Self-Service Big Data Preparation in the Age of Hadoop

A CONVERSATION WITH SAS
As enterprises take on the challenge of deploying, managing, and trying to wrest value from big data technologies such as Hadoop, they’re coming to grips with what might seem like an insurmountable problem: big data management and the lack of specialized skills required to prepare that data. No, that isn’t an oxymoron. It is possible to manage big data without advanced coding skills.

IT organizations are discovering that the difficulty of managing information at big data scale is compounded by the limitations of the Hadoop platform itself and finding resources with the required skill sets. Hadoop combines a distributed file system—HDFS, the Hadoop Distributed File System—with a baked-in parallel processing engine called MapReduce that is time-consuming to create and manage. This one-two punch is disruptive in that it brings processing to data where it lives (i.e., to file blocks distributed across HDFS) instead of bringing data to processing, which is how it’s done in the classic paradigm. The thing is, Hadoop is not actually a database management system. Unlike a database, Hadoop lacks features or safeguards designed to address data consistency, data integrity, and other key data management (DM) concerns.

What’s an IT organization to do? Is it simply impossible to promote good data management and good data governance at big-data scale? SAS Institute Inc. (SAS) doesn’t think so.

Officials say SAS’s new Data Loader for Hadoop is designed to extend the aegis of data preparation in the world of big data and Hadoop to business users. SAS Data Loader for Hadoop permits business users—along with data scientists and IT personnel—to access, profile, transform, and cleanse Hadoop data with minimal training. Under the hood, it supports SQL query access to data stored in Hadoop and helps automate the movement of data into and out of the Hadoop platform. Power users and SAS coders can run SAS statistical, data mining, and data-quality code and parallelize it across a Hadoop cluster.

In other words, SAS Data Loader for Hadoop extends an automated, self-service data management layer to the Hadoop platform. “This isn’t the stuff
of rocket science," says Tamara Dull, director of emerging technologies for SAS. Think of it, instead, as a pragmatic approach to big data management. According to Dull, “big” data poses some of the same challenges as does “small” data: some (usually fractional) proportion of it must be vetted to meet critical quality and consistency standards; a much larger proportion of it can be managed less stringently.

“Not all data—big or small—is created equally, and, therefore, doesn’t need to be treated equally. In other words, one size does not fit all,” she says.

Take data quality, for example. The pragmatic issues that govern data quality in traditional data management—Whose data quality? Quality at what cost?—are no less applicable in big data management. No single standard of data quality is “right” for each and every class of user, in all contexts, at all times. If a data quality program produces immaculately clean, impeccably consistent data but takes days or weeks to do so, is it worth it? Sometimes, analysts don’t need to be so fussy about the quality of the data they’re working with—even if the alternative is for them to sit around waiting for IT to procure the data they’ve asked for.

The Promise and the Peril of Hadoop

Why is there so much hoopla about Hadoop?

For one thing, Hadoop’s one-two punch of distributed storage and parallel processing enables fundamentally new applications and use cases at significantly lower costs. This has less to do with the “size” of data, or for that matter with the advent of “new” types of data, than with Hadoop’s ability to store, organize, and analyze data of all shapes and sizes. “This new age of big data is forcing organizations to look at their data—big and small, existing and potential—in a new light,” Dull says, stressing that “we can’t afford to fall into the trap of believing that the ‘data’ in ‘big data’ is something new, because it’s not.”

Dull ticks through a litany of the most popular “new” big data types: e-mail, photos, videos, spreadsheets, PDF documents, satellite images, social media data, blogs, audio files, GPS data, call center transcripts, open data, RSS feeds, and clickstream data.

“Are any of these data sources new? Of course not. We’ve been collecting this data for many, many years. So what’s new? Why all the hype? The simple answer is technology. We now have big data technology ... that allows us to store all this data at a fraction of the cost and process it in a fraction of the time—as compared to our existing, traditional systems,” she says.

“Not only that, we can combine, fuse, and integrate all this data—regardless of volume or format—like never before. As a result, we are able to get more creative with our data analysis, discover more insights, answer more and better business questions, and subsequently improve the bottom line for our organization.”

For the most part, these new use cases and best practices complement and extend the capabilities of a traditional BI analytic infrastructure. Because it lacks critical data management features—and because, moreover, it isn’t actually a database—Hadoop is a comparatively poor platform for (ad hoc) SQL query. It might be far from sexy, but the vast majority of day-to-day business decision making will continue to be powered by traditional BI reporting and analytics. That said, there are several critical ways in which Hadoop complements or extends the data warehouse.

A data quality program—and good data governance—should be pragmatic enough to meet the legitimate needs of all potential users. In the same way, data governance shouldn’t be used as a straw man: the data warehouse and its well-managed data management infrastructure evolved over time to control the consistency, accuracy, and accountability of information.

Big data won’t be governed overnight, but it will be governed.

Because Hadoop is itself an impoverished platform for data management, Dull argues, SAS Data Loader for Hadoop is a critical enabling technology for big data management.

If a data quality program produces immaculately clean, impeccably consistent data but takes days or weeks to do so, is it worth it?

A data quality program—and good data governance—should be pragmatic enough to meet the legitimate needs of all potential users. In the same way, data governance shouldn’t be used as a straw man: the data warehouse and its well-managed data management infrastructure evolved over time to control the consistency, accuracy, and accountability of information.

Big data won’t be governed overnight, but it will be governed.

Because Hadoop is itself an impoverished platform for data management, Dull argues, SAS Data Loader for Hadoop is a critical enabling technology for big data management.
First, from a data management perspective, Hadoop can function as a cost-effective and practically elastic distributed storage resource. It is orders of magnitude cheaper to store data in Hadoop than a data warehouse. Thomas Davenport, a senior adviser at Deloitte Analytics and a senior research fellow at MIT’s Center for Digital Business, puts the cost of Hadoop-based storage at 23 cents per gigabyte per month and that of online storage in a data warehouse at $19 per gigabyte per month.

Hadoop isn’t just an inexpensive storage sink, however. Much like a massively parallel processing (MPP) database system, which also distributes storage across clustered nodes, Hadoop doubles as a parallel data processing platform. However, unlike an MPP system, Hadoop is a distributed file system with a general-purpose parallel processing engine baked into it. This combination is new and valuable, inasmuch as it turns the traditional decision support model on its head. After all, a database—MPP or otherwise—expects to structure and persist data in a rigid, schematic way; it has to “know” what it’s storing before it stores it.

Hadoop, on the other hand, is a schema-on-read data processing platform: it can land and store data of any and all kinds without having to “know” anything about it; schema can be derived once it’s accessed for processing. Hadoop is also much less costly—and much less fussy—than an MPP database: you don’t have to worry as much about configuring, sizing, and balancing new Hadoop nodes. For these reasons and others, organizations are starting to use Hadoop as a long-term repository in which to land, prepare, and persist data.

This is why vendors such as SAS are so excited about Hadoop, says Dull, who argues that the Hadoop platform fundamentally changes how SAS customers can acquire, manage, and analyze data. After all, even though Hadoop is a comparatively poor SQL query platform, it supports a range of advanced NoSQL analytics use cases for which the data warehouse itself is fundamentally unsuited.

“Big data has expanded how SAS works with data. Traditionally, the process has been for SAS customers to move their data [in]to SAS to do their analysis. This can be a costly process, not to mention the time and drag it can have on the network. Now with the volumes of data big data brings, moving data from server to server isn’t ideal or even recommended in some cases. The less you have to move the data around, the better,” Dull points out. “SAS customers have traditionally had to take their data to the processing. Customers can now take the processing to their data.”

The challenge for a vendor such as SAS, with its four decades of data management and analytic expertise, is to accommodate the new, advantageous, and disruptive capabilities that Hadoop introduces while at the same time promoting the required levels of data quality, consistency, and control. “This is why SAS takes a holistic approach to big data management,” says Dull. Its support for Hadoop spans the data-to-decision life cycle—from data management to data discovery, analytics model development to deployment.
“You can access and integrate data from Hadoop, push SAS processing and data quality functions into the Hadoop cluster, or lift data into memory and perform distributed data processing, data exploration, analytic computations, and more,” she says.

Shoring up Hadoop's Shortcomings

From a data management perspective, Dull argues, SAS Data Loader for Hadoop addresses several of the Hadoop platform’s most critical data management defects. Remember, Hadoop’s storage layer is a file system, not a database management system. Its default data management feature set consists of Hive—a SQL-like interpreter that compiles HiveQL queries into MapReduce or Apache Tez jobs—and HCatalog, a metadata services catalog. (This is changing, but it will be years before Hadoop’s data management feature set matches that of a mature (R)DBMS.)

Hive is an interpreter for Hadoop’s baked-in parallel computing layer; as such, it has nothing to do with how Hadoop actually stores and organizes data in HDFS. At the file system level, for example, Hadoop persists data in any of several data serialization formats (including Avro data files and SequenceFiles); as text, CSV, JSON, or XML files; and in formats such as Parquet files (a column store format) or Record Columnar Files, both of which attempt to optimize how HDFS organizes table data.

Most data warehouse platforms, and virtually all MPP analytic appliances, use optimized technologies to accelerate data loading. A good example is Teradata’s Parallel Data Transporter, which can ingest data at the rate of more than a dozen terabytes an hour. In contrast, vanilla Hadoop doesn’t have a native bulk loader that’s optimized for ingesting and organizing tabular or relational data. Instead, the best or most efficient ways to get data in and out of Hadoop tend to entail lots of programming.

These include:

- Writing MapReduce jobs in Java or other languages that translate ETL operations into map and reduce operations
- Writing jobs in PigLatin for Hadoop’s Pig framework
- Writing SQL in Hive Query Language (HiveQL)
- Exploiting programmer-oriented technologies, such as Cascading

“The Apache Hadoop project does two functions extremely well: store any kind of data—structured, semi-structured, and unstructured—and process the data. The process of loading data into Hadoop is straightforward: data is ingested and stored in Hadoop’s Distributed File System through custom … scripts … typically written by programmers,” SAS’s Dull notes. “While this process is effective and allows organizations to store and process data at just a fraction of the cost as compared to traditional relational systems, it is new technology that requires new skills. It also requires a lot of custom development work, and it’s not really designed for the technically savvy data professional.”

More important, Dull notes, vanilla Hadoop cannot address data quality and other critical data management issues. “Hadoop pales in comparison when you match its capabilities to that of a tried-and-true data management solution, so even though you can load lots of data very quickly into HDFS, you shouldn’t really expect [the Hadoop platform] to be the gatekeeper for all data coming into it. That’s where a tool such as SAS Data Loader for Hadoop comes in handy,” she explains.

Organizations today are faced with unprecedented volumes of often dirty data, and a gap in the skills and resources needed to manage that data.

Working with Hadoop can seem comparatively primitive, at least from the perspective of developers or DBAs accustomed to using SQL to describe, manipulate, and access data in a database. Until recently, Cloudera, Hortonworks, MapR, and other Hadoop distributions lacked mature, functional graphical user interfaces. True, these vendors now deliver attractive GUI management consoles that provide at least some measure of visibility into what’s happening in a Hadoop cluster. These tools are typically able to identify problems; they’re much less useful at explaining why something went wrong, however.
The absence of a mature, functional UI, to say nothing of a lack of critical data management features, together limit Hadoop’s potential usefulness, at least as a robust platform for data management. Administrative access to Hadoop was, and to some extent still is, largely programmatic. Programming in Hadoop requires data-processing-specific knowledge and skills, along with the requisite proficiency in Java, Pig Latin, Python, Clojure, Scala, and other languages.

“The dearth of Hadoop-specific skills is a limiting factor for many would-be adopters,” says Matthew Magne, global product marketing manager for data management with SAS. Skills of this kind tend to be highly specialized and in high demand.

“Organizations need a way to overcome the gap in available Hadoop skills, speed access to big data, and improve performance so that they can empower business users, reduce the data provisioning burden on IT, and increase productivity for all involved,” he argues. “Organizations today are faced with unprecedented volumes of often dirty data, and a gap in the skills and resources needed to manage that data.”

Recasting Hadoop for Self Service

SAS takes a different tack with its Data Loader for Hadoop product. Apache Hadoop is rightly viewed as a playground for programmers and programming-savvy data scientists. Data Loader recasts Hadoop as a self-service playground for business analysts and savvy business users by automating (or by making self-serviceable) many otherwise-manual tasks in Hadoop. It shields self-service users from the complexity of working in the Hadoop environment by providing a simple point-and-click user interface that permits business users to create, run, and save “directives” (i.e., reusable data management tasks).

Data Loader recasts Hadoop as a self-service playground for business analysts and savvy business users by automating (or by making self-serviceable) many otherwise-manual tasks in Hadoop.

“[It] allows business analysts and data scientists to perform data integration, data quality, and data preparation tasks on big data without requiring specialized skills or training,” says Dull, citing its wizard-driven UI. “Data Loader for Hadoop provides self-service access to big data and unburdens IT by allowing users access to profile, manage, and cleanse data on the Hadoop cluster.”

SAS Data Loader for Hadoop exposes a familiar programmatic interface for data scientists and data quality developers. Instead of coding in Java, Pig Latin, Clojure, Scala, or in other, even more arcane languages, it’s possible to program using familiar SAS code or SAS data-quality functions and to execute them in Hadoop. In fact, Dull points out, Data Loader for Hadoop supports a full range of SAS data quality functions, including profiling, deduplicating, standardizing, parsing, generating match codes, and identification analysis.
SAS “brings the math to the data” and runs as an “embedded process” in Hadoop—much like it does in Teradata, a parallel processing platform optimized for analytic query performance. This means Hadoop schedules and parallelizes SAS workloads much like it does MapReduce or Tez jobs—with a twist. Because SAS’s embedded process engine is optimized for analytic and data quality processing, SAS workloads will actually run much faster in Hadoop than will the equivalent MapReduce or Tez workloads. (Hadoop MapReduce is a general-purpose parallel processing engine; it is not optimized for ETL or SQL analytic workloads.)

“The beauty of these two technologies is that the business users employing them need have no knowledge of what they are or how they work,” Magne explains. “The SAS Embedded Process is a lightweight SAS execution engine that ... runs natively on Hadoop, Teradata, and other relational databases. It provides stability and consistency to a Hadoop ecosystem that is often in flux. You don’t have to write MapReduce code in order to be able to execute code on the Hadoop cluster in parallel for improved performance. Underneath the hood, SAS Data Loader leverages the SAS Embedded Process to parallelize the work and realize improved performance.”

Data Quality Second to None

Data quality is arguably SAS Data Loader’s killer app, Magne argues. Out-of-the-box Hadoop has an impoverished—if not quite nonexistent—data quality feature set. Data Loader opens up SAS’s best-of-breed data quality feature set to the Hadoop platform. (See Gartner Magic Quadrant for Data Quality Tools press release.)

For example, Magne notes, there’s the SAS Quality Knowledge Base (QKB), a set of prebuilt data quality rules that can be customized by customers building on years of experience as a leader in this space. “This is at the core of a lot of our data quality technology at SAS and really separates us from the pack,” he points out, explaining that the QKB features prebuilt rules that, for example, allow parsing blocks of text into address fields such as street, name, ZIP code, as well as for standardizing state codes. It can also guess what type of entity (organization or person, for example) a string of text represents through identity analysis.

“The user simply chooses ‘parse’ to glean sense from a block of semi-structured text. Underneath the hood, [SAS data quality] intelligence is being brought to bear,” Magne says.

“Bringing this level of data quality to Hadoop is a real win for our customers. Being able to leverage the same set of data quality rules across both Hadoop and traditional data sources creates greater reuse of skill sets and continuity across the organization.”

SAS Data Loader supports both loading into and extraction out of Hadoop. It permits users to copy data from relational data sources or SAS data sets into Hadoop—and vice versa. Comparatively few technologies are able to automate read-write access to Hadoop, and fewer still are able to automate ETL or ELT on Hadoop itself (such as Data Loader for Hadoop).

However, that’s just what SAS technology does. For business users, Data Loader supports drag-and-drop data movement to and from the Hadoop environment. This is particularly useful in data discovery scenarios: a self-serving user can quickly prepare data for visualization by copying it from Hadoop to SAS’s LASR in-memory analytic server. This emphasis on self-service is no accident, says Dull. “Self-service is the key to improving the productivity of business users and data scientists and bridging the Hadoop skills gap.”

In the same way, SAS Data Loader for Hadoop supports SQL access to Hadoop. Users can view, filter, aggregate, join, and query data in Hadoop. This is particularly useful when Hadoop is used as a long-term repository or archive for OLTP data (e.g., as a so-called Hadoop “data lake”). Even though the concept of a data lake is tendentious—skeptics deride it as a “data swamp”—it does have value as a persistent store for raw OLTP data.

Think about it: it isn’t at all uncommon for a data warehouse—which contains a conformed subset of OLTP data—to lack critical information that data scientists or business analysts need. This data isn’t actually “missing”; it lives, or it once lived, in raw transaction data. A well-curated Hadoop data lake gives IT a means to persist and manage raw data from OLTP systems, along with other types of data—
streaming, time-series, text, and so on—that can’t cost-effectively be managed by a data warehouse. If or when an analyst needs it, it’s there for the getting without requiring IT to add yet another column of data. “The idea is that instead of storing your data in a structured data store, such as a data warehouse, you store it in a [typically Hadoop-based] data lake in its original format,” Dull says.

Even though the concept of a data lake is tendentious—skeptics deride it as a “data swamp”—it does have value as a persistent store for raw OLTP data.

Hadoop only makes sense as a data lake, data archive, or data-what-have-you if you have a cost-effective way to move stuff in and out of it. Dull says SAS Data Loader gives you just this.

First, it makes it easy to persist and retrieve your raw transaction data. Only a fraction of this data is actually used in day-to-day BI reporting and analytics. From the perspective of a business analyst or data scientist, however, this data—far from being superfluous—is nothing less than invaluable: your quants are going to clamor for as much of it as possible. If you can store raw data at trifling cost in Hadoop, and if your analysts can access and retrieve it themselves, why not let them do so?

Second, SAS Data Loader for Hadoop both simplifies and automates parallel processing in the Hadoop environment. As we’ve noted, writing data processing code for MapReduce and Tez requires a deep knowledge of data engineering—(how many business analysts know what a directed acyclic graph is?)—and expertise in Java, Pig Latin, Python, and other languages.

Third, Data Loader lets you treat Hadoop as a queryable archive: you can query data in situ without having to prepare or extract it.

Fourth, Data Loader can help accelerate your advanced analytic efforts. Remember, Hadoop can be used to store and analyze data types—for example, audio, video, and image data—that the SQL language and the data warehouse itself simply were not designed to cost-effectively store, manipulate, and analyze. After all, you don’t “query” an image file; you run algorithms against it to perform the equivalent of feature extraction. Once you’ve done that, you persist the results of your analysis in Hadoop itself or (via SAS Data Loader for Hadoop) to a traditional (R)DBMS platform.

“SAS Data Loader for Hadoop provides business analysts and data scientists the ability to load data from relational databases and SAS data sets to and from Hadoop without writing MapReduce code,” Dull explains. “Users can then transform, cleanse, or load that data in-memory [e.g., in SAS LASR] for visualization or use it for other analytic processes.”

Leveraging What’s New and Different

The availability of a big data platform such as Hadoop radically disrupts the data management status quo. For the most part, SAS’s Magne argues, enterprise data integration vendors have attempted to co-opt or redirect this disruption by retooling their existing products—the plumbing of which is, in some cases, two decades old—to run with or in Hadoop. Like their predecessors, the Hadoop-ready versions of these tools are designed for information technologists: they’re hard to use, have their own programming/scripting languages, and offer next to nothing in the way of self-service amenities.

For established data integration players, the priority isn’t to use Hadoop’s strengths—it’s to safeguard lucrative market segments and revenue streams, including those of their professional services teams, business partners, integrators, and certified consultants.

Meanwhile, in the teeming Hadoop ecosystem, several start-up vendors market Hadoop-based “ETL” tools that help to automate data profiling, preparation, and extraction in Hadoop. These tools typically have extremely limited support for traditional (R) DBMS platforms, however. More to the point, Magne contends, they lack robust data integration and data quality features.

“SAS Data Loader for Hadoop is one of the few purpose-built tools that is designed to permit business users and data scientists to manage big data on Hadoop,” he says.
Self-servicing business users and data scientists is key. According to Magne, data preparation is a time-consuming, exhaustive process that entails considerable back-and-forth between and among business analysts, data scientists, and IT workers. Experts reckon that data scientists, for example, squander the overwhelming majority of their time identifying, acquiring, and preparing data for analysis. SAS Data Loader exposes a data discovery and profiling environment. It supports point-and-click data selection and drag-and-drop data movement to and from Hadoop. In other words, says Magne, it empowers business users and data scientists to serve themselves and have a hand in provisioning data. From IT’s perspective, that’s a two-pronged win-win, he argues.

Conclusion

Hadoop is a powerful data processing and analytic platform that complements and extends traditional decision support practices. In addition, Hadoop promotes new—and hitherto impracticable—applications and use cases. It is genuinely disruptive.

Hadoop is not, however, a technological silver bullet. Buying into Hadoop isn’t going to suddenly supercharge your analytics. Business value isn’t an abstract thing: there’s no innate business value in using a platform such as Hadoop to manage non-traditional types of data—e.g., e-mail messages; image, video, and audio files; social media data; blogs; and call center transcripts, among others.

What you’re doing with Hadoop and big data must, instead, be linked explicitly to business value—i.e., to tangible business problems. “If you’re not using big data to improve your business (e.g., revenues, profits, operational efficiencies, decision making), then don’t do big data. It’s not worth the time, money, or hassle,” Dull says.

“Many big data projects failed early on because these projects weren’t focused on addressing or solving a particular business problem. They were more of a tire-kicking exercise for teams to get familiar with this new technology. For a big data project to succeed, it must be focused on solving a business problem.”

Ironically enough, as organizations grapple with the challenge of wresting value out of their big data investments, they’re finding that they don’t have a good enough handle on their “small” (or transactional) data, Dull points out. “They question whether they should hold off on integrating big data into their ecosystem—and even more notable, whether they need big data at all to move their company forward,” she says, noting that this is a false choice.

If you’re not using big data to improve your business, then don’t do big data.

A technology such as SAS Data Loader for Hadoop, combined with SAS data quality and data integration...
technology, permits an organization to manage the complete data management life cycle.

The key is to approach big data quality and big data governance iteratively, pragmatically, and—above all—patiently, Dull concludes. “Data quality is part of the larger data governance discussion, and our recommendation from the start has been to tackle data governance one big data project at a time. Determine which components of data governance (i.e., decision-making bodies, workflow, decision rights, data quality, and rules of engagement) make sense for each big data project,” she says. “Not all governance components will apply and that’s okay. After you have a few big data projects under your belt, you should have a solid data governance framework that supports all data, big and small. You don’t want to develop a data governance or data quality initiative just for big data. You want to extend your existing initiatives to include big data.”
www.sas.com

Data is more than facts and figures. It’s the lifeblood of your business. That’s why success depends on getting to relevant data quicker and making strategic decisions faster.

The leader in business analytics software and services, SAS helps customers at more than 75,000 sites use their data to identify what’s working, fix what isn’t, and discover new opportunities—faster than ever before.

Plus, it doesn’t matter how big, small, or complex your data is—or how big or small your business is. SAS provides everything you need to derive valuable insights.

SAS offers an integrated, flexible presentation layer for the full breadth of analytics capabilities: statistics, predictive analytics, data and text mining, forecasting, and optimization. You can use diverse data, evaluate more scenarios, and run more models to answer questions you never even had time to ask.

In addition, we provide the best in self-service data management. Not a data scientist? Not a problem. Our latest offering, SAS® Data Loader for Hadoop, provides an intuitive user interface that enables business users to prepare, integrate, and cleanse big data—without the need for writing code or asking for help from IT.

SAS gives you THE POWER TO KNOW®.

- Learn about SAS Data Loader for Hadoop
- Webinar: Self-Service Big Data Preparation for Business Users

© 2015 by TDWI, a division of 1105 Media, Inc. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. E-mail requests or feedback to info@tdwi.org.

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies.