O'REILLY®

Getting Data Right

Tackling the Challenges of Big Data Volume and Variety

Jerry Held, Michael Stonebraker, Thomas H. Davenport, Ihab Ilyas, Michael L. Brodie, Andy Palmer & James Markarian

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World helps you put big data, cutting-edge data science, and new business fundamentals to work.

- Learn new business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # D2044

Getting Data Right

Tackling the Challenges of Big Data Volume and Variety

Jerry Held, Michael Stonebraker, Thomas H. Davenport, Ihab Ilyas, Michael L. Brodie, Andy Palmer, and James Markarian

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Getting Data Right

by Jerry Held, Michael Stonebraker, Thomas H. Davenport, Ihab Ilyas, Michael L. Brodie, Andy Palmer, and James Markarian

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
Production Editor: Nicholas Adams
Copyeditor: Rachel Head
Proofreader: Nicholas Adams

Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

September 2016: First Edition

Revision History for the First Edition

2016-09-06: First Release

While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-93553-8

[LSI]

Table of Contents

Introduction

Jerry Held

Companies have invested an estimated $3–4 trillion in IT over the last 20-plus years, most of it directed at developing and deploying single-vendor applications to automate and optimize key business processes. And what has been the result of all of this disparate activity? Data silos, schema proliferation, and radical data heterogeneity.

With companies now investing heavily in big data analytics, this entropy is making the job considerably more complex. This complexity is best seen when companies attempt to ask "simple" questions of data that is spread across many business silos (divisions, geographies, or functions). Questions as simple as "Are we getting the best price for everything we buy?" often go unanswered because on their own, top-down, deterministic data unification approaches aren't prepared to scale to the variety of hundreds, thousands, or tens of thousands of data silos.

The diversity and mutability of enterprise data and semantics should lead CDOs to explore—as a complement to deterministic systems—a new bottom-up, probabilistic approach that connects data across the organization and exploits big data variety. In managing data, we should look for solutions that find siloed data and connect it into a unified view. "Getting Data Right" means embracing variety and transforming it from a roadblock into ROI. Throughout this report, you'll learn how to question conventional assumptions, and explore alternative approaches to managing big data in the enterprise. Here's a summary of the topics we'll cover:

Chapter 1, The Solution: Data Curation at Scale

Michael Stonebraker, 2015 A.M. Turing Award winner, argues that it's impractical to try to meet today's data integration demands with yesterday's data integration approaches. Dr. Stonebraker reviews three generations of data integration products, and how they have evolved. He explores new third-generation products that deliver a vital missing layer in the data integration "stack"—data curation at scale. Dr. Stonebraker also highlights five key tenets of a system that can effectively handle data curation at scale.

Chapter 2, An Alternative Approach to Data Management

In this chapter, Tom Davenport, author of *Competing on Analytics* and *Big Data at Work* (Harvard Business Review Press), proposes an alternative approach to data management. Many of the centralized planning and architectural initiatives created throughout the 60 years or so that organizations have been managing data in electronic form were never completed or fully implemented because of their complexity. Davenport describes five approaches to realistic, effective data management in today's enterprise.

Chapter 3, Pragmatic Challenges in Building Data Cleaning Systems

Ihab Ilyas of the University of Waterloo points to "dirty, inconsistent data" (now the norm in today's enterprise) as the reason we need new solutions for quality data analytics and retrieval on large-scale databases. Dr. Ilyas approaches this issue as a theoretical and engineering problem, and breaks it down into several pragmatic challenges. He explores a series of principles that will help enterprises develop and deploy data cleaning solutions at scale.

Chapter 4, Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery

Michael Brodie, research scientist at MIT's Computer Science and Artificial Intelligence Laboratory, is devoted to understanding data science as an emerging discipline for data-intensive analytics. He explores data science as a basis for the Fourth Paradigm of engineering and scientific discovery. Given the potential risks and rewards of data-intensive analysis and its breadth of application, Dr. Brodie argues that it's imperative we get it right. In this chapter, he summarizes his analysis of more than 30 large-scale use cases of data science, and reveals a body of principles and techniques with which to measure and improve the correctness, completeness, and efficiency of data-intensive analysis.

Chapter 5, From DevOps to DataOps
Tamr Cofounder and CEO Andy Palmer argues in support of "DataOps" as a new discipline, echoing the emergence of "DevOps," which has improved the velocity, quality, predictability, and scale of software engineering and deployment. Palmer defines and explains DataOps, and offers specific recommendations for integrating it into today's enterprises.

Chapter 6, Data Unification Brings Out the Best in Installed Data Management Strategies
Former Informatica CTO James Markarian looks at current data management techniques such as extract, transform, and load (ETL); master data management (MDM); and data lakes. While these technologies can provide a unique and significant handle on data, Markarian argues that they are still challenged in terms of speed and scalability. Markarian explores adding data unification as a frontend strategy to quicken the feed of highly organized data. He also reviews how data unification works with installed data management solutions, allowing businesses to embrace data volume and variety for more productive data analysis.

CHAPTER 1

The Solution: Data Curation at Scale

Michael Stonebraker, PhD

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, focusing on the 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 datasets directly for analyses and applications.

First-generation data integration systems—data warehouses—arrived on the scene in the 1990s. Major retailers took the lead, assembling, customer-facing data (e.g., item sales, products, customers) in data stores and mining it to make better purchasing 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 the 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, wherein 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 five key tasks:

1. Ingesting data sources
2. Cleaning errors from the data (–99 often means null)
3. Transforming attributes into other ones (for example, euros to dollars)
4. Performing schema integration to connect disparate data sources
5. Performing entity consolidation to remove duplicates

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

While many of these are still in use today, 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 is leading business analysts to want to curate additional data sources. Data on everything from the weather to customs records to real estate transactions to political campaign contributions is 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 tens of data sources most second-generation tools are equipped to handle.

Architecture

Second-generation tools typically are designed for central IT departments. A professional programmer will 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 businesspeople in line-of-business organizations can answer these kinds of questions. However, businesspeople are usually not in the same organizations 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 earlier. 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, we must move to an architecture with the computer running an automatic process, asking a human for help only when required. It's also important that this process 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 coexist with second-generation systems that are currently in place, 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 the third-generation systems. Table 1-1 illustrates the key characteristics of the three types of curation 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 warehouses	Data warehouses or Data marts	Data lakes and self-service data analytics
Users	IT/programmers	IT/programmers	Data scientists, data stewards, data owners, business analysts

	First generation	Second generation	Third generation
	1990s	2000s	2010s
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 data 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 company'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. Insertions and deletions in these sources generate 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 consultants 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 semiautomatic 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 begin producing 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 familiar with 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 the location could be a food court. There is no good way to know which record is correct 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 make this determination.

For all of these reasons, any third-generation data curation system must be able to ask the right human expert for input when it is unsure of the answer. The system must also avoid overloading the experts that are involved.

Tenet 4: Data Curation Must Fit into the Enterprise Ecosystem

Every enterprise has a computing infrastructure in place. This includes a collection of database management systems 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 data from corporate databases, use legacy data cleaning tools, and export data to legacy data systems. Hence, an open environment is required wherein callouts are available to existing systems. In addition, adapters to common input and export formats are 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, linked 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 intranet to locate relevant data sources. Collectively, this represents a schema for “finding” enterprise data sources.

Taken together, 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 features.

CHAPTER 2

An Alternative Approach to Data Management

Thomas H. Davenport

For much of the 60 years or so that organizations have been managing data in electronic form, there has been an overpowering desire to subdue it through centralized planning and architectural initiatives.

These initiatives have had a variety of names over the years, including the most familiar: "information architecture," "information engineering," and "master data management." Underpinning them has been a set of key attributes and beliefs:

- Data needs to be centrally controlled.
- Modeling is an approach to controlling data.
- Abstraction is a key to successful modeling.
- An organization's information should all be defined in a common fashion.
- Priority is on efficiency in information storage (a given data element should only be stored once).
- Politics, ego, and other common human behaviors are irrelevant to data management (or at least not something that organizations should attempt to manage).

Each of these statements has at least a grain of truth in it, but taken together and to their full extent, I have come to believe that they simply don't work as the foundation for data management. I rarely find business users who believe they work either, and this dissatisfaction has been brewing for a long time. For example, in the 1990s I interviewed a marketing manager at Xerox Corporation who had also spent some time in IT at the same company. He explained that the company had "tried information architecture" for 25 years, but got nowhere—they always thought they were doing it incorrectly.

Centralized Planning Approaches

Most organizations have had similar results from their centralized architecture and planning approaches.

Not only do centralized planning approaches waste time and money, but they also drive a wedge between those who are planning them and those who will actually use the information and technology. Regulatory submissions, abstract meetings, and incongruous goals can lead to months of frustration, without results.

The complexity and detail of centralized planning approaches often mean that they are never completed, and when they are finished, managers frequently decide not to implement them. The resources devoted to central data planning are often redeployed into other IT projects of more tangible value. If by chance they are implemented, they are typically hopelessly out of date by the time they go into effect.

As an illustration of how the key tenets of centralized information planning are not consistent with *real organizational behavior*, let's look at one: the assumption that all information needs to be common.

Common Information

Common information—agreement within an organization on how to define and use key data elements—is a useful thing, to be sure. But it's also helpful to know that uncommon information—information definitions that suit the purposes of a particular group or individual—can also be useful to a particular business function, unit, or work group. Companies need to strike a balance between these two desirable goals.

After speaking with many managers and professionals about common information, and reflecting on the subject carefully, I formulated "Davenport's Law of Common Information" (you can Google it, but don't expect a lot of results). If by some strange chance you haven't heard of Davenport's Law, it goes like this:

> The more an organization knows or cares about a particular business entity, the less likely it is to agree on a common term and meaning for it.

I first noticed this paradoxical observation at American Airlines more than a decade ago. Company representatives told me during a research visit that they had 11 different usages of the term "airport." As a frequent traveler on American Airlines planes, I was initially a bit concerned about this, but when they explained it, the proliferation of meanings made sense. They said that the cargo workers at American Airlines viewed anyplace you can pick up or drop off cargo as the airport; the maintenance people viewed anyplace you can fix an airplane as the airport; the people who worked with the International Air Transport Authority relied on their list of international airports, and so on.

Information Chaos

So, just like Newton being hit on the head with an apple and discovering gravity, the key elements of Davenport's Law hit me like a brick. This was why organizations were having so many problems creating consensus around key information elements. I also formulated a few corollaries to the law, such as:

> If you're not arguing about what constitutes a "customer," your organization is probably not very passionate about customers.

Davenport's Law, in my humble opinion, makes it much easier to understand why companies all over the world have difficulty establishing common definitions of key terms within their organizations.

Of course, this should not be an excuse for organizations to allow alternative meanings of key terms to proliferate. Even though there is a good reason why they proliferate, organizations may have to limit—or sometimes even stop—the proliferation of meanings and agree on one meaning for each term. Otherwise they will continue to find that when the CEO asks multiple people how many employees a company has, he/she will get different answers. The proliferation of meanings, however justifiable, leads to information chaos.

But Davenport's Law offers one more useful corollary about how to stop the proliferation of meanings. Here it is:

> A manager's passion for a particular definition of a term will not be quenched by a data model specifying an alternative definition.

If a manager has a valid reason to prefer a particular meaning of a term, he/she is unlikely to be persuaded to abandon it by a complex, abstract data model that is difficult to understand in the first place, and is likely never to be implemented.

Is there a better way to get adherence to a single definition of a term?

Here's one final corollary:

> Consensus on the meaning of a term throughout an organization is achieved not by data architecture, but by data *arguing*.

Data modeling doesn't often lead to *dialog*, because it's simply not comprehensible to most nontechnical people. If people don't understand your data architecture, it won't stop the proliferation of meanings.

What Is to Be Done?

There is little doubt that something needs to be done to make data integration and management easier. In my research, I've conducted more than 25 extended interviews with data scientists about what they do, and how they go about their jobs. I concluded that a more appropriate title for data scientists might actually be "data plumbers." It is often so difficult to extract, clean, and integrate data that data scientists can spend up to 90% of their working time doing those tasks. It's no wonder that big data often involves "small math" —after all the preparation work, there isn't enough time left to do sophisticated analytics.

This is not a new problem in data analysis. The dirty little secret of the field is that someone has always had to do a lot of data preparation before the data can be analyzed. The problem with big data is partly that there is a large volume of it, but mostly that we are often trying to integrate multiple sources. Combining multiple data sources means that for each source, we have to determine how to clean, format, and integrate its data. The more sources and types of data there are, the more plumbing work is required.

So let's assume that data integration and management are necessary evils. But what particular approaches to them are most effective? Throughout the remainder of this chapter, I'll describe five approaches to realistic, effective data management:

1. Take a federal approach to data management.
2. Use all the new tools at your disposal.
3. Don't model, catalog.
4. Keep everything simple and straightforward.
5. Use an ecological approach.

Take a Federal Approach to Data Management

Federal political models—of which the United States is one example —don't try to get consensus on every issue. They have some laws that are common throughout the country, and some that are allowed to vary from state to state or by region or city. It's a hybrid approach to the centralization/decentralization issue that bedevils many large organizations. Its strength is its practicality, in that it's easier to get consensus on some issues than on all of them. If there is a downside to federalism, it's that there is usually a lot of debate and discussion about which rights are federal, and which are states' or other units' rights. The United States has been arguing about this issue for more than 200 years.

While federalism does have some inefficiencies, it's a good model for data management. It means that some data should be defined commonly across the entire organization, and some should be allowed to vary. Some should have a lot of protections, and some should be relatively open. That will reduce the overall effort required to manage data, simply because not everything will have to be tightly managed.

Your organization will, however, have to engage in some "data arguing." Hashing things out around a table is the best way to resolve key issues in a federal data approach. You will have to argue about which data should be governed by corporate rights, and which will be allowed to vary. Once you have identified corporate data, you'll then have to argue about how to deal with it. But I have found that if managers feel that their issues have been fairly aired, they are more likely to comply with a policy that goes against those issues.

Use All the New Tools at Your Disposal

We now have a lot of powerful tools for processing and analyzing data, but up to now we haven't had them for cleaning, integrating, and "curating" data. ("Curating" is a term often used by librarians, and there are typically many of them in pharmaceutical firms who manage scientific literature.) These tools are sorely needed and are beginning to emerge. One source I'm close to is a startup called Tamr, which aims to help "tame" your data using a combination of machine learning and crowdsourcing. Tamr isn't the only new tool for this set of activities, though, and I am an advisor to the company, so I would advise you to do your own investigation. The founders of Tamr (both of whom have also contributed to this report) are Andy Palmer and Michael Stonebraker. Palmer is a serial entrepreneur and incubator founder in the Boston area.

Stonebraker is the database architect behind INGRES, Vertica, VoltDB, Paradigm4, and a number of other database tools. He's also a longtime computer science professor, now at MIT. As noted in his chapter of this report, we have a common view of how well-centralized information architecture approaches work in large organizations.

In a research paper published in 2013, Stonebraker and several co-authors wrote that they had tested "Data-Tamer" (as it was then known) in three separate organizations. They found that the tool reduced the cost of data curation in those organizations by about 90%.

I like the idea that Tamr uses two separate approaches to solving the problem. If the data problem is somewhat repetitive and predictable, the machine learning approach will develop an algorithm that will do the necessary curation. If the problem is a bit more ambiguous, the crowdsourcing approach can ask people who are familiar with the data (typically the owners of that data source) to weigh in on its quality and other attributes. Obviously the machine learning approach is more efficient, but crowdsourcing at least spreads the labor around to the people who are best qualified to do it. These two approaches are, together, more successful than the top-down approaches that many large organizations have employed.

A few months before writing this chapter, I spoke with several managers from companies who are working with Tamr. Thomson Reuters is using the technology to curate "core entity master" data—creating clear and unique identities of companies and their parents and subsidiaries. Previous in-house curation efforts, relying on a handful of data analysts, found that 30–60% of entities required manual review. Thomson Reuters believed manual integration would take up to six months to complete, and would identify 95% of duplicate matches (precision) and 95% of suggested matches that were, in fact, different (recall).

Thomson Reuters looked to Tamr's machine-driven, human-guided approach to improve this process. After converting the company's XML files to CSVs, Tamr ingested three core data sources—factual data on millions of organizations, with more than 5.4 million records. Tamr deduplicated the records and used "fuzzy matching" to find suggested matches, with the goal of achieving high accuracy rates while reducing the number of records requiring review. In order to scale the effort and improve accuracy, Tamr applied machine learning algorithms to a small training set of data and fed guidance from Thomson Reuters' experts back into the system.

The "big pharma" company Novartis is also using Tamr. Novartis has many different sources of biomedical data that it employs in research processes, making curation difficult. Mark Schreiber, then an "informatician" at Novartis Institutes for Biomedical Research (he has since moved to Merck), oversaw the testing of Tamr going all the way back to its academic roots at MIT. He is particularly interested in the tool's crowdsourcing capabilities, as he wrote in a blog post:

> The approach used gives you a critical piece of the workflow bridging the gap between the machine learning/automated data improvement and the curator. When the curator isn't confident in the prediction or their own expertise, they can distribute tasks to your data producers and consumers to ask their opinions and draw on their expertise and institutional memory, which is not stored in any of your data systems.

I also spoke with Tim Kasbe, the COO of Gloria Jeans, which is the largest "fast fashion" retailer in Russia and Ukraine. Gloria Jeans has tried out Tamr on several different data problems, and found it particularly useful for identifying and removing duplicate loyalty program records. Here are some results from that project:

> We loaded data for about 100,000 people and families and ran our algorithms on them and found about 5,000 duplicated entries. A

> portion of these represented people or families that had signed up for multiple discount cards. In some cases, the discount cards had been acquired in different locations or different contact information had been used to acquire them. The whole process took about an hour and did not need deep technical staff due to the simple and elegant Tamr user experience. Getting to trustworthy data to make good and timely decisions is a huge challenge this tool will solve for us, which we have now unleashed on all our customer reference data, both inside and outside the four walls of our company.

I am encouraged by these reports that we are on the verge of a breakthrough in this domain. But don't take my word for it—do a proof of concept with one of these types of tools.

Don't Model, Catalog

One of the paradoxes of IT planning and architecture is that those activities have made it more difficult for people to find the data they need to do their work. According to Gartner, much of the roughly $3–4 trillion invested in enterprise software over the last 20 years has gone toward building and deploying software systems and applications to automate and optimize key business processes in the context of specific functions (sales, marketing, manufacturing) and/or geographies (countries, regions, states, etc.). As each of these idiosyncratic applications is deployed, an equally idiosyncratic data source is created. The result is that data is extremely heterogeneous and siloed within organizations.

For generations, companies have created "data models," "master data models," and "data architectures" that lay out the types, locations, and relationships of all the data that they have now and will have in the future. Of course, those models rarely get implemented exactly as planned, given the time and cost involved. As a result, organizations have no guide to what data they actually have in the present and how to find it. Instead of creating a data model, they should create a *catalog* of their data—a straightforward listing of what data exists in the organization, where it resides, who's responsible for it, and so forth.

One reason why companies don't create simple catalogs of their data is that the result is often somewhat embarrassing and irrational. Data is often duplicated many times across the organization. Different data is referred to by the same term, and the same data by different terms. A lot of data that the organization no longer needs is still hanging around, and data that the organization could really benefit from is nowhere to be found. It's not easy to face up to all of the informational chaos that a cataloging effort can reveal.

Perhaps needless to say, however, cataloging data is worth the trouble and initial shock at the outcome. A data catalog that lists what data the organization has, what it's called, where it's stored, who's responsible for it, and other key metadata can easily be the most valuable information offering that an IT group can create.

Cataloging Tools

Given that IT organizations have been more preoccupied with modeling the future than describing the present, enterprise vendors haven't really addressed the catalog tool space to a significant degree. There are several catalog tools for individuals and small businesses, and several vendors of ETL (extract, transform, and load) tools have some cataloging capabilities built into their own tools. Some also tie a catalog to a data governance process, although "governance" is right up there with "bureaucracy" as a term that makes many people wince.

At least a few data providers and vendors are actively pursuing catalog work, however. One company, Enigma, has created a catalog for public data, for example. The company has compiled a set of public databases, and you can simply browse through its catalog (for free if you are an individual) and check out what data you can access and analyze. That's a great model for what private enterprises should be developing, and I know of some companies (including Tamr, Informatica, Paxata, and Trifacta) that are developing tools to help companies develop their own catalogs.

In industries such as biotech and financial services, for example, you increasingly need to know what data you have—and not only so you can respond to business opportunities. Industry regulators are also concerned about what data you have and what you are doing with it. In biotech companies, for example, any data involving patients has to be closely monitored and its usage controlled, and in financial services firms there is increasing pressure to keep track of customers' and partners' "legal entity identifiers," and to ensure that dirty money isn't being laundered.

If you don't have any idea of what data you have today, you're going to have a much tougher time adhering to the demands from regulators. You also won't be able to meet the demands of your marketing, sales, operations, or HR departments. Knowing where your data is seems perhaps the most obvious tenet of information management, but thus far, it has been among the most elusive.

Keep Everything Simple and Straightforward

While data management is a complex subject, traditional information architectures are generally more complex than they need to be. They are usually incomprehensible not only to nontechnical people, but also to the technical people who didn't have a hand in creating them. From IBM's Business Systems Planning—one of the earliest architectural approaches—up through master data management (MDM), architectures feature complex and voluminous flow diagrams and matrices. Some look like the circuitry diagrams for the latest Intel microprocessors. MDM has the reasonable objective of ensuring that all important data within an organization comes from a single authoritative source, but it often gets bogged down in discussions about who's in charge of data and whose data is most authoritative.

It's unfortunate that information architects don't emulate architects of physical buildings. While they definitely require complex diagrams full of technical details, good building architects don't show those blueprints to their clients. For clients, they create simple and easy-to-digest sketches of what the building will look like when it's done. If it's an expensive or extensive building project, they may create three-dimensional models of the finished structure.

More than 30 years ago, Michael Hammer and I created a new approach to architecture based primarily on "principles." These are simple, straightforward articulations of what an organization believes and wants to achieve with information management; the equivalent of a sketch for a physical architect. Here are some examples of the data-oriented principles from that project:

- Data will be owned by its originator but will be accessible to higher levels.
- Critical data items in customer and sales files will conform to standards for name, form, and semantics.
- Applications should be processed where data resides.

We suggested that an organization's entire list of principles—including those for technology infrastructure, organization, and applications, as well as data management—should take up no more than a single page. Good principles can be the drivers of far more detailed plans, but they should be articulated at a level that facilitates understanding and discussion by businesspeople. In this age of digital businesses, such simplicity and executive engagement is far more critical than it was in 1984.

Use an Ecological Approach

I hope I have persuaded you that enterprise-level models (or really models at any level) are not sufficient to change individual and organizational behavior, with respect to data. But now I will go even further and argue that neither models nor technology, policy, or any other single factor is enough to move behavior in the right direction. Instead, organizations need a broad, ecological approach to data-oriented behaviors.

In 1997 I wrote a book called *Information Ecology: Mastering the Information and Knowledge Environment* (Oxford University Press). It was focused on this same idea—that multiple factors and interventions are necessary to move an organization in a particular direction with regard to data and technology management. Unlike engineering-based models, ecological approaches assume that technology alone is not enough to bring about the desired change, and that with multiple interventions an environment can evolve in the right direction. In the book, I describe one organization, a large UK insurance firm called Standard Life, that adopted the ecological approach and made substantial progress on managing its customer and policy data. Of course, no one—including Standard Life—ever achieves perfection in data management; all one can hope for is progress.

In *Information Ecology*, I discussed the influence on a company's data environment of a variety of factors, including staff, politics, strategy, technology, behavior and culture, process, architecture, and the external information environment. I'll explain the lesser-known aspects of this model briefly.

Staff, of course, refers to the types of people and skills that are present to help manage information. *Politics* refers primarily to the type of political model for information that the organization employs; as noted earlier, I prefer federalism for most large companies. *Strategy* is the company's focus on particular types of information and particular objectives for it. *Behavior and culture* refers to the particular information behaviors (e.g., not creating new data sources and reusing existing ones) that the organization is trying to elicit; in the aggregate they constitute "information culture." *Process* involves the specific steps that an organization undertakes to create, analyze, disseminate, store, and dispose of information. Finally, the *external information environment* consists of information sources and uses an outside of organization's boundaries that the organization may use to improve its information situation. Most organizations have architectures and technology in place for data management, but they have few, if any, of these other types of interventions.

I am not sure that these are now (or ever were) the only types of interventions that matter, and in any case the salient factors will vary across organizations. But I am quite confident that an approach that employs multiple factors to achieve an objective (for example, to achieve greater use of common information) is more likely to succeed than one focused only on technology or architectural models.

Together, the approaches I've discussed in this chapter comprise a common-sense philosophy of data management that is quite different from what most organizations have employed. If for no other reason, organizations should try something new because so many have yet to achieve their desired state of data management.

CHAPTER 3

Pragmatic Challenges in Building Data Cleaning Systems

Ihab Ilyas

Acquiring and collecting data often introduces errors, including missing values, typos, mixed formats, replicated entries of the same real-world entity, and even violations of business rules. As a result, "dirty data" has become the norm, rather than the exception, and most solutions that deal with real-world enterprise data suffer from related pragmatic problems that hinder deployment in practical industry and business settings.

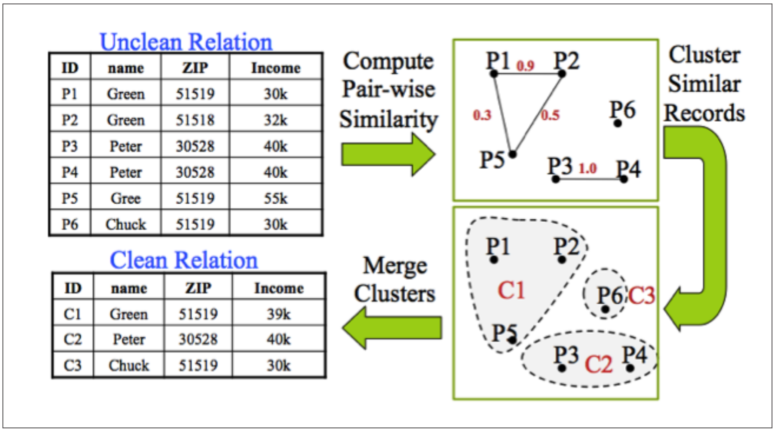
In the field of big data, we need new technologies that provide solutions for quality data analytics and retrieval on large-scale databases that contain inconsistent and dirty data. Not surprisingly, developing pragmatic data quality solutions is a challenging task, rich with deep theoretical and engineering problems. In this chapter, we discuss several of the pragmatic challenges caused by dirty data, and a series of principles that will help you develop and deploy data cleaning solutions.

Data Cleaning Challenges

In the process of building data cleaning software, there are many challenges to consider. In this section, we'll explore seven characteristics of real-world applications, and the often-overlooked challenges they pose to the data cleaning process.

1. Scale

One of the building blocks in data quality is record linkage and consistency checking. For example, detecting functional dependency violations involves (at least) quadratic complexity algorithms, such as those that enumerate all pairs of records to assess if there is a violation (e.g., Figure 3-1 illustrates the process of determining that if two employee records agree on the zip code, they have to be in the same city). In addition, more expensive activities, such as clustering and finding the minimum vertex, work to consolidate duplicate records or to accumulate evidence of data errors. Given the complexity of these activities, cleaning large-scale data sets is prohibitively expensive, both computationally and in terms of cost. (In fact, scale renders most academic proposals inapplicable to real-world settings.) Large-scale blocking and hashing techniques are often used to trade off the *complexity* and *recall* of detected anomalies, and *sampling* is heavily used in both assessing the quality of the data and producing clean data samples for analytics.



Unclean Relation

ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k

Figure 3-1. Expensive operations in record deduplication

2. Human in the Loop

Data is not born an orphan, and enterprise data is often treated as an asset guarded by "data owners" and "custodians." Automatic changes are usually based on *heuristic objectives*, such as introducing minimal changes to the data, or trusting a specific data source over others. Unfortunately, these objectives cannot lead to viable deployable solutions, since oftentimes human-verified or trusted updates are necessary to actually change the underlying data.

A major challenge in developing an enterprise-adoptable solution is allowing only *trusted* fixes to data errors, where "trusted" refers to expert interventions or verification by master data or knowledge bases. The high cost involved in engaging data experts and the heterogeneity and limited coverage of reference master data make trusted fixes a challenging task. We need to *judiciously* involve experts and knowledge bases (reference sources) to repair erroneous data sets.

Effective user engagement in data curation will necessarily involve different roles of humans in the data curation loop: data scientists are usually aware of the final questions that need to be answered from the input data, and what tools will be used to analyze it; business owners are the best to articulate the value of the analytics, and hence control the cost/accuracy trade-off; while domain experts are uniquely qualified to answer data-centric questions, such as whether or not two instances of a product are the same (Figure 3-2).

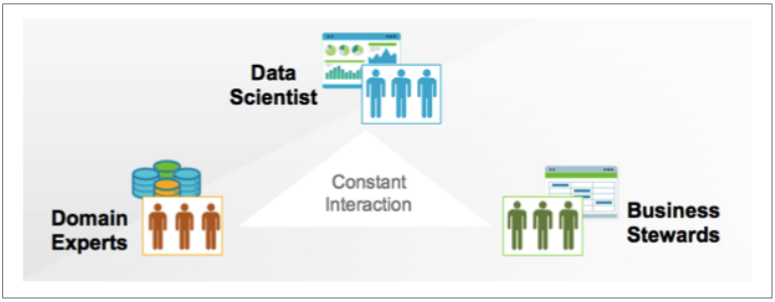
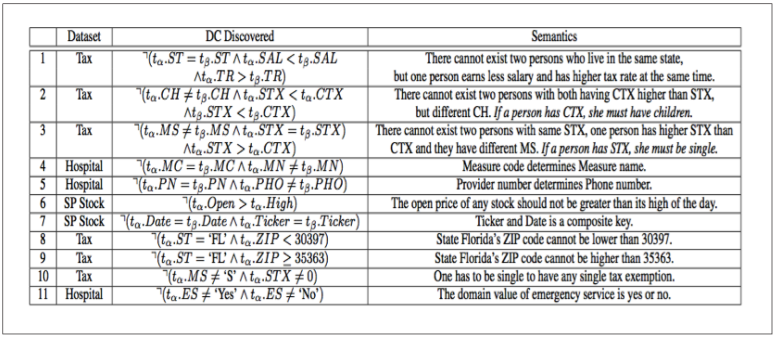


Figure 3-2. Humans in the loop

What makes things even more interesting is that enterprise data is often protected by layers of access control and policies to guide who can see what. Solutions that involve humans or experts have to adhere to these access control policies during the cleaning process. While that would be straightforward if these policies were explicitly and succinctly represented to allow porting to the data curation stack, the reality is that most of these access controls are embedded and hardwired in various applications and data access points. To develop a viable and effective human-in-the-loop solution, full awareness of these access constraints is a must.

3. Expressing and Discovering Quality Constraints

While data repairing is well studied for closed-form integrity constraints formulae (such as functional dependency or denial constraints), real-world business rules are rarely expressed in these rather limited languages. Quality engineers often require running scripts written in imperative languages to encode the various business rules (Figure 3-3). Having an extensible cleaning platform that allows for expressing rules in these powerful languages, yet limiting the interface to rules that are interpretable and practical to enforce, is a hard challenge. What is even more challenging is discovering these high-level business rules from the data itself (and ultimately verifying them via domain experts). Automatic business and quality constraints discovery and enforcement can play a key role in continually monitoring the health of the source data and pushing data cleaning activities upstream, closer to data generation and acquisition.

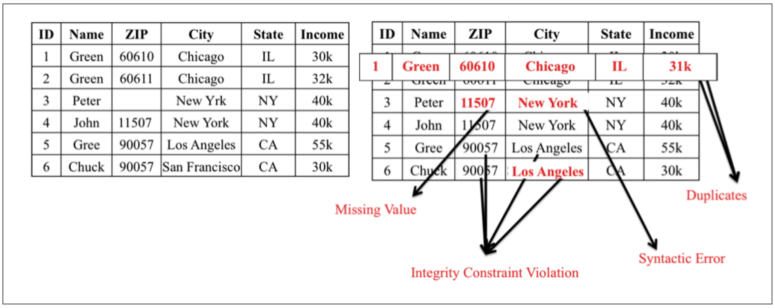


	Dataset	DC Discovered	Semantics
1	Tax	$\neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.SAL < t_\beta.SAL$ $\wedge t_\alpha.TR > t_\beta.TR)$	There cannot exist two persons who live in the same state, but one person earns less salary and has higher tax rate at the same time.
2	Tax	$\neg(t_\alpha.CH \neq t_\beta.CH \wedge t_\alpha.STX < t_\alpha.CTX$ $\wedge t_\beta.STX < t_\beta.CTX)$	There cannot exist two persons with both having CTX higher than STX, but different CH. *If a person has CTX, she must have children.*
3	Tax	$\neg(t_\alpha.MS \neq t_\beta.MS \wedge t_\alpha.STX = t_\beta.STX)$ $\wedge t_\alpha.STX > t_\alpha.CTX)$	There cannot exist two persons with same STX, one person has higher STX than CTX and they have different MS. *If a person has STX, she must be single.*
4	Hospital	$\neg(t_\alpha.MC = t_\beta.MC \wedge t_\alpha.MN \neq t_\beta.MN)$	Measure code determines Measure name.
5	Hospital	$\neg(t_\alpha.PN = t_\beta.PN \wedge t_\alpha.PHO \neq t_\beta.PHO)$	Provider number determines Phone number.
6	SP Stock	$\neg(t_\alpha.Open > t_\alpha.High)$	The open price of any stock should not be greater than its high of the day.
7	SP Stock	$\neg(t_\alpha.Date = t_\beta.Date \wedge t_\alpha.Ticker = t_\beta.Ticker)$	Ticker and Date is a composite key.
8	Tax	$\neg(t_\alpha.ST = \text{'FL'} \wedge t_\alpha.ZIP < 30397)$	State Florida's ZIP code cannot be lower than 30397.
9	Tax	$\neg(t_\alpha.ST = \text{'FL'} \wedge t_\alpha.ZIP \geq 35363)$	State Florida's ZIP code cannot be higher than 35363.
10	Tax	$\neg(t_\alpha.MS \neq \text{'S'} \wedge t_\alpha.STX \neq 0)$	One has to be single to have any single tax exemption.
11	Hospital	$\neg(t_\alpha.ES \neq \text{'Yes'} \wedge t_\alpha.ES \neq \text{'No'})$	The domain value of emergency service is yes or no.

Figure 3-3. Sample business rules expressed as denial constraints

4. Heterogeneity and Interaction of Quality Rules

Data anomalies are rarely due to one type of error; dirty data often includes a collection of duplicates, business rules violations, missing values, misaligned attributes, and unnormalized values. Most available solutions focus on one type of error to allow for sound theoretical results, or for a practical scalable solution. These solutions cannot be applied independently because they usually conflict on the same data. We have to develop "holistic" cleaning solutions that compile heterogeneous constraints on the data, and identify the most problematic data portions by accumulating "evidence of errors" (Figure 3-4).



ID	Name	ZIP	City	State	Income
1	Green	60610	Chicago	IL	30k
2	Green	60611	Chicago	IL	32k
3	Peter		New Yrk	NY	40k
4	John	11507	New York	NY	40k
5	Gree	90057	Los Angeles	CA	55k
6	Chuck	90057	San Francisco	CA	30k

ID	Name	ZIP	City	State	Income
1	Green	60610	Chicago	IL	31k
2	Green	60611	Chicago	IL	32k
3	Peter	11507	New York	NY	40k
4	John	11507	New York	NY	40k
5	Gree	90057	Los Angeles	CA	55k
6	Chuck	90057	Los Angeles	CA	30k

Figure 3-4. Data cleaning is holistic

5. Data and Constraints Decoupling and Interplay

Data and integrity constraints often interplay and are usually decoupled in space and time, in three different ways. First, while errors are born with the data, they are often discovered much later in applications, where more business semantics are available; hence, constraints are often declared and applied much later, and in multiple stages in the data processing life cycle. Second, detecting and fixing errors at the source, rather than at the application level, is important in order to avoid updatability restrictions and to prevent future errors. Finally, data cleaning rules themselves are often inaccurate; hence, a cleaning solution has to consider "relaxing" the rules to avoid overfitting and to respond to business logic evolution. Cleaning solutions need to build on causality and responsibility results, in order to reason about the errors in data sources. This allows for identifying the most problematic data, and logically summarizing data anomalies using predicates on the data schema and accompanying provenance information.

6. Data Variety

Considering only *structured* data limits the complexity of detecting and repairing data errors. Most current solutions are designed to work with one type of structured data—tables—yet businesses and modern applications process a large variety of data sources, most of which are *unstructured*. Oftentimes, businesses will extract the important information and store it in structured data warehouse tables. Delaying the quality assessment until *after* this information is extracted and loaded into data warehouses becomes inefficient and inadequate. More effective solutions are likely to push data quality constraints to the information extraction subsystem to limit the amount of dirty data pumped into the business intelligence stack and to get closer to the sources of errors, where more context is available for trusted and high-fidelity fixes (Figure 3-5).

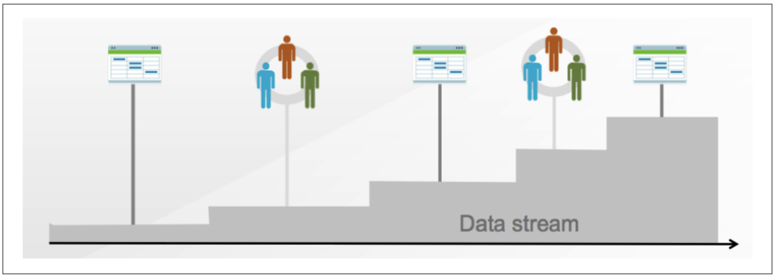


Figure 3-5. Iterative by design

7. Iterative by Nature, Not Design

While most cleaning solutions insist on "one-shot cleaning," data typically arrives and is handled incrementally, and quality rules and schema are continuously evolving. One-shot cleaning solutions cannot sustain large-scale data in a continuously changing enterprise environment, and are destined to be abandoned. The cleaning process is iterative by nature, and has to have incremental algorithms at its heart. This usually entails heavy collection and maintenance of data *provenance* (e.g., metadata that describes the sources and the types of changes the data is going through), in order to keep track of data "states." Keeping track of data states allows algorithms and human experts to add knowledge, to change previous beliefs, and even to roll back previous actions.

Building Adoptable Data Cleaning Solutions

With hundreds of research papers on the topic, data cleaning efforts in *industry* are still pretty much limited to one-off solutions that are a mix of consulting work, rule-based systems, and ETL scripts. The data cleaning challenges we've reviewed in this chapter present real obstacles in building cleaning platforms. Tackling all of these challenges in one platform is likely to be a very expensive software engineering exercise. On the other hand, ignoring them is likely to produce throwaway system prototypes.

Adoptable data cleaning solutions can tackle at least a few of these pragmatic problems by:

1. Having humans or experts in the loop as a first-class cleaning process for training models and verification
2. Focusing on scale from the start, and not as an afterthought (which will exclude most naïve brute-force techniques currently used in problems like deduplication and schema mapping)
3. Realizing that curation is a continuous incremental process that requires a mix of incremental algorithms and a full-fledged provenance management system in the backend, to allow for controlling and revising decisions long into the curation life cycle
4. Coupling data cleaning activities to data consumption endpoints (e.g., data warehouses and analytics stacks) for more effective feedback

Building practical, deployable data cleaning solutions for big data is a hard problem that is full of both engineering and algorithmic challenges; however, being programmatic does not mean being unprincipled.

CHAPTER 4

Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery

Michael L. Brodie

Over the past two decades, Data-Intensive Analysis (DIA)—also referred to as Big Data Analytics—has emerged not only as a basis for the *Fourth Paradigm* of engineering and scientific discovery, but as a basis for discovery in most human endeavors for which data is available. Though the idea originated in the 1960s, widespread deployment has occurred only recently, thanks to the emergence of big data and massive computing power. Data-Intensive Analysis is still in its infancy in its application and our understanding of it, and likewise in its development. Given the potential risks and rewards of DIA, and its breadth of application, it is imperative that we get it right.

The objective of this new Fourth Paradigm is more than simply acquiring data and extracting knowledge. Like its predecessor, the scientific method, the objective of the Fourth Paradigm is to *investigate phenomena* by acquiring new knowledge, and *to integrate it with* and use it to correct previous knowledge. It is now time to identify and understand the fundamentals. In my research, I have analyzed more than 30 large-scale use cases to understand current practical aspects, to gain insight into the fundamentals, and to address the fourth "V" of big data—*veracity*, or the accuracy of the data and the resulting analytics.

Data Science: A New Discovery Paradigm That Will Transform Our World

Big data has opened the door to profound change—to new ways of reasoning, problem solving, and processing that in turn bring new opportunities and challenges. But as was the case with its predecessor discovery paradigms, establishing this emerging Fourth Paradigm and the underlying principles and techniques of data science may take decades.

To better understand DIA and its opportunities and challenges, my research has focused on DIA use cases that are at *very large scale*—in the range where theory and practice may break. This chapter summarizes some key results of this research, related to understanding and defining data science as *a body of principles and techniques with which to measure and improve the correctness, completeness, and efficiency of Data-Intensive Analysis*.

Significance of DIA and Data Science

Data science is transforming discovery in many human endeavors, including healthcare, manufacturing, education, financial modeling, policing, and marketing [1][2]. It has been used to produce significant results in areas from particle physics (e.g., Higgs Boson), to identifying and resolving sleep disorders using Fitbit data, to recommenders for literature, theatre, and shopping. More than 50 national governments have established data-driven strategies as an official policy, in science and engineering [3] as well as in healthcare (e.g., the US National Institutes of Health and President Obama's Precision Medicine Initiative for "delivering the right treatments, at the right time, every time to the right person"). The hope, supported by early results, is that data-driven techniques will accelerate the discovery of treatments to manage and prevent chronic diseases that are more precise and are tailored to specific individuals, as well as being dramatically lower in cost.

Data science is being used to radically transform entire domains, such as medicine and biomedical research—as is stated as the purpose of the newly created Center for Biomedical Informatics at Harvard Medical School. It is also making an impact in economics [4], drug discovery [5], and many other domains. As a result of its successes and potential, data science is rapidly becoming a subdiscipline of most academic areas. These developments suggest a strong belief in the potential value of data science—but can it deliver?

The early successes and clearly stated expectations of data science are truly remarkable; however, its actual deployment, like that of many hot trends, is far less extensive than it might appear. According to Gartner's 2015 survey of Big Data Management and Analytics, 60% of the Fortune 500 companies claim to have deployed data science, yet less than 20% have implemented consequent significant changes, and less than 1% have optimized its benefits. Gartner concludes that 85% of these companies will be unable to exploit big data in 2015. The vast majority of deployments address tactical aspects of existing processes and static business intelligence, rather than realizing the power of data science by discovering previously unforeseen value and identifying strategic advantages.

Illustrious Histories: The Origins of Data Science

Data science is in its infancy. Few individuals or organizations understand the potential of and the paradigm shift associated with data science, let alone understand it conceptually. The high rewards and equally high risks, and its pervasive application, make it imperative that we better understand data science—its models, methods, processes, and results.

Data science is inherently multidisciplinary. Its principal components include mathematics, statistics, and computer science—especially areas of *artificial intelligence* such as machine learning, data management, and high-performance computing. While these disciplines need to be evaluated in the new paradigm, they have long, illustrious histories.

Data analysis developed over 4,000 years ago, with origins in Babylon (17th–12th c. BCE) and India (12th c. BCE). Mathematical analysis originated in the 17th c. around the time of the Scientific Revolution. While statistics has its roots in the 5th c. BCE and the 18th c. CE, its application in data science originated in 1962 with John W. Tukey [6] and George Box [7]. These long, illustrious histories suggest that data science draws on well-established results that took decades or centuries to develop. To what extent do they (e.g., statistical significance) apply in this new context?

Data science constitutes a new paradigm in the sense of Thomas S. Kuhn's scientific revolutions [8]. Data science's predecessor paradigm, the scientific method, was approximately 2,000 years in the development, starting with Aristotle (384–322 BCE) and continuing through Ptolemy (1st c. CE) and the Bacons (13th, 16th c. CE). Today, data science is emerging following the ~1,000-year development of its three predecessor paradigms of scientific and engineering discovery: theory, experimentation, and simulation [9]. Data science has been developing over the course of the last 50 years, but it changed qualitatively in the late 20th century with the emergence of big data—data whose *volumes*, *velocities*, and *variety* current technologies, let alone humans, cannot handle efficiently. This chapter addresses another characteristic that current technologies and theories do not handle well—*veracity*.

What Could Possibly Go Wrong?

Do we understand the risks of recommending the wrong product, let alone the wrong medical diagnoses, treatments, or drugs? The risks of a result that fails to achieve its objectives may include losses in time, resources, customer satisfaction, customers, and potentially business collapse. The vast majority of data science applications face such small risks, however, that veracity has received little attention.

Far greater risks could be incurred if incorrect data science results are acted upon in *critical contexts*, such as drug discovery [10] and personalized medicine. Most scientists in these contexts are well aware of the risks of errors, and go to extremes to estimate and minimize them. The announcement of the "discovery" of the Higgs boson at CERN's Large Hadron Collider (LHC) on July 4, 2012 might have suggested that the results were achieved overnight—they were not. The results took 40 years to achieve and included data science techniques developed over a decade and applied over big data by two independent projects, ATLAS and CMS, each of which were subsequently peer-reviewed and published [11][12] with a further year-long verification. To what extent do the vast majority of data science applications concern themselves with verification and error bounds, let alone understand the verification methods applied at CERN? Informal surveys of data scientists conducted in my research at data science conferences suggest that 80% of customers never ask for error bounds.

The existential risks of applying data science have been called out by world-leading authorities in institutions such as the Organisation for Economic Co-operation and Development and in the artificial

Intelligence (AI) [13][14][15][16] and legal [17] communities. The most extreme concerns have been stated by the Future of Life Institute, which has the objective of safeguarding life, developing optimistic visions of the future, and mitigating "existential risks facing humanity" from AI.

Given the potential risks and rewards of DIA and its breadth of application across conventional, empirical scientific and engineering domains, as well as across most human endeavors, we had better get this right! The scientific and engineering communities place high trust in their existing discovery paradigms, with well-defined measures of likelihood and confidence within relatively precise error estimates. Can we say the same for modern data science as a discovery paradigm, and for its results? A simple observation of the formal development of the processes and methods of its predecessors suggests that we cannot. Indeed, we do not know if, or under what conditions, the constituent disciplines—like statistics—may break down.

Do we understand DIA to the extent that we can assign probabilistic measures of likelihood to its results? With the scale and emerging nature of DIA-based discovery, how do we estimate the correctness and completeness of analytical results relative to a hypothesized discovery question? The underlying principles and techniques may not apply in this new context.

In summary, we do not yet understand DIA adequately to quantify the probability or likelihood that a projected outcome will occur within estimated error bounds. While the researchers at CERN used data science and big data to identify results, verification was ultimately empirical, as it must be in drug discovery [10] and other critical areas until analytical techniques are developed and proven robust.

Do We Understand Data Science?

Do we even understand what data science methods compute or how they work? Human thought is limited by the human mind. According to Miller's Law [4], the human mind (short-term working memory) is capable of holding on to less than 10 (7 +/– 2) concepts at one time. Hence, humans have difficulty understanding complex models involving more than 10 variables. The conventional process is to imagine a small number of variables, and then abstract or encapsulate that knowledge into a model that can subsequently be augmented with more variables. Thus, most scientific theories develop *slowly* over time into complex models. For example, Newton's model of physics was extended over the course of three centuries, through Bohr, Einstein, Heisenberg, and more, up to Higgs—to form the Standard Model of particle physics. Scientific discovery in particle physics is wonderful, but it has taken over 300 years. Due to its complexity, *no physicist has claimed to understand the entire Standard Model.*

When humans analyze a problem, they do so with models with a limited number of variables. As the number of variables increases, it becomes increasingly difficult to understand the model and the potential combinations and correlations. Hence, humans limit the scale of their models and analyses—which are typically theory-driven—to a level of complexity that they can comprehend.

But what if the phenomenon is arbitrarily complex or beyond immediate human conception? I suspect that this is addressed iteratively, with one model (theory) being abstracted as the base for another more complex theory, and so on (standing on the shoulders of those who have gone before), as the development of quantum physics from the discovery of elementary particles. That is, once the human mind understands a model, it can form the basis of a more complex model. This development under the scientific method scales at a rate limited by human conception, thus limiting the number of variables and the complexity. This is error-prone, since phenomena may not manifest at a certain level of complexity. Models correct at one scale may be wrong at a larger scale, or vice versa—a model wrong at one scale (and hence discarded) may become correct at a higher scale (a more complex model).

Machine learning algorithms can identify correlations between thousands, millions, or even billions of variables. This suggests that it is difficult or even impossible for humans to understand what (or how) these algorithms discover. Imagine trying to understand such a model that results from selecting some subset of the correlations on the assumption that they may be causal, and thus constitute a model of the phenomenon with high confidence of being correct with respect to some hypotheses, with or without error bars.

Cornerstone of a New Discovery Paradigm

The Fourth Paradigm—eScience supported by data science—is paradigmatically different from its predecessor discovery paradigms. It provides revolutionary new ways [8] of thinking, reasoning, and processing; new modes of inquiry, problem solving, and decision making. It is not the Third Paradigm augmented by big data, but something profoundly different. Losing sight of this difference forfeits its power and benefits, and loses the perspective that it is "a revolution that will transform how we live, work, and think" [2].

Paradigm shifts are difficult to notice as they emerge. There are several ways to describe the trend. There is a shift of resources, from (empirically) discovering causality (*why the phenomenon occurs*)—the heart of the scientific method—to discovering interesting correlations (*what might have occurred*). This shift involves moving from a *strategic perspective* driven by human-generated hypotheses (theory-driven, top-down) to a *tactical perspective* driven by observations (data-driven, bottom-up).

Seen at their extremes, the scientific method involves testing hypotheses (theories) posed by scientists, while data science can be used to generate hypotheses to be tested based on significant correlations among variables that are identified *algorithmically* in the data. In principle, vast amounts of data and computing resources can be used to accelerate discovery simply by outpacing human thinking in both power and complexity.

Data science is rapidly gaining momentum due to the development of ever more powerful computing resources and algorithms, such as *deep learning*. Rather than optimizing existing processes, data science can be used to identify patterns that suggest *unforeseen solutions*.

However, even more compelling is the idea that goes one step beyond the simple version of this shift—namely, a symbiosis of *both* paradigms. For example, data science can be used to offer several highly probable hypotheses or correlations, from which we select those with acceptable error estimates that are worthy of subsequent empirical analysis. In turn, empiricism can be used to pursue these hypotheses until some converge and some diverge, at which point data science can be applied to refine or confirm the converging hypotheses, and the cycle starts again. Ideally, one would optimize the combination of theory-driven empirical analysis with data-driven analysis to accelerate discovery to a rate neither on its own could achieve.

While data science is a cornerstone of a new discovery paradigm, it may be conceptually and methodologically more challenging than its predecessors, since it involves everything included in its predecessor, paradigms—modeling, methods, processes; measures of correctness, completeness, and efficiency—in a much more complex context, namely that of big data. Following well-established developments, we should try to find the fundamentals of data science—its principles and techniques—to help manage the complexity and guide its understanding and application.

Data Science: A Perspective

Since data science is in its infancy and is inherently multidisciplinary, there are naturally many definitions that emerge and evolve with the discipline. As definitions serve many purposes, it is reasonable to have multiple definitions, each serving different purposes. Most definitions of data science attempt to define *why* (its purpose), *what* (constituent disciplines), and *how* (constituent actions of discovery workflows).

A common definition of data science is *the activity of extracting knowledge from data*. While simple, this does not convey the larger goal of data science or its consequent challenges. A DIA activity is far more than a collection of actions, or the mechanical processes of acquiring and analyzing data. Like its predecessor paradigm, the scientific method, the purpose of data science and DIA activities is to *investigate phenomena* by *acquiring new knowledge, and correcting and integrating it with previous knowledge*—continually evolving our understanding of the phenomena, based on newly available data. We seldom start from scratch. Hence, discovering, understanding, and integrating data must precede extracting knowledge, and all at massive scale—i.e., largely by automated means.

The scientific method that underlies the Third Paradigm is a body of principles and techniques that provide the formal and practical bases of scientific and engineering discovery. The principles and techniques have been developed over hundreds of years, originating with Plato, and are still evolving today, with significant unresolved issues such as statistical significance (i.e., P values) and reproducibility.

While data science had its origins 50 years ago with Tukey [16] and Box [7], it started to change *qualitatively* less than two decades ago, with the emergence of big data and the consequent paradigm shift we've explored. The focus of this research into *modern* data science is on veracity—the ability to estimate the correctness, completeness, and efficiency of an end-to-end DIA activity and of its results. Therefore, we will use the following definition, in the spirit of Provost and Fawcett [5]:

> *Data science* is a body of principles and techniques for applying data-intensive analysis to investigate phenomena, acquire new knowledge, and correct and integrate previous knowledge with measures of correctness, completeness, and efficiency of the derived results, with respect to some predefined (top-down) or emergent (bottom-up) specification (scope, question, hypothesis).

Understanding Data Science from Practice

Methodology to Better Understand DIA

Driven by a passion for understanding data science in practice, my year-long and ongoing research study has investigated over 30 very large-scale big data applications—most of which have produced or are daily producing significant value. The use cases include particle physics; astrophysics and satellite imagery; oceanography; economics; information services; several life sciences applications in pharmaceuticals, drug discovery, and genetics; and various areas of medicine including precision medicine, hospital studies, clinical trials, and intensive care unit and emergency room medicine.

The aim of this study is to investigate relatively well understood, successful use cases where *correctness is critical* and the big data context is at massive scale; such use cases constitute less than 5% of all deployed big data analytics projects. The focus is on these use cases, as we do not know where errors (outside normal scientific and analytical errors) may arise. There is a greater likelihood that established disciplines such as statistics and data management might break at very large scale, where errors due to failed fundamentals may be more obvious.

The breadth and depth of the use cases revealed strong, significant emerging trends, some of which are listed below. These confirmed for some use case owners solutions and directions that they were already pursuing, and suggested to others directions that they could not have seen without the perspective of 30+ use cases.

DIA Processes

A *Data-Intensive-Activity* is an analytical process that consists of applying sophisticated analytical methods to large data sets that are stored under some analytical models (Figure 4-1). While this is the typical view of data science projects or DIA use cases, this analytical component of the DIA activity constitutes ~20% of an end-to-end DIA pipeline or workflow. Thus, currently it consumes ~20% of the resources required to complete a DIA analysis.

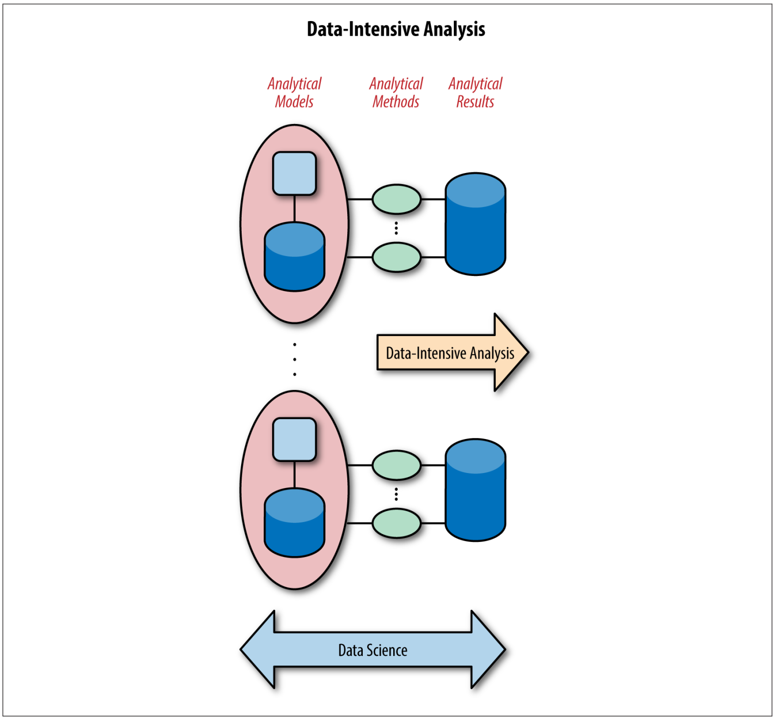


Figure 4-1. Conventional view of Data-Intensive Analysis

An end-to-end DIA activity (Figure 4-2) involves two data management processes that precede the DIA process, namely *raw data acquisition and curation* and *analytical data acquisition*. Raw data acquisition and curation starts with discovering and understanding data in data sources and ends with integrating and storing in a repository curated data that represents entities in the domain of interest, and metadata about those entities. Analytical data acquisition starts with discovering and understanding data within the shared repository and ends with storing the resulting information—specific entities and interpretations—into an analytical model to be used by the subsequent DIA process.

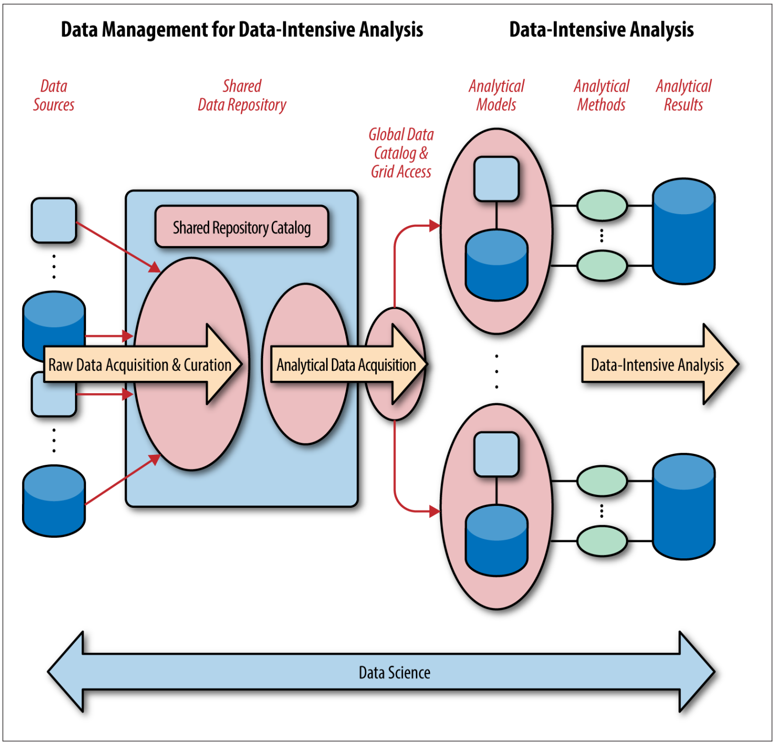

Data Management for Data-Intensive Analysis
Data-Intensive Analysis
Data Sources
Shared Data Repository
Analytical Models
Analytical Methods
Analytical Results
Global Data Catalog & Grid Access
Shared Repository Catalog
Raw Data Acquisition & Curation
Analytical Data Acquisition
Data-Intensive Analysis
Data Science

Figure 4-2. End-to-end Data-Intensive Analysis workflow

Sophisticated algorithms, such as machine learning algorithms, largely automate DIA processes, which have to be automated to process such large volumes of data using complex algorithms. Currently, raw data acquisition and curation and analytical data acquisition processes are far less automated, typically requiring 80% or more of the total resources to complete.

This understanding leads us to the following definitions:

Data-Intensive Discovery (DID)

The activity of using big data to investigate phenomena, to acquire new knowledge, and to correct and integrate previous knowledge. The "-Intensive" is added when the data is "at scale." Theory-driven DID is the application of human-generated scientific, engineering, or other hypotheses to big data. Data-driven DID employs automatic hypothesis generation.

Data-Intensive Analysis (DIA)

The process of analyzing big data with analytical methods and models.

DID goes beyond the Third Paradigm of scientific or engineering discovery by investigating scientific or engineering hypotheses using DIA. A DIA activity is an experiment over data, thus requiring all aspects of a scientific experiment—e.g., experimental design, expressed over data (a.k.a. *data-based empiricism*).

DIA process (workflow or pipeline)

A sequence of operations that constitute an end-to-end DIA activity, from the acquisition of the source data to the quantified, qualified result.

Currently, ~80% of the effort and resources required for the entire DIA activity are dedicated to the two data management processes—areas where scientists/analysts are not experts. Emerging technologies, such as those for data curation at scale, aim to flip that ratio from 80:20 to 20:80, to *let scientists do science and analysts do analysis*. This requires an understanding of the data management processes and their correctness, completeness, and efficiency, in addition to those of the DIA process. Another obvious consequence of the present imbalance is that proportionally, 80% of the errors that could arise in DIA may arise in the data management processes, prior to DIA even starting.

Characteristics of Large-Scale DIA Use Cases

The focus of my research is successful, very large-scale, multiyear projects with 100s–1,000s of ongoing DIA activities. These activities are supported by a *DIA ecosystem*, consisting of a community of users (e.g., over 5,000 scientists in the ATLAS and CMS projects at CERN and similar numbers of scientists using the worldwide Cancer Genome Atlas) and technology (e.g., *science gateways*, collectively referred to in some branches of science as *networked science*).

Some significant trends that have emerged from the analysis of these use cases are listed, briefly, below.

The typical view of data science appears to be based on the vast majority (~95%) of DIA use cases. While they share some characteristics with those in this study, there are fundamental differences, such as the concern for and due diligence associated with *veracity*.

Based on this study data, analysis appears to fall into three classes. *Conventional data analysis* over "small data" accounts for at least 95% of all data analysis, often using Microsoft Excel. DIA over big data has two subclasses: *simple DIA*, including the vast majority of DIA use cases mentioned above, and *complex DIA*, such as the use cases analyzed in this study, which are characterized by complex analytical models and a corresponding plethora of analytical methods. The models and methods are as complex as the phenomena being analyzed.

The most widely used DIA tools for simple cases profess to support analyst self-service in point-and-click environments, with some claiming "point us at the data and we will find the patterns of interest for you." This model is infeasible in the use cases analyzed, which are characterized not only by being machine-driven and human-guided, but by extensive attempts to optimize this *man–machine symbiosis* for scale, cost, and precision (too much human-in-the-loop leads to errors; too little leads to nonsense).

DIA ecosystems are inherently *multidisciplinary* (ideally interdisciplinary), *collaborative*, and *iterative*. Not only does DIA (Big Data Analytics) require multiple disciplines—e.g., genetics, statistics and machine learning—but so too do the data management processes—e.g., data management, domain and machine-learning experts for data curation, statisticians for sampling, and so on.

In large-scale DIA ecosystems, a DIA is a *virtual experiment* [18]. Far from claims of simplicity and point-and-click self-service, most large-scale DIA activities reflect the complexity of the analysis at hand and are the result of long-term (months to years) experimental designs. These designs necessarily involve greater complexity than their empirical counterparts, to deal with scale, significance, hypotheses, null hypotheses, and deeper challenges, such as determining causality from correlations and identifying and dealing with biases, and often irrational human intervention.

Finally, *veracity* is one of the most significant challenges and critical requirements of all DIA ecosystems studied. While there are many complex methods in conventional data science to estimate veracity, most owners of the use cases studied expressed concern for adequately estimating veracity in modern data science. Most assume that all data is imprecise, and hence require *probabilistic measures* and *error bars* and *likelihood* estimates for all results. More basically, most DIA ecosystem experts recognize that errors can arise across an end-to-end DIA activity and are investing substantially in addressing these issues in both the DIA process and the data management processes, which currently require significant human guidance.

An objective of this research is to discover the extent to which the characteristics of very large-scale, complex DIAs described here also apply to simple DIAs. There is a strong likelihood that they apply directly, but are difficult to detect—that is, that the principles and techniques of DIA apply equally to simple and complex DIA.

Looking Into a Use Case

Due to the detail involved, there is not space in this chapter (or this report) to fully describe a single use case considered in this study. However, let's look into a single step of a use case, involving a virtual experiment conducted at CERN in the ATLAS project.

The heart of empirical science is experimental design. It starts by identifying, formulating, and verifying a worthy hypothesis to pursue. This first complex step typically involves a multidisciplinary team, called the *collaborators* for this virtual experiment, often from around the world and for more than a year. We will consider the second step, the construction of the control or background model (executable software and data) that creates the background (e.g., an executable or testable model and a given data set) required as the basis within which to search (analyze) for "signals" that would represent the phenomenon being investigated in the hypothesis. This control completely excludes the data of interest. That is, the data of interest (the signal region) is "blinded" completely so as not to bias the experiment. The background (control) is designed using software that simulates relevant parts of the Standard Model of particle physics, plus data from ATLAS selected with the appropriate signatures with the data of interest blinded.

Over time, ATLAS contributors have developed simulations of many parts of the Standard Model. Hence, constructing the model required for the background involves selecting and combining relevant simulations. If there is no simulation for some aspect that is required, then it must be requested or built by hand. Similarly, if there is no relevant data of interest in the experimental data repository, it must be requested from a subsequent capture from the detectors when the LHC is next fired up in the appropriate energy levels. This comes from a completely separate team running the (non-virtual) experiment.

The development of the background is approximately a one-person-year activity, as it involves the experimental design, the design and refinement of the model (software simulations), the selection of methods and tuning to achieve the correct signature (i.e., get the right data), the verification of the model (observing expected outcomes when tested), and dealing with errors (statistical and systematic) that arise from the hardware or process. The result of the background phase is a model approved by the collaborative to represent the background required by the experiment with the signal region blinded. The model is an "application" that runs on the ATLAS "platform" using ATLAS resources—libraries, software, simulations, and data, drawing on the ROOT framework, CERN's core modeling and analysis infrastructure. It is verified by being executed under various testing conditions.

This is an incremental or iterative process, each step of which is reviewed. The resulting design document for the Top Quark experiment was approximately 200 pages of design choices, parameter settings, and results—*both positive and negative*! All experimental data and analytical results are probabilistic. All results have error bars; in particle physics they must be at least 5 sigma to be accepted. This explains the year of iteration in which analytical models are adjusted, analytical methods are selected and tuned, and results are reviewed by the collaborative. The next step is the actual virtual experiment. This too takes months. Surprisingly, once the data is unblinded (i.e., synthetic data is replaced in the region of interest with experimental data), the experimenter—often a PhD candidate —gets one and only one execution of the "verified" model over the experimental data.

Hopefully, this portion of a use case illustrates that Data-Intensive Analysis is a complex but critical tool in scientific discovery, used with a well-defined understanding of veracity. It must stand up to scrutiny that evaluates whether the experiment—consisting of all models, methods, and data with probabilistic results and error bounds better than 5 sigma—is adequate to be accepted by *Science* or *Nature* as demonstrating that the hypothesized correlation is causal.

Research for an Emerging Discipline

The next step in this research to better understand the theory and practice of the emerging discipline of data science, to understand and address its opportunities and challenges, and to guide its development is given in its definition. Modern data science builds on conventional data science and on all of its constituent disciplines required to design, verify, and operate end-to-end DIA activities, including both data management and DIA processes, in a DIA ecosystem for a shared community of users. Each discipline must be considered with respect to what it contributes to investigating phenomena, acquiring new knowledge, and correcting and integrating new with previous knowledge. Each operation must be understood with respect to the level of correctness, completeness, and efficiency that can be estimated.

This research involves identifying relevant principles and techniques. *Principles* concern the theories that are established formally—e.g., mathematically—and possibly demonstrated empirically. *Techniques* involve the application of wisdom [19]; i.e., domain knowledge, art, experience, methodologies, and practice—often called *best practices*. The principles and techniques, especially those established for conventional data science, must be verified and, if required, extended, augmented, or replaced for the new context of the Fourth Paradigm—especially its volumes, velocities, and variety. For example, new departments at MIT, Stanford, and the University of California, Berkeley, are conducting such research under what some are calling *21st-century statistics*.

A final, stimulating challenge is what is called *metamodeling* or *metatheory*. This area emerged in the physical sciences in the 1980s and subsequently in statistics and machine learning and is now being applied in other areas. Metamodeling arises when using multiple analytical models and multiple analytical methods to analyze different perspectives or characteristics of the same phenomenon. This extremely natural and useful methodology, called *ensemble modeling*, is required in many physical sciences, statistics, and AI, and should be explored as a fundamental modeling methodology.

Acknowledgment

I gratefully acknowledge the brilliant insights and improvements proposed by Prof Jennie Duggan of Northwestern University and Prof Thilo Stadelmann of the Zurich University of Applied Sciences.

References

[1] M.I. Jordan and T.M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349(6245), 255–260.

[2] V. Mayer-Schönberger and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt.

[3] National Science Foundation. 2008. Accelerating discovery in science and engineering through petascale simulations and analysis (PetaApps). Posted July 28, 2008.

[4] G.A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2), 81–97.

[5] F. Provost and T. Fawcett. 2013. Data science and its relationship to big data and data-driven decision making. *Big Data* 1(1), 51–59.

[6] J.W. Tukey. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33(1), 1–67.

[7] G.E.P. Box. 2012. Science and statistics. *Journal of the American Statistical Association* 71(356), 791–799. Reprint of original from 1962.

[8] T.S. Kuhn. 1996. *The Structure of Scientific Revolutions*. 3rd ed. Chicago, IL: University of Chicago Press.

[9] J. Gray. 2009. Jim Gray on eScience: A transformed scientific method. In A.J.G. Hey, S. Tansley, and K.M. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

[10] S. Spangler et al. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM*

SIGKDD international conference on knowledge discovery and data mining (KDD ’14). New York: ACM.

[11] G. Aad et al. 2012. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters* B716(1), 1–29.

[12] V. Khachatryan et al. 2012. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters* B716(1), 30–61.

[13] J. Bohannon. 2015. Fears of an AI pioneer. *Science*, 349(6245), 252.

[14] S.J. Gershman, E.J. Horvitz, and J.B. Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349(6245), 273–278.

[15] E. Horvitz and D. Mulligan. 2015. Data, privacy, and the greater good. *Science* 349(6245), 253–255.

[16] J. Stajic, R. Stone, G. Chin, and B. Wible. 2015. Rise of the machines. *Science* 349(6245), 248–249.

[17] N. Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. Tow Center.

[18] J. Duggan and M. Brodie. 2015. Hephaestus: Data reuse for accelerating scientific discovery. In CIDR 2015.

[19] B. Yu. 2015. Data wisdom for data science. ODBMS.org.

CHAPTER 5

From DevOps to DataOps

Andy Palmer

Why It's Time to Embrace "DataOps" as a New Discipline

Over the past 10 years, the technology industry has experienced the emergence of "DevOps." This new set of practices and tools have improved the velocity, quality, predictability, and scale of software engineering and deployment. Starting at the large Internet companies, the trend toward DevOps is now transforming the way that systems are developed and managed inside the enterprise—often dovetailing with enterprise cloud adoption initiatives. Regardless of your opinion about on-prem versus multitenant cloud infrastructure, the adoption of DevOps is improving how quickly new features and functions are delivered at scale for end users.

There is a lot to learn from the evolution of DevOps, across the modern Internet as well as within the modern enterprise—most notably for those who work with data every day.

At its core, DevOps is about the combination of software engineering, quality assurance, and technology operations (Figure 5-1). DevOps emerged because traditional systems management (as opposed to software development management) was not adequate to meet the needs of modern, web-based application development and deployment.

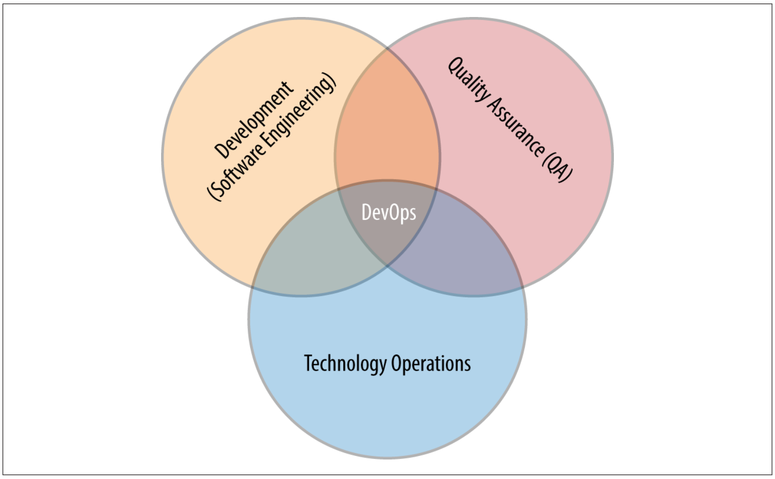


Figure 5-1. DevOps in the enterprise

From DevOps to DataOps

It's time for data engineers and data scientists to embrace a new, similar discipline—let's call it "DataOps"—that at its core addresses the needs of data professionals inside the modern enterprise.

Two trends are creating the need for DataOps:

1. *The democratization of analytics* is giving more individuals access to cutting-edge visualization, data modeling, machine learning, and statistical tools.
2. *The implementation of "built-for-purpose" database engines* is improving the performance and accessibility of large quantities of data, at unprecedented velocity. The techniques to improve beyond legacy relational DBMSs vary across markets, and this has driven the development of specialized database engines such as StreamBase, Vertica, VoltDB, and SciDB.

   More recently, Google made its massive Cloud Bigtable database (the same one that powers Google Search, Maps, YouTube, and Gmail) available to everyone in a scalable NoSQL database service through the Apache HBase API.

Together, these trends create pressure from both "ends of the stack." From the top of the stack, users want access to more data in more combinations. From the bottom of the stack, more data is available than ever before—some aggregated, but much of it not. The only way for data professionals to deal with the pressure of heterogeneity from both the top and bottom of the stack is to embrace a new approach to managing data This new approach blends operations and collaboration. The goal is to organize and deliver data from many sources, to many users, *reliably*. At the same time, it's essential to maintain the *provenance* required to support reproducible data flows.

Defining DataOps

DataOps is a *data management method* used by data engineers, data scientists, and other data professionals that emphasizes:

- Communication
- Collaboration
- Integration
- Automation

DataOps acknowledges the *interconnected nature* of data engineering, integration, quality, and security and privacy. It aims to help an organization rapidly deliver data that accelerates analytics, and to enable previously impossible analytics.

The "ops" in DataOps is very intentional. The operation of infrastructure required to support the quantity, velocity, and variety of data available in the enterprise today is radically different from what traditional data management approaches have assumed. The nature of DataOps embraces the need to manage *many* data sources and *many* data pipelines, with a wide variety of transformations.

Changing the Fundamental Infrastructure

While people have been managing data for a long time, we're at a point now where the quantity, velocity, and variety of data available to a modern enterprise can no longer be managed without a significant change in the fundamental infrastructure. The design of this infrastructure must focus on:

- The thousands of sources that are not centrally controlled, and which frequently change their schemas without notification

(much in the way that websites change frequently without notifying search engines)

- Treating these data sources (especially tabular data sets) as if they were websites being published inside of an organization

DataOps challenges preconceived notions of how to *engage with* the vast quantities of data being collected every day. Satisfying the enormous appetite for this data requires that we sort it in a way that is rapid, interactive, and flexible. The key to DataOps is that you don't have to theorize and manage your data schemas up front, with a misplaced idealism about how the data should look.

DataOps Methodology

Using DataOps methodology, you start with the data as it is and work from the bottom up. You work with it, integrate it, uncover insights along the way, and find more data and more data sources that support or add to what you have discovered. Eventually, you come away with more quality outcomes than if you had tried to sort through the information from the top down with a specific goal in mind.

DataOps methodology brings a more agile approach to interrogating and analyzing data, on a very large scale. At some point, what you want is all the data. If you have *all* the data in a clear, comprehensible format, then you can actually see things that other people can't see. But you can't reach that monumental goal by simply declaring that you're going to somehow conjure up all of the data in one place—instead, you have to continually iterate, execute, evaluate, and improve, just like when you are developing software.

If you want to do a better job with the quality of the data you are analyzing, you've got to develop *information-seeking behaviors*. The desire to look at more information and use more data sources gives you better signals from the data and uncovers more potential sources of insight. This creates a virtuous cycle: as data is utilized and processed, it becomes well organized and accessible, allowing more data to emerge and enter the ecosystem.

Any enterprise data professional knows that data projects can quickly become insurmountable if they rely heavily on manual processes. DataOps requires automating many of these processes to quickly incorporate new data into the existing knowledge base.

First-generation DataOps tools (such as Tamr's Data Unification platform) focus on making agile data management easier.

Integrating DataOps into Your Organization

Much of what falls under the umbrella of big data analytics today involves idiosyncratic and manual processes for breaking down data. Often, companies will have hundreds of people sifting through data for connections, or trying to find overlap and repetition. Despite the investment of these resources, new sources of data actually make this work harder—much, much harder—which means more data can *limit* instead of improve outcomes. DataOps tools will eliminate this hypolinear relationship between data sources and the amount of resources required to manage them, making data management automated and truly scalable.

To integrate this revolutionary data management method into an enterprise, you need two basic components. The first is cultural—enterprises need to create an environment of communication and cooperation among data analytics teams. The second component is technical—workflows will need to be automated with technologies like machine learning to recommend, collect, and organize information. This groundwork will help radically simplify administrative debt and vastly improve the ability to manage data as it arrives.

The Four Processes of DataOps

As illustrated in Figure 5-2, four processes work together to create a successful DataOps workflow:

- Engineering
- Integration
- Quality
- Security

Within the context of DataOps, these processes work together to create meaningful methods of handling enterprise data. Without them, working with data becomes expensive, unwieldy, or—worse—unsecure.

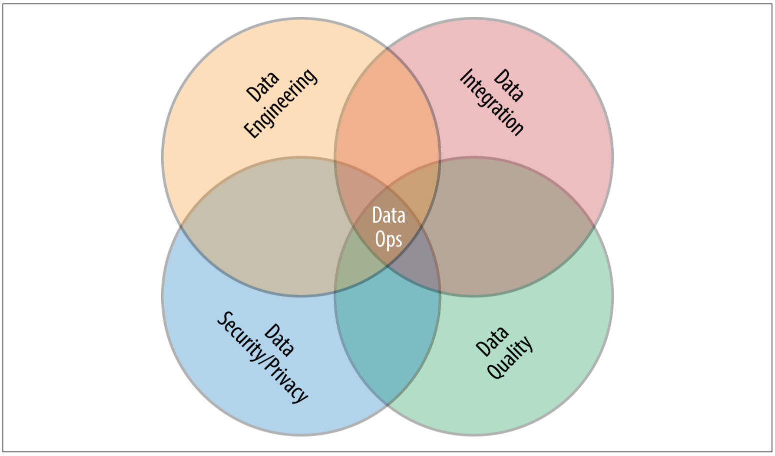


Figure 5-2. Four processes of DataOps

Data Engineering

Organizations trying to leverage all the possible advantages derived from mining their data need to move quickly to create *repeatable* processes for productive analytics. Instead of starting with a specific analytic in mind and working through a manual process to get to that endpoint, the data sifting experience should be optimized so that the most traditionally challenging, but least impactful, aspects of data analysis are automated.

Take, for example, the management of customer information in a CRM database or other database product. Sorting through customer data to make sure that the information is accurate is a challenge that many organizations either address manually—which is bad—or don't address at all, which is worse. No company should be expected to have bad data or be overwhelmed by working with its data in an age when machine learning can be used as a balm to these problems.

The central problem of previous approaches to data management was the lack of automation. The realities of manually bringing together data sources restricted projects' goals and therefore limited the focus of analytics—and if the analytical outcomes did not match the anticipated result, the whole effort was wasted. Moving to DataOps ensures that foundational work for one project can give a jump-start to the next, which expands the scope of analytics.

A bias toward automation is even more critical when addressing the huge variety of data sources that enterprises have access to. Only enterprises that engineer with this bias will truly be able to be data-driven—because only these enterprises will begin to approach that lofty goal of gaining a handle on all of their data.

To serve your enterprise customers the right way, you have to deliver the right data. To do this, you need to engineer a process that automates getting the right data to your customers, and to make sure that the data is well integrated for those customers.

Data Integration

Data integration is the mapping of physical data entities, in order to be able to differentiate one piece of data from another.

Many data integration projects fail because most people and systems lack the ability to differentiate data correctly for a *particular use case*. There is no one schema to rule them all; rather, you need the ability to flexibly create new logical views of your data within the context of your users' needs. Existing processes that enterprises have created usually merge information too literally, leading to inaccurate data points. For example, often you will find repetitive customer names or inaccurate email data for a CRM project; or physical attributes like location or email addresses may be assigned without being validated.

Tamr's approach to data integration is "machine driven, human guided." The "machines" (computers running algorithms) organize certain data that is similar and should be integrated into one data point. A small team of skilled analysts validate whether the data is right or wrong. The feedback from the analysts informs the machines, continually improving the quality of automation over time. This cycle can remove inaccuracies and redundancies from data sets, which is vital to finding value and creating new views of data for each use case.

This is a key part of DataOps, but it doesn't work if there is nothing actionable that can be drawn from the data being analyzed. That value depends on the quality of the data being examined.

Data Quality

Quality is purely subjective. DataOps moves you toward a system that recruits users to improve data quality in a bottom-up, bidirectional way. The system should be bottom-up in the sense that data quality is not some theoretical end state imposed from on high, but rather is the result of real users engaging with and improving the data. It should be bidirectional in that the data can be manipulated and dynamically changed.

If a user discovers some weird pattern or duplicates while analyzing data, resolving these issues immediately is imperative; your system must give users this ability to submit instant feedback. It is also important to be able to manipulate and add more data to an attribute as correlating or duplicate information is uncovered.

Flexibility is also key—the user should be open to what the data reveals, and approach the data as a way to feed an initial conjecture.

Data Security

Companies usually approach data security in one of two ways—either they apply the concept of access control, or they monitor usage.

The idea of an *access control* policy is that there has to be a way to trace back who has access to which information. This ensures that sensitive information rarely falls into the wrong hands. Actually *implementing* an access control policy can slow down the process of data analysis, though—and this is the existing infrastructure for most organizations today.

At the same time, many companies don't worry about who has access to which sets of data. They want data to flow freely through the organization; they put a policy in place about how information can be used, and they watch what people use and don't use. However, this leaves companies potentially susceptible to malicious misuse of data.

Both of these data protection techniques pose a challenge to combining various data sources, and make it tough for the right information to flow freely.

As part of a system that uses DataOps, these two approaches need to be combined. There needs to be some access control and use monitoring. Companies need to manage who is using their data and why, and they also always need to be able to trace back how people are using the information they may be trying to leverage to gain new big data insights. This framework for managing the security of your data is necessary if you want to create a broad data asset that is also protected. Using both approaches—combining some level of access control with usage monitoring—will make your data more fluid and secure.

Better Information, Analytics, and Decisions

By incorporating DataOps into existing data analysis processes, a company stands to gain a more granular, better-quality understanding of the information it has and how best to use it. The most effective way to maximize a system of data analytics is through viewing data management not as an unwieldy, monolithic effort, but rather as a fluid, incremental process that aligns the goals of many disciplines.

If you balance out the four processes we've discussed (engineering, integration, quality, and security), you'll empower the people in your organization and give them a game-changing way to interact with data and to create analytical outcomes that improve the business.

Just as the movement to DevOps fueled radical improvements in the overall quality of software and unlocked the value of information technology to many organizations, DataOps stands to radically improve the quality and access to information across the enterprise, unlocking the true value of enterprise data.

CHAPTER 6

Data Unification Brings Out the Best in Installed Data Management Strategies

James Markarian

Companies are now investing heavily in technology designed to control and analyze their expanding pools of data, reportedly spending $44 billion for big data analytics alone in 2014. In relation, data management software now accounts for over 40 percent of the total spend on software in the US With companies focusing on strategies like ETL (extract, transform, and load), MDM (master data management), and data lakes, it's critical to understand that while these technologies can provide a unique and significant handle on data, they still fall short in terms of speed and scalability—with the potential to delay or fail to surface insights that can propel better decision making.

Data is generally too siloed and too diverse for systems like ETL, MDM, and data lakes, and analysts are spending too much time finding and preparing data manually. On the other hand, the nature of this work defies complete automation. *Data unification* is an emerging strategy that catalogs data sets, combines data across the enterprise, and publishes the data for easy consumption. Using data unification as a *frontend strategy* can quicken the feed of highly organized data into ETL and MDM systems and data lakes, increasing the value of these systems and the insights they enable. In this chapter, we'll explore how data unification works with installed data management solutions, allowing businesses to embrace data volume and variety for more productive data analyses.

Positioning ETL and MDM

When enterprise data management software first emerged, it was built to address data variety and scale. ETL technologies have been around in some form since the 1980s. Today, the ETL vendor market is full of large, established players, including Informatica, IBM, and SAP, with mature offerings that boast massive installed bases spanning virtually every industry. ETL makes short work of repackaging data for a different use—for example, taking inventory data from a car parts manufacturer and plugging it into systems at dealerships that provide service, or cleaning customer records for more efficient marketing efforts.

Extract, Transform, and Load

Most major applications are built using ETL products, from finance and accounting applications to operations. ETL products have three primary functions for integrating data sources into single, unified datasets for consumption:

1. *Extracting* data from data sources within and outside of the enterprise
2. *Transforming* the data to fit the particular needs of the target store, which includes conducting joins, rollups, lookups, and cleaning of the data
3. *Loading* the resulting transformed dataset into a target repository, such as a data warehouse for archiving and auditing, a reporting tool for advanced analytics (e.g., business intelligence), or an operational database/flat file to act as reference data

Master Data Management

MDM arrived shortly after ETL to create an authoritative, top-down approach to data verification. A centralized dataset serves as a "golden record," holding the approved values for all records. It performs exacting checks to assure the central data set contains the most up-to-date and accurate information. For critical business decision making, most systems depend on a consistent definition of "master data," which is information referring to core business operational elements. The primary functions of master data *management* include:

- *Consolidating* all master data records to create a comprehensive understanding of each entity, such as an address or dollar figure
- *Establishing survivorship*, or selecting the most appropriate attribute values for each record
- *Cleansing* the data by validating the accuracy of the values
- *Ensuring compliance* of the resulting single "good" record related to each entity as it is added or modified

Clustering to Meet the Rising Data Tide

Enterprise data has changed dramatically in the last decade, creating new difficulties for products that were built to handle mostly static data from relatively few sources. These products have been extended and overextended to adjust to modern enterprise data challenges, but the workaround strategies and patches that have been developed are no match for current expectations.

Today's tools, like Hadoop and Spark, help organizations reduce the cost of data processing and give companies the ability to host massive and diverse datasets. With the growing popularity of Hadoop, a significant number of organizations have been creating data lakes, where they store data derived from structured and unstructured data sources in its raw format.

Upper management and shareholders are challenging their companies to become more competitive using this data. Businesses need to integrate massive information silos—both archival and streaming—and accommodate sources that change constantly in content and structure. Further, every organizational change brings new demand for data integration or transformation. The cost in time and effort to make all of these sources analysis-ready is prohibitive.

There is a chasm between the data we can access thanks to Hadoop and Spark and the ordered information we need to perform analysis. While Hadoop, ETL, and MDM technologies (as well as many others) prove to be useful tools for storing and gaining insight from data, collectively they can't resolve the problem of bringing massive and diverse datasets to bear on time-sensitive decisions.

Embracing Data Variety with Data Unification

Data variety isn't a problem; it is a natural and perpetual state. While a single data format is the most effective starting point for analysis, data comes in a broad spectrum of formats for good reason. Data sets typically originate in their most useful formats, and imposing a single format on data negatively impacts that original usefulness.

This is the central struggle for organizations looking to compete through better use of data. The value of analysis is inextricably tied to the amount and quality of data used, but data siloed throughout the organization is inherently hard to reach and hard to use. The prevailing strategy is to perform analysis with the data that is easiest to reach and use, putting expediency over diligence in the interest of using data before it becomes out of date. For example, a review of suppliers may focus on the largest vendor contracts, focusing on small changes that might make a meaningful impact, rather than accounting for all vendors in a comprehensive analysis that returns five times the savings.

Data unification represents a philosophical shift, allowing data to be raw and organized at the same time. Without changing the source data, data unification prepares the varying data sets for any purpose through a combination of automation and human intelligence.

The process of unifying data requires three primary steps:

1. *Catalog:* Generate a central inventory of enterprise metadata. A central, platform-neutral record of metadata, available to the entire enterprise, provides visibility of what relevant data is available. This enables data to be grouped by logical entities (customers, partners, employees), making it easier for companies to discover and uncover the data necessary to answer critical business questions.
2. *Connect:* Make data across silos ready for comprehensive analysis at any time while resolving duplications, errors, and inconsistencies among the source data's attributes and records. Scalable data connection enables data to be applied to more kinds of business problems. This includes matching multiple entities by taking into account relationships between them.

3. *Publish:* Deliver the prepared data to the tools used within the enterprise to perform analysis—from a simple spreadsheet to the latest visualization tools. This can include functionality that allows users to set custom definitions and enrich data on the fly. Being able to manipulate external data as easily as if it were their own allows business analysts to use that data to resolve ambiguities, fill in gaps, enrich their data with additional columns and fields, and more.

Data Unification Is Additive

Data unification has significant value on its own, but when added to an IT environment that already includes strategies like ETL, MDM, and data lakes, it turns those technologies into the best possible versions of themselves. It creates an ideal data set for these technologies to perform the functions for which they are intended.

Data Unification and Master Data Management

The increasing volume and frequency of change pertaining to data sources poses a big threat to MDM speed and scalability. Given the highly manual nature of traditional MDM operations, managing more than a dozen data sources requires a large investment in time and money. Consequently, it's often very difficult to economically justify scaling the operation to cover all data sources. Additionally, the speed at which data sources are integrated is often contingent on how quickly employees can work, which will be at an increasingly unproductive rate as data increases in volume.

Further, MDM products are very deterministic and up-front in the generation of matching rules. It requires manual effort to understand what constitutes potential matches, and then define appropriate rules for matching. For example, in matching addresses, there could be thousands of rules that need to be written. This process becomes increasingly difficult to manage as data sources become greater in volume; as a result, there's the risk that by the time new rules (or rule changes) have been implemented, business requirements will have changed.

Using data unification, MDM can include the long tail of data sources as well as handle frequent updates to existing sources—reducing the risk that the project requirements will have changed before the project is complete. Data unification, rather than replacing MDM, works in unison with it as a system of reference, recommending new "golden records" via matching capability and acting as a repository for keys.

Data Unification and ETL

ETL is highly manual, slow, and not scalable to the number of sources used in contemporary business analysis. Integrating data sources using ETL requires a lot of up-front work to define requirements, target schemas, and establish rules for matching entities and attributes. After all of this work is complete, developers need to manually apply these rules to match source data attributes to the target schema, as well as to deduplicate or cluster entities that appear in many variations across various sources.

Data unification's probabilistic matching provides a far better engine than ETL's rules when it comes to matching records across all of these sources. Data unification also works hand-in-hand with ETL as a system of reference to suggest transformations at scale, particularly for joins and rollups. This results in a faster time-to-value and more scalable operation.

Changing Infrastructure

Additionally, data unification solves the biggest challenges associated with changing infrastructure—namely, unifying datasets in Hadoop to connect and clean the data so that it's ready for analytics. Data unification creates integrated, clean datasets with unrivaled speed and scalability. Because of the scale of business data today, it is very expensive to move Hadoop-based data outside of the data lake. Data unification can handle all of the large-scale processing within the data lake, eliminating the need to replicate the entire data set.

Data unification delivers more than technical benefits. In unifying enterprise data, enterprises can also unify their organizations. By cataloging and connecting dark, disparate data into a unified view, for example, organizations illuminate what data is available for analysts, and who controls access to the data. This dramatically reduces discovery and prep effort for business analysts and "gatekeeping" time for IT.

Probabilistic Approach to Data Unification

The probabilistic approach to data unification is reminiscent of Google's full-scale approach to web search and connection. This approach draws from the best of machine and human learning to find and connect hundreds or thousands of data sources (both visible and dark), as opposed to the few that are most familiar and easiest to reach with traditional technologies.

The first step in using a probabilistic approach is to catalog all metadata available to the enterprise in a central, platform-neutral place using both machine learning and advanced collaboration capabilities. The data unification platform automatically connects the vast majority of sources while resolving duplications, errors, and inconsistencies among source data. The next step is critical to the success of a probabilistic approach—where algorithms can't resolve connections automatically, the system must call for expert human guidance. It's imperative that the system work with people in the organization familiar with the data, to have them weigh in on mapping and improving the quality and integrity of the data. While expert feedback can be built into the system to improve the algorithms, it will always play a role in this process. Using this approach, the data is then provided to analysts in a ready-to-consume condition, eliminating the time and effort required for data preparation.

About the Authors

Jerry Held has been a successful entrepreneur, executive, and investor in Silicon Valley for over 40 years. He has been involved in managing all growth stages of companies from conception to multibillion dollar global enterprises.

He is currently CEO of Held Consulting LLC and a mentor at Studio 9+, a Silicon Valley incubator. Dr. Held is chairman of Tamr, MemSQL, and Software Development Technologies. He serves on the boards of NetApp (NTAP), Informatica (INFA), Kalio, and Copia. From 2006 to 2010, he served as executive chairman of Vertica Systems (acquired by HP) and lead independent director of Business Objects from 2002 to 2008 (acquired by SAP).

In 1998, Dr. Held was "CEO-in-residence" at the venture capital firm Kleiner Perkins Caufield & Byers. Through 1997, he was senior vice president of Oracle Corporation's server product division, leading a division of 1,500 people and helping the company grow revenues from $1.5 billion to $6 billion annually. Prior to Oracle, he spent 18 years at Tandem Computers, where he was a member of the executive team that grew Tandem from a startup to a $2 billion company. Throughout his tenure at Tandem, Dr. Held was appointed to several senior management positions, including chief technology officer, senior vice president of strategy, and vice president of new ventures. He led the initial development of Tandem's relational database products.

Dr. Held received a B.S. in electrical engineering from Purdue, an M.S. in systems engineering from the University of Pennsylvania, and a Ph.D. in computer science from the University of California, Berkeley, where he led the initial development of the INGRES relational database management system. He also attended the Stanford Business School's Executive Program.

Dr. Held is also a member of the board of directors of the Tech Museum of Innovation.

Michael Stonebraker is an adjunct professor at MIT CSAIL and a database pioneer who specializes in database management systems and data integration. He was awarded the 2014 A.M. Turing Award (known as the “Nobel Prize of computing”) by the Association for Computing Machinery for his “fundamental contributions to the concepts and practices underlying modern database systems as well as their practical application through nine start-up companies that he has founded.”

Professor Stonebraker has been a pioneer of database research and technology for more than 40 years, and is the author of scores of papers in this area. Before joining CSAIL in 2001, he was a professor of computer science at the University of California Berkeley for 29 years. While at Berkeley, he was the main architect of the INGRES relational DBMS; the object-relational DBMS POSTGRES; and the federated data system Mariposa. After joining MIT, he was the principal architect of C-Store (a column store commercialized by Vertica), H-Store, a main memory OLTP engine (commercialized by VoltDB), and SciDB (an array engine commercialized by Paradigm4). In addition, he has started three other companies in the big data space, including Tamr, oriented toward scalable data integration. He also co-founded the Intel Science and Technology Center for Big Data, based at MIT CSAIL.

Tom Davenport is the President’s Distinguished Professor of Information Technology and Management at Babson College, the co-founder of the International Institute for Analytics, a Fellow of the MIT Center for Digital Business, and a Senior Advisor to Deloitte Analytics. He teaches analytics and big data in executive programs at Babson, Harvard Business School, MIT Sloan School, and Boston University. He pioneered the concept of “competing on analytics” with his best-selling 2006 *Harvard Business Review* article (and his 2007 book by the same name). His most recent book is *Big Data@Work*, from Harvard Business Review Press. It surprises no one that Tom has once again branched into an exciting new topic. He has extended his work on analytics and big data to its logical conclusion–what happens to us humans when smart machines make many important decisions? Davenport and Julia Kirby, his frequent editor at *Harvard Business Review*, published the lead/cover article in the HBR June 2015 issue. Called “Beyond Automation,” it’s the first article to focus on how individuals and organizations can add value to the work of cognitive technologies. It argues for “augmentation”—people and machines working alongside each other—over automation. Davenport and Kirby will also publish a book on this topic with Harper Business in 2016.

Professor Davenport has written or edited seventeen books and over 100 articles for *Harvard Business Review*, *Sloan Management Review*, the *Financial Times*, and many other publications. He also writes a weekly column for the *Wall Street Journal*'s Corporate Technology section. Tom has been named one of the top three business/technology analysts in the world, one of the 100 most influential people in the IT industry, and one of the world's top fifty business school professors by *Fortune* magazine.

Tom earned a Ph.D. from Harvard University in social science and has taught at the Harvard Business School, the University of Chicago, Dartmouth's Tuck School of Business, Boston University, and the University of Texas at Austin.

Ihab Ilyas is a Professor in the Cheriton School of Computer Science at the University of Waterloo. He received his PhD in computer science from Purdue University, West Lafayette. He holds BS and MS degrees in computer science from Alexandria University. His main research is in the area of database systems, with special interest in data quality, managing uncertain data, rank-aware query processing, and Information extraction. From 2011 to 2013 he has been on leave leading the Data Analytics Group at the Qatar Computing Research Institute. Ihab is a recipient of an Ontario Early Researcher Award, a Cheriton Faculty Fellowship, an NSERC Discovery Accelerator Award, and a Google Faculty Award. He is also an ACM Distinguished Scientist. Ihab is a co-founder of Tamr, a startup focusing on large-scale data integration and cleaning.

Michael L. Brodie has over 40 years experience in research and industrial practice in databases, distributed systems, integration, artificial intelligence, and multi-disciplinary problem solving. He is concerned with the "big picture" aspects of information ecosystems, including business, economic, social, applied, and technical aspects. Dr. Brodie is a Research Scientist, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology; advises startups; serves on Advisory Boards of national and international research organizations; and is an adjunct professor at the National University of Ireland, Galway and at the University of Technology, Sydney.

For over 20 years he served as Chief Scientist of IT, Verizon, a Fortune 20 company, responsible for advanced technologies, architectures, and methodologies for IT strategies and for guiding industrial scale deployments of emerging technologies. His current research and applied interests include big data, data science, and data curation at scale, and the related start, Tamr. He has also served on several National Academy of Science committees.

Dr. Brodie holds a PhD in Databases from the University of Toronto and a Doctor of Science (honoris causa) from the National University of Ireland. He has two amazing children Justin Brodie-Kommit (b.3/1/1990) and Kayla Kommit (b. 1/19/1995).

Andy Palmer is co-founder and CEO of Tamr, a data analytics start-up—a company he founded with fellow serial entrepreneur and 2014 Turing Award winner Michael Stonebraker, PhD, adjunct professor at MIT CSAIL; Ihab Ilyas, University of Waterloo; and others. Previously, Palmer was co-founder and founding CEO of Vertica Systems, a pioneering big data analytics company (acquired by HP). He also founded Koa Labs, a shared start-up space for entrepreneurs in Cambridge's Harvard Square. During his career as an entrepreneur, Palmer has served as founding investor, BOD member or advisor to more than 50 start-up companies in technology, healthcare and the life sciences. He also served as Global Head of Software and Data Engineering at Novartis Institutes for BioMedical Research (NIBR) and as a member of the start-up team and Chief Information and Administrative Officer at Infinity Pharmaceuticals. Additionally, he has held positions at Bowstreet, pcOrder.com, and Trilogy.

James Markarian is the former CTO of Informatica, where he spent 15 years leading the data integration technology and business as the company grew from a startup to over a $1 billion revenue company. He has spoken on data and integration at Strata, Hadoop World, TDWI and numerous other technical and investor events. Currently James is an investor in and advisor to many startup companies including Tamr, DxContinuum, Waterline, StreamSets, and EnerAllies. Previously, he was an an Entrepreneur in Residence (EIR) at Khosla Ventures, focussing on integration and business intelligence. He got his start at Oracle in 1988 where he was variously a developer, manager and a member of the company-wide architecture board. James has a B.A. in Computer Science and B.A. and M.A. in Economics from Boston University.