O'REILLY®

Compliments of tamr

Getting Data Right

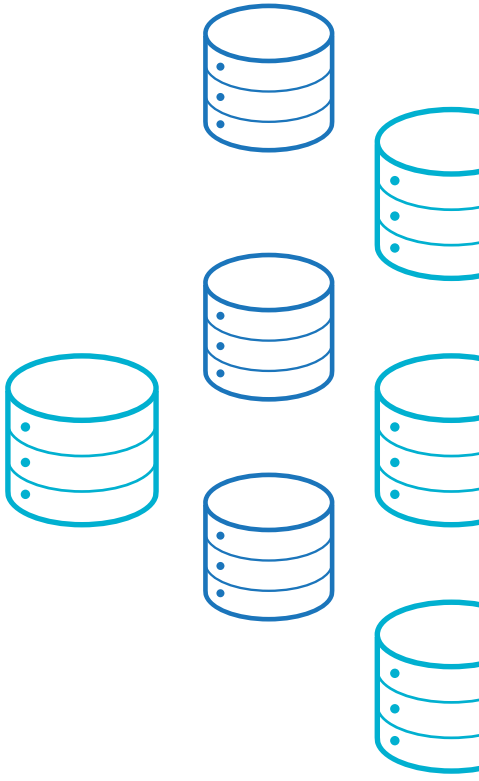
Tackling the Challenges of Big Data Volume and Variety

Edited by Shannon Cutt

tamr

unified data. optimized decisions.

Fuel your decision-making with all available data



In unifying enterprise data for better analytics, Tamr unifies enterprise organizations – bringing business and IT together on mission critical questions and the information needed to answer them.

This Preview Edition of *Getting Data Right*, *Chapter 4*, is a work in progress.
The final book is currently scheduled for release in fall 2015.

Getting Data Right

Tackling The Challenges of Big Data Volume and Variety

Edited by Shannon Cutt

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Getting Data Right
Edited by Shannon Cutt

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com .

Editor: Shannon Cutt **Cover Designer:** Randy Comer

June 2015: First Edition

Revision History for the First Edition
2015-06-17: First Release

978-1-491-93529-3

[LSI]

Table of Contents

1. The Solution: Data Curation at Scale. 5
Three Generations of Data Integration Systems 5
Five Tenets for Success 9

CHAPTER 1

The Solution: Data Curation at Scale

—Michael Stonebraker, Ph.D.

Integrating data sources isn't a new challenge. But the challenge has intensified in both importance and difficulty, as the volume and variety of usable data - and enterprises' ambitious plans for analyzing and applying it - have increased. As a result, trying to meet today's data integration demands with yesterday's data integration approaches is impractical.

In this chapter, we look at the three generations of data integration products and how they have evolved. We look at new third-generation products that deliver a vital missing layer in the data integration "stack": data curation at scale. Finally, we look at five key tenets of an effective data curation at scale system.

Three Generations of Data Integration Systems

Data integration systems emerged to enable business analysts to access converged data sets directly for analyses and applications.

First-generation data integration systems - data warehouses - arrived on the scene in the 1990s. Led by the major retailers, customer-facing data (e.g., item sales, products, customers) were assembled in a data store and used by retail buyers to make better purchase decisions. For example, pet rocks might be out of favor

while Barbie dolls might be "in." With this intelligence, retailers could discount the pet rocks and tie up the Barbie doll factory with a big order. Data warehouses typically paid for themselves within a year through better buying decisions.

First-generation data integration systems were termed **ETL** (Extract, Transform and Load) products. They were used to assemble the data from various sources (usually fewer than 20) into the warehouse. But enterprises underestimated the "T" part of the process - specifically, the cost of data curation (mostly, data cleaning) required to get heterogeneous data into the proper format for querying and analysis. Hence, the typical data warehouse project was usually substantially over-budget and late because of the difficulty of data integration inherent in these early systems.

This led to a second generation of ETL systems, whereby the major ETL products were extended with data cleaning modules, additional adapters to ingest other kinds of data, and data cleaning tools. In effect, the ETL tools were extended to become **data curation** tools.

Data curation involves:

- ingesting data sources
- cleaning errors from the data (-99 often means null)
- transforming attributes into other ones (for example, Euros to dollars)
- performing schema integration to connect disparate data sources
- performing entity consolidation to remove duplicates

In general, data curation systems followed the architecture of earlier first-generation systems: toolkits oriented toward professional programmers. In other words, they were programmer productivity tools.

Second-generation data curation tools have two substantial weaknesses:

Scalability
: Enterprises want to curate "the long tail" of enterprise data. They have several thousand data sources, everything from company budgets in the CFO's spreadsheets to peripheral operational systems. There is "business intelligence gold" in the long

tail, and enterprises wish to capture it - for example, for cross-selling of enterprise products. Furthermore, the rise of public data on the web leads business analysts to want to curate additional data sources. Anything from weather data to customs records to real estate transactions to political campaign contributions are readily available. However, in order to capture long-tail enterprise data as well as public data, curation tools must be able to deal with hundreds to thousands of data sources rather than the typical few tens of data sources.

Architecture
: Second-generation tools typically are designed for central IT departments. A professional programmer does not know the answers to many of the data curation questions that arise. For example, are "rubber gloves" the same thing as "latex hand protectors?" Is an "ICU50" the same kind of object as an "ICU?" Only business people in line-of-business organizations can answer these kinds of questions. However, business people are usually not in the same organization as the programmers running data curation projects. As such, second-generation systems are not architected to take advantage of the humans best able to provide curation help.

These weaknesses led to a third generation of data curation products, which we term **scalable data curation** systems. Any data curation system should be capable of performing the five tasks noted above. However, first- and second-generation ETL products will only scale to a small number of data sources, because of the amount of human intervention required.

To scale to hundreds or even thousands of data sources, a new approach is needed - one that:

1. Uses statistics and machine learning to make automatic decisions wherever possible.
2. Asks a human expert for help only when necessary.

Instead of an architecture with a human controlling the process with computer assistance, move to an architecture with the computer running an automatic process, asking a human for help only when required. And ask the right human: the data creator or owner (a business expert) not the data wrangler (a programmer).

Obviously, enterprises differ in the required accuracy of curation, so third-generation systems must allow an enterprise to make tradeoffs between accuracy and the amount of human involvement. In addition, third-generation systems must contain a crowdsourcing component that makes it efficient for business experts to assist with curation decisions. Unlike Amazon's Mechanical Turk, however, a data-curation crowdsourcing model must be able to accommodate a hierarchy of experts inside an enterprise as well as various kinds of expertise. Therefore, we call this component an **expert sourcing system** to distinguish it from the more primitive crowdsourcing systems.

In short: a third-generation data curation product is an automated system with an expert sourcing component. Tamr is an early example of this third generation of systems.

Third-generation systems can co-exist with currently-in-place second-generation systems, which can curate the first tens of data sources to generate a composite result that in turn can be curated with the "long tail" by third-generation systems.

Table 1-1. Evolution of Three Generations of Data Integration Systems

	First Generation 1990s	Second Generation 2000s	Third Generation 2010s
Approach	ETL	ETL+>Data Curation	Scalable Data Curation
Target Data Environment(s)	Data Warehouse	Data Warehouses or Data Marts	Data Lakes & Self-Service Data Analytics
Users	IT/Programmers	IT/Programmers	Data Scientists, Data Stewards, Data Owners, Business Analysts
Integration Philosophy	Top-down/rules-based/IT-driven	Top-down/rules-based/IT-driven	Bottom-up/demand-based/business-driven
Architecture	Programmer productivity tools (task automation)	Programming productivity tools (task automation with machine assistance)	Machine-driven, human-guided process
Scalability (# of data sources)	10s	10s to 100s	100s to 1000s+

To summarize: ETL systems arose to deal with the transformation challenges in early data warehouses. They evolved into second-generation data curation systems with an expanded scope of offerings. Third-generation data curation systems, which have a very different architecture, were created to address the enterprise's need for data source scalability.

Five Tenets for Success

Third-generation scalable data curation systems provide the architecture, automated workflow, interfaces and APIs for data curation at scale. Beyond this basic foundation, however, are five tenets that are desirable in any third-generation system.

Tenet 1: Data curation is never done

Business analysts and data scientists have an insatiable appetite for more data. This was brought home to me about a decade ago during a visit to a beer company in Milwaukee. They had a fairly standard data warehouse of sales of beer by distributor, time period, brand and so on. I visited during a year when *El Niño* was forecast to disrupt winter weather in the US. Specifically, it was forecast to be wetter than normal on the West Coast and warmer than normal in New England. I asked the business analysts: "Are beer sales correlated with either temperature or precipitation?" They replied, "We don't know, but that is a question we would like to ask." However temperature and precipitation were not in the data warehouse, so asking was not an option.

The demand from warehouse users to correlate more and more data elements for business value leads to additional data curation tasks. Moreover, whenever a company makes an acquisition, it creates a data curation problem (digesting the acquired's data). Lastly, the treasure trove of public data on the web (such as temperature and precipitation data) is largely untapped, leading to more curation challenges.

Even without new data sources, the collection of existing data sources is rarely static. Hence, inserts and deletes to these sources generates a pipeline of incremental updates to a data curation system. Between the requirements of new data sources and updates to existing ones, it is obvious that data curation is never done, ensuring that

any project in this area will effectively continue indefinitely. Realize this and plan accordingly.

One obvious consequence of this tenet concerns consultants. If you hire an outside service to perform data curation for you, then you will have to rehire them for each additional task. This will give the consultant a guided tour through your wallet over time. In my opinion, you are much better off developing in-house curation competence over time.

Tenet 2: A PhD in AI can't be a requirement for success

Any third-generation system will use statistics and machine learning to make automatic or semi-automatic curation decisions. Inevitably, it will use sophisticated techniques such as T-tests, regression, predictive modeling, data clustering, and classification. Many of these techniques will entail training data to set internal parameters. Several will also generate recall and/or precision estimates.

These are all techniques understood by data scientists. However, there will be a shortage of such people for the foreseeable future, until colleges and universities produce substantially more than at present. Also, it is not obvious that one can "retread" a business analyst into a data scientist. A business analyst only needs to understand the output of SQL aggregates; in contrast, a data scientist is typically knowledgeable in statistics and various modeling techniques.

As a result, most enterprises will be lacking in data science expertise. Therefore, any third-generation data curation product must use these techniques internally, but not expose them in the user interface. Mere mortals must be able to use scalable data curation products.

Tenet 3: Fully automatic data curation is not likely to be successful

Some data curation products expect to run fully automatically. In other words, they translate input data sets into output without human intervention. Fully automatic operation is very unlikely to be successful in an enterprise for a variety of reasons. First, there are curation decisions that simply cannot be made automatically. For example, consider two records; one stating that restaurant X is at

location Y while the second states that restaurant Z is at location Y. This could be a case where one restaurant went out of business and got replaced by a second one or it could be a food court. There is no good way to know the answer to this question without human guidance.

Second, there are cases where data curation must have high reliability. Certainly, consolidating medical records should not create errors. In such cases, one wants a human to check all (or maybe just some) of the automatic decisions. Third, there are situations where specialized knowledge is required for data curation. For example, in a genomics application one might have two terms: ICU50 and ICE50. An automatic system might suggest that these are the same thing, since the lexical distance between the terms is low. However, only a human genomics specialist can decide this question.

For these reasons, any third-generation data curation system must be able to ask a human expert - the right human expert - when it is unsure of the answer. Therefore, one must have multiple domains in which a human can be an expert. Within a single domain, humans have a variable amount of expertise, from a novice level to enterprise expert. Lastly, one must avoid overloading the humans that it is scheduling. Therefore, when considering a third generation data curation system, look for an embedded expert system with levels of expertise, load balancing and multiple expert domains.

Tenet 4: Data curation must fit into the enterprise ecosystem

Every enterprise has a computing infrastructure in place. This includes a collection of DBMSs storing enterprise data, a collection of application servers and networking systems, and a set of installed tools and applications. Any new data curation system must fit into this existing infrastructure. For example, it must be able to extract from corporate databases, use legacy data cleaning tools, and export data to legacy data systems. Hence, an open environment is required whereby callouts are available to existing systems. In addition, adapters to common input and export formats is a requirement. Do not use a curation system that is a closed ”black box.”

Tenet 5: A scheme for "finding" data sources must be present

A typical question to ask CIOs is "How many operational data systems do you have?". In all likelihood, they do not know. The enterprise is a sea of such data systems connected by a hodgepodge set of connectors. Moreover, there are all sorts of personal datasets, spreadsheets and databases, as well as datasets imported from public web-oriented sources. Clearly, CIOs should have a mechanism for identifying data resources that they wish to have curated. Such a system must contain a data source catalog with information on a CIO's data resources, as well as a query system for accessing this catalog. Lastly, an "enterprise crawler" is required to search a corporate internet to locate relevant data sources. Collectively, this represents a schema for "finding" enterprise data sources.

Collectively, these five tenets indicate the characteristics of a good third generation data curation system. If you are in the market for such a product, then look for systems with these characteristics.