O'REILLY®

Self-Service Analytics

Making the Most of Data Access

Sandra Swanson

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn 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 # 15420

Self-Service Analytics

Making the Most of Data Access

Sandra Swanson

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Self-Service Analytics
by Sandra Swanson

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: Tim McGovern
Interior Designer: David Futato
Cover Designer: Randy Comer
Cover Image: Michael Seeley

January 2016: First Edition

Revision History for the First Edition
2016-01-15: First Release
2016-03-23: Second Release

978-1-491-93900-0

[LSI]

Table of Contents

Self-Service Analytics

More than ever before, organizations are swimming in oceans of data. But that doesn't necessarily lead to a surge in business insights. Companies estimate that they are only analyzing about 12% of their data, according to Forrester Research. To help build a stronger data-driven culture, organizations are turning to self-service analytics. This approach provides data access to more people within the company, and allows them to combine disparate sources of data and create their own customized analysis. "It's an approach to analytics that enables the person to access and work with data, without the dependence on someone from the IT department," says Jean-Michel Franco, Director of Product Marketing for Talend. "It lets you find the information you need, so you can be autonomous; it cuts out waiting for someone not only to create your own reports and dashboards, but also to collect, shape, and connect the datasets that are needed for your analysis."

To Provide the Right Tools, Watch and Listen

Tom Schenk, Chief Data Officer for the City of Chicago, has personally observed the benefits of increased access to data. The city has 33,000 employees spread across 30 departments, from garbage collection to public-safety services like police and fire departments, to libraries and building inspectors. "It's absolutely necessary in a large organization like ours to allow individual users access to data, to be able to answer questions for their commissioner or their boss," he says. Although not all 33,000 employees access that data, hundreds of them do. "It enables fundamental things like performance metrics, for departments that use it to drive decision making," he says.

That move to self-service has allowed city employees to be more responsive to their own departments or divisions, instead of waiting on someone else to provide the data they need.

Schenk notes that the real power of self-service will come from multi-variant analytics, allowing users to look at an array of variables and tease out correlations. His organization is working toward providing that capability, particularly in the realm of predictive analytics. The City of Chicago already uses predictive analytics to help identify which restaurants are most likely to have food violations (based on variables such as the weather and complaints about garbage in nearby streets). That's vital, considering there are only three dozen inspectors and more than 15,000 food establishments. "Right now, it takes a lot of human intervention and a lot of time to do these sort of research projects," says Schenk. It's becoming possible to take a self-service approach instead, with machine learning and other techniques that do some of the analytical heavy lifting. "We would like to get to that point, so we won't have to spend as many hours getting it done," he says—noting that almost every city department has responsibility for doing some sort of inspection. A self-service approach would significantly improve the efficiency of those inspections.

For effective self-service, one of the greatest challenges is ensuring that users have the right tools to facilitate data exploration. "Having a completeness of toolsets is key in order to allow those individuals to navigate data and communicate with data," says Schenk. The best way to achieve that is not by just offering a variety of tools, but also offering tools that are actually needed. That requires listening closely to users, says Schenk. He recommends setting up advisory groups to get constant feedback from users. "For instance, mapping is very important for running a city operation, but in other organizations, that may be superfluous," he says. Schenk also notes it's critical to try tools out, not just buy-and-deploy after a quick demo. "If you are looking at a visualization tool that might make sense, don't just take one and implement it," he says. Pilot a handful, and see what works best for users.

Also, watch for employees who use tools in ways they weren't designed for, as a clue for unmet needs. Schenk has seen that happen several times and notes that it represents a deeper underlying issue. One Chicago report developer, for example, went to great and impressive lengths to create a dashboard-like report. This took some

significant time and talent, but clearly marked where there wasn't a sufficient dashboard application—which would have saved time and let the developer focus on the data—available to them. "It was just representing that we didn't have the right toolset for them," he says. "We keep an eye out for how we can do a better job to make it easier for those departments." End-user service is what's crucial here, he says—because without listening to the user, attempts at self-service analytics will not go well.

Sumeet Singh, Senior Director of Product Management for Cloud and Big Data Platforms at Yahoo isn't a fan of the term "self-service," because it doesn't capture an important aspect of democratizing data. "For widespread use, what matters is how easy it is to use," he says. For Yahoo's data platform, end users range from very savvy, data-trained engineers to sales and marketing employees who aren't as knowledgeable.

To facilitate that ease-of-use, Singh says his organization has become "tool agnostic," meaning employees can bring many different types of BI and analytics tools to the platform. "You can use SAP, Excel, Tableau, whatever you want." That's important, because the learning curve for each tool can vary greatly. This approach allows employees to use tools within their comfort zone. "I call this *data to desktop*—we will bring data to your desktop in whatever form or fashion you want to consume that data," he says.

When Yahoo's platform wasn't so easy to use, employees would contact the company's central reporting team with their requests. Depending on the complexity of those requests, it could take six months to turn around a customized report solution. Now it can happen in 10 seconds. "There's a world of difference between a self-serve environment and one that is custom and request based, where you have a central team that has knowledge of data and reporting tools, and is building reports for people across the company," he says. "That model just wasn't viable, and didn't allow us to move at the speed which we needed."

Data-centric Tools Shift to Line-of-Business Users

As more organizations focus on data-driven decision making, that has prompted a growing demand for data access. Jean-Michel

Franco of Talend sees those data-centric tools shifting to line-of-business users. "If you are a marketing department, you want to make sure all of your marketing decisions can be challenged with data," says Franco. His company provides data integration capabilities that help organizations make their information ready for users to consume.

"You need more and more access to data, simply to do your job—and you can't be dependent on a third party if it's part of your daily job." Franco compares that with the financial responsibilities of managers—they need the ability to autonomously manage their P & L, but must also comply with corporate rules. "The same thing is happening with data," he says.

Beyond access and analysis, self-service data preparation is the next frontier; it's an emerging but swiftly growing market. Gartner has predicted that: "By 2017, most business users and analysts in organizations will have access to self-service tools to prepare data for analysis." This represents a further shift in power from IT to business units, with the rewards of faster and more customized provisioning of data.

That shift toward more widespread access to data also reflects organizations' efforts to offer customers additional guidance when needed. Franco notes that one of Talend's clients is a company that provides healthcare services, and it needs to provide personalized healthcare guidance to customers. "The assistants need to be able to say, *According to your health plan, you should go to this hospital*—so those assistants need a lot of data at their fingertips to provide the best advice, and they need to access it in an agile way." Customers now expect more guidance from a number of industries, he says. To achieve that, organizations require more data and more access for employees.

Create a Path for More "What If" Exploration

The Financial Industry Regulatory Authority (FINRA) is a nonprofit organization that regulates the securities industry; it must balance the need for speed and accuracy with massive amounts of data. It monitors financial markets, looking for fraud and manipulation—which requires watching nearly 6 billion shares traded daily and processing approximately 6 terabytes of data daily, bringing in data-

sets from different equity exchanges as well as options exchanges and fixed income markets.

About two years ago, FINRA started to update platforms, and self-service analytics was part of the overall strategy behind that. The organization has a couple of main lines of business—market regulation and member regulation—but they each have different work groups with very specific focuses, such as insider trading or market manipulation or compliance. That means some users might look for activity that took place in half a second, while others will scrutinize activity during the course of a year. "There is a whole variety and uniqueness of questions," says Scott Donaldson, Senior Director for Market Regulation Technology at FINRA. "We had a legacy platform where you would bring in the data and create analytic models up on top of that," says Donaldson. "By the time you get it built, the user says, 'Oh, we want to ask this other question.' And it's very, very time-consuming. All of these information requests basically were little technology projects."

With the updated platform, FINRA gives employees the ability to answer their own questions with the right data—and without picking up the phone to call IT. To that end, it developed an application called Diver, which allows users to obtain slices of data from the trillions of records in FINRA's data ocean. These chunks of data—which FINRA calls private data marts—could contain 100 records, or several billion, depending on the user's query.

Once users have that dataset, they can probe it and follow a line of investigation. "Our internal phrase is, users want to have dialogue with data," says Donaldson. "When you're working with it, you want to be able to interrogate it." If analysts can quickly obtain the full picture of what happened to order over time, it helps inform their decisions as to whether a rule violation occurred. "It gives them more intuitive exploratory analysis," says Donaldson. Users now have the ability to ask more "what if" questions, which is vital when trying to determine if fraud or manipulation occurred. "Completeness and accuracy is extremely important," he continues. "Although you're looking at an order in a particular point in time, you need to view it in context of all the other orders or what's happening on that market or other exchanges." That means users need to be able to query and build context on multiple levels: what are the various market conditions at the time, for example, and is there a pattern or practice from a particular firm that might constitute market manip-

ulation? "What we're doing is lowering the barrier to entry, so users are able to do more complex analysis at scale." With self service, requests that might have taken hours or days for IT to complete can now be executed by the user in seconds.

The Benefits of Metadata

FINRA tracks the metadata, from ETL to the user's last interaction with the data. Ultimately, that improves the user experience, says Donaldson, because it allows FINRA to learn more about how data helps employees do their job.

"At the end-user perspective, we're tracking everything from what query parameters people included, and tracking what operations they performed on data—filtered it by this, sorted it by that," says Donaldson. "We provide that to them, so if they need to come back a year from now and reproduce those steps, it's there."

Donaldson and his team also observe what users do with data—and look for patterns that could simplify the user experience. "If people are always doing an aggregation step or summary, then maybe that's something we should summarize for them," he says. "It's about agility and adapting. We are constantly monitoring and reviewing the actions of users, trying to see what future feature we want to offer in the platform, what other data models do we want to do. When you see 9 out of 10 users doing the same thing, you can say, 'We could automate that for you.' It drives a level of efficiency back into the platform."

About three years ago, Yahoo started encouraging users to register their data in a central metastore. This allows employees to browse individual clusters to see what datasets are available, and they can also search for common words they might associate with certain datasets (such as "audience" for clickstream data).

"Once we have registered all of the company's data in the central metastore, we can expose the catalog in a very central fashion," says Yahoo's Singh. The schema, the semantics, all kinds of details about the data are transparently available to employees. "But I'm not exposing data," notes Singh. "I'm just exposing information about data to people." If employees find a dataset that could be valuable for their work, they can request access from the same portal they use to browse and search datasets.

The number of employees who use the platform has exploded during the past two years, says Singh, and that wouldn't have been possible without Yahoo's emphasis on ease of use. "Years ago, I didn't ever have a person in marketing or sales using the platform, because it was considered too techie or too difficult," says Singh. "But that has changed. They don't go to the reporting guy and say, 'Pull this report." They can do it themselves. We even have a senior vice president who runs Hadoop jobs—that's the power of this thing."

For Singh, the most valuable part of isn't the data itself—it's the ability to track what happens to it. When needed, he can monitor individuals' activities on the platform down to the granular level of who opened a particular file and what that person did with it. Yahoo has billions of files, 3,000 users, and 600 terabytes of storage space—which would make the self-service environment a very difficult problem without the right capability to audit the platform. "Being able to answer questions like 'Who accessed this data set in the last three months?'—and answer it very quickly—is extremely important. It helps us meet a lot of regulatory compliance, for example, and ad hoc audits for consumer privacy."

The vast majority of organizations don't have an easy way to do those audits quickly, and that's a big problem, says Singh. "That ability to audit is critical as you go more and more into self-service," he says. "You are making it extremely easy for people to access and to work with data, but one of the pitfalls is you expose yourself to greater risk." The best way to resolve conflict between security and access, he says, is audit and monitoring.

Develop a Culture of Data Literacy

To democratize access to data, organizations don't need to transform employees into data scientists. But it is important to foster baseline data literacy, particularly for decision makers, says Carl Anderson, Director of Data Science at eyewear retailer Warby Parker. The company has expanded access to store data, allowing each store manager to see how foot traffic correlates with sales and number of staff on hand, for example. "If you have an open service culture, you want everyone to have skills to use business intelligence tools," says Anderson. And that means broad data literacy among decision makers. They are the ones taking the analysis and making important

strategic decisions—and many don't have a data background, notes Anderson.

There are a few skills that decision makers and managers should have. He's created a class to teach those skills to Warby Parker managers and decision makers, and the company also has an internal book on experimental design that it uses quite broadly. "The decision makers should be data skeptical and at least have a firm grounding in experimental design. You want them to be able to see if people stretched interpretation of the data too far." These classes are one way Warby Parker enables a common language and broad ability to evaluate data analysis.

Don't Overlook Data Governance

When organizations consider providing increased access to data, they first need to take a close look at their data governance processes to ensure data security and quality. Report accuracy can represent a big problem in self-service environments. Often, it's only detected in management meetings when the users realize their self-service BI data doesn't match, notes Kurt Schlegel, Research Vice President at Gartner. He studies how organizations report and analyze data for performance management and decision support. "The problem stems from end users creating reports that look correct—because the SQL statement has returned rows and columns—but conceal logical flaws. These are most frequently caused by overestimated or underestimated measures, based on mistakes in joining fact and dimension tables," says Schlegel. Because they're hard to identify, those problems can persist undetected in flawed reports that drive important business decisions, he says.

One potential answer: Provide business users with only simple semantic layers to create reports—single fact tables with basic dimensions, says Schlegel. Another approach he suggests: Build a cultural understanding that ad hoc reports and analyses are not meant to be used as systems-of-record to run the business. "Several companies have achieved success by enabling users to create their own reports and even to share them in public folders, but in these companies it is culturally understood that unless these reports go through a rigorous validation process (usually carried out by the central BI team), they should be treated as containing preliminary

results," he says. "Only once validated are they put in a system-of-record folder."

FINRA's Donaldson underscores the need for rigorous data governance within a self-service environment. "I think there's a misconception sometimes that self-service means, 'We'll just point tools at a bunch of different datasets and the users can bring it all together and do whatever they want with it," he says. But that leads to a couple big, problematic questions, says Donaldson: How are you governing that? And how do you make sure people are interpreting the data in the same way? "Not everyone is level-set in terms of what is in the data and the semantics of what the data actually means. So it's very important for us to have safeguards and guard rails around this," he says. There are nuances in the type of orders within different market systems, and some FINRA users might not be aware of those subtle differences, so Donaldson's team has to normalize and standardize that to treat data in equal fashion. Because FINRA is charged with identifying violations that could ultimately go to arbitration or result in legal fines, accuracy is critical. "The semantics of that data is extremely important, and you have to provide that along with the self-service analytics, to make sure that it's all being held in a consistent manner," says Donaldson.

People will use the data if they trust the data, he says. That means organizations have to fit their governance with whatever standards their industry dictates. "I deal in highly regulated industry—I can't do a rounding error," says Donaldson. "If you're in marketing, you can deal with probablistics—if customers buy this shirt, they might want these shoes. The type of activity you're looking for defines the data governance aspect of it."

Most organizations talk about governance as a pain point. But if they focused on governance from the outset, they'd recognize it can be a blessing, says Singh. "If you design your process around governance correctly, it can be amazingly helpful." Some basic steps can help organizations handle data governance. "One is very simple classification—like levels of sensitivity, and what are the implications if someone gains access," he says. For example, with certain financial data, it might be OK for a person to access it, as long as that person agrees to be treated as an insider who will have blackouts for trading stock.

Collaboration with IT

When organizations consider self-service, they should remember that it doesn't eliminate the need for IT (or any other organizations in charge of delivering the data). Instead, it requires a new way of collaborating. "Sometimes self-service can look a bit scary, especially to IT people," says Talend's Franco. That can stem from the misconception that self-service is a completely do-it-yourself data experience that can lead to chaos in terms of control and security. "Self-service data preparation tools allow analysts to create data inventories that can include some data delivered by IT as *sanctioned sources,* meaning that they are ready for consumption, quality roofed, and protected against fraudulent usage," he continues. "Through these capabilities, the user gets more than the raw material, but rather value-added datasets that are ready for reuse, while IT can control what is being done with the data. So I think the distinction between self-service and do-it-yourself is important, because it introduces the idea that there's the need for data governance."

The self-service environment creates a different role for IT. "It used to be, if you need info, IT will provide that," says Franco. With self-service, IT is there to provide guidance and help the user gain autonomy. IT acts as a change agent to help users achieve their data goals, while adhering to rules and standards. "Sometimes self-service is seen as conflict of, *Is it line-of-business, or is it IT?* The real answer is, it should be both," says Franco.

Stay Central and Avoid Silos

In most companies, data is fragmented and siloed—and it's a disaster, says Yahoo's Singh. That's something organizations should avoid when they create a self-service environment. At the outset, companies need to ask: "How do we build this platform in a way that maximizes value out of every ounce of data we have? Most companies don't think about that," says Singh. It's paramount to have the company's data in one place. For self-service, that's the first step toward providing ease of use around data. "It's a lot easier to discover in a single platform than in silos of data, given the complex data organizations can collect. Some call it a data lake, or enterprise data hub—but the idea is consolidation, so users have a way to easily discover data throughout the organization."

Singh's organization is ingrained in Hadoop and open source, but he realizes that's not the case for all companies. "As long as you focus on keeping data centralized when you go about building the platform, I think organizations will do the right tool selection."

At his organization, greater access to data has led to more effective discussions about business decisions. "For some issues that would have involved 2 hours' worth of argument, now it takes 10 seconds," he says. "You can say, 'Here's what happened to engagement when we released this feature.' I'm not saying it's led to success of certain products, but at least you're very clear on what works and what doesn't. More decisions are now being driven by data, versus intuition or experience or a hunch. And that's a huge value to the organization."

About the Author

Sandra Swanson jumped out of a plane once, for the sake of journalism (it was a skydiving story). Since then, she's found plenty of invigorating work that doesn't require plummeting toward the ground at 120 mph. As a Chicago-based writer, she's covered technology, science, and business for dozens of publications. She's also keen on topics that intersect in unexpected ways. One example: "Toys That Inspired Scientific Breakthroughs," a story she wrote for ScientificAmerican.com. Connect with her on Twitter (*@saswanson*) or at *www.saswanson.com*.