Contact Us

To learn more about how DataHex Data Library can unlock your potential for seamless collaboration and drive innovation, visit us at **rozettatechnology.com** or email us at **enquries@rozettatechnology.com**