CITO Research

Advancing the craft of technology leadership

Hadoop: Data Storage Locker or Agile Analytics Platform? It's Up to You.

SPONSORED BY

CONTENTS

Introduction 1

How Hadoop Becomes a Data Storage Locker 1

Transforming Hadoop into an Agile Analytics Platform 4

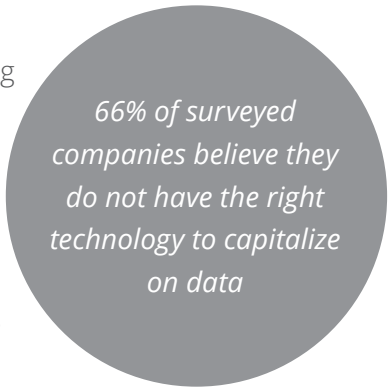
Investment in Data Transformation for Hadoop Delivers 10x Productivity Gains 8

Conclusion 10

Introduction

Why is Hadoop so enticing to businesses? As an open source repository, Hadoop is a cutting edge and disruptive technology. It has the capacity to handle quantities of data that traditional repositories simply cannot, and its storage is drastically cheaper than traditional data warehouses. These factors contribute to extremely high expectations from business users. Companies expect a return on their investment from Hadoop in the range of three to four dollars for every dollar invested.

Yet, reality is starkly different. At present, Hadoop users are achieving a return of 55 cents per dollar invested. This is a tenuous situation for businesses, as the flow of big data will not slow down in order for them to learn how to better utilize Hadoop. The problem is only going to be compounded in the next decade: the amount of big data is doubling every two years and is estimated to grow from 4.4 ZB in 2014 to 44 ZB in 2020. That's as many pieces of data as there are stars in the universe.[1]



Despite this opportunity, a Bain and Company study found that 66% of surveyed companies believe they do not have the right technology to capitalize on data.[2] Consequently, even companies that recognize how powerful data could be do not have the knowledge, expertise, or tools to make that vision a reality.

Hadoop lowers the barrier to storing data, but it doesn't necessarily lower the barriers to creating value from data. This CITO Research paper will describe what's required to turn Hadoop into a productive platform for agile analytics.

How Hadoop Becomes a Data Storage Locker

Hadoop's economics are transformational. The cost per gigabyte of data makes Hadoop an attractive data storage solution for many different applications and types of data.

On its own, Hadoop can't parse the meaning of the data it is collecting. And while Hadoop can amass multi-structured data, it is not designed to transform it or help companies decide whether or not that data is useful.

The result is that frequently, Hadoop gathers data that lies fallow. Sometimes this amassing of data is tellingly referred to as "Hadumping."

[1] http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[2] http://www.bain.com/Images/BAIN%20_BRIEF_The_value_of_Big_Data.pdf

Store Now, Understand Later

The cross-that-bridge-when-we-come-to-it strategy of landing data in Hadoop and figuring out what to do with it later has led to Hadoop functioning more like a data storage locker than an agile analytics platform.

One of Hadoop's strengths—the fact that it doesn't need a predefined schema to load data into it—also feeds into its weakness. For meaning to be applied to data when it is read (referred to as schema-on-read), users need to understand the data and then add some context to transform raw data into insights.

Data has the potential to be valuable, but companies need tools to explore and extract that value. In truth, this is not a new problem: Even in the world of traditional data warehousing with structured data repositories and rigid, top-down governance, 90% of business data went unused. Obviously, with ever more data on hand, and Hadoop to store it, more data is going unused than ever before.

Variety: The V You Need to Worry About

Big data is often framed in terms of the three Vs: volume, variety, and velocity. CITO Research believes that variety is the most problematic of the three Vs. Here's why.

For schema on read to work, that is, to apply understanding to data stored in Hadoop, someone has to understand what the data means. For each dataset, this application of meaning must happen again. It doesn't matter how large the dataset is; what's really both a problem and an opportunity is how many datasets are coming in. The meaning of each of these datasets must be specified. In addition to evaluating the meaning of each dataset coming in, determining the relationships between those datasets is often the critical breakthrough point to uncovering business value in the data.

> That understanding means the challenge is not volume (once you know what the data means, you can read it) but variety (figuring out what the data means).

A VP of Data at a marketing data provider echoed this sentiment. "There are a great variety of sources and all sizes and shapes and flavors of the data, and we have to understand them up front. We can't process them and then decide whether they are relevant. A lot of pre-analysis happens with the data before we even accept it for modeling," she said.

Datasets are coming in at high velocity, but knowing what to do with them requires dealing first with what those data sources mean. The variety of data is problematic because the question of meaning must be answered each time a new type of data appears. If data is stored first and understood later, the pile of data to deal with at some future point only gets larger.

Expertise Is Scarce

To date, self-service access to Hadoop has been more dream than reality. The current framework for assigning meaning to data in Hadoop requires analysts to rely on development experts for their workflows. This reliance on Hadoop experts bogs down processes, creates bottlenecks, and makes it difficult for people who can supply the needed business context for the data—in other words, who have native understanding of the data from a business perspective—to directly explore and interact with the data.

The need to transform data into usable forms is so acute that the fastest growing category of specialist is now the data engineer, not the data scientist. As of September 2014, LinkedIn had nearly 21,000 postings for jobs with "data engineer" in the title, compared to just over 11,000 for jobs with "data scientist" in the title.

Data Preparation Is Time-Consuming

Companies are forced to devote far too much of their time to preparing data, often repeating steps without business context. These tasks include the type of wrangling, munging, and hand-coding exercises that devour time, whether that involves joins of disparate datasets or just getting all the data into the same format.

Here is a sampling of a few common data preparation problems:

- **Problems stemming from business logic.** For example, "price" might include taxes and shipping in some data sources but not in others.

- **Missing values and outliers.** When missing values or outliers (such as latitude and longitude in the middle of the ocean) show up, what should the person working with the data do? Should the rest of the data for those records be included in the models or should data records with missing values be omitted entirely? The answer can be highly specific to the use case for the data.

- **Derived values.** The data may contain answers, but it may take work to get at those answers. Consider the task of figuring out how long a user spent on a website, which requires sessionizing data from weblogs to ascertain the activities of a particular user. The definition of a session is a derived value calculated by defining the starting time and ending time. Finding a business definition of a session is an inexact science. It requires some experimentation and observation of user behaviors to determine the session length appropriate to analyze for the business questions being asked. This iterative process, when executed by a non-business expert, is defined by a lot of extra trial and error.

Data preparation is notoriously time consuming; data scientists say that these types of activities consume some 50 to 80% of their time.[3]

Transforming Hadoop into an Agile Analytics Platform

When companies consider what kinds of platforms to adopt to get the most out of their data and their Hadoop implementation, they should focus on the following factors to achieve long-term success.

Agile versus Waterfall

What does it mean for analytics to be agile? It means that you need a workflow that is iterative and dynamic and allows users to discover insights naturally (see Figure 1).

Consider the waterfall methodology, in which the expected output was a report. The report was designed to answer to a question or a group of questions, defined in advance. The output of the analysis was often a static KPI or single chart for storytelling. By making that starting assumption, much data is thrown out immediately to drive to a single answer. This simplifies data management, but has the downside of removing data from the analysis that might highlight an unexpected insight that resides in that dataset.

Agile analytics gives you the ability to explore, try new things, and then change your mind. It's an exploratory test-and-learn approach in which questions lead to more questions and then eventually to discovered answers. The nature of agile analytics is iterative. The following scenarios demonstrate the need for agility.

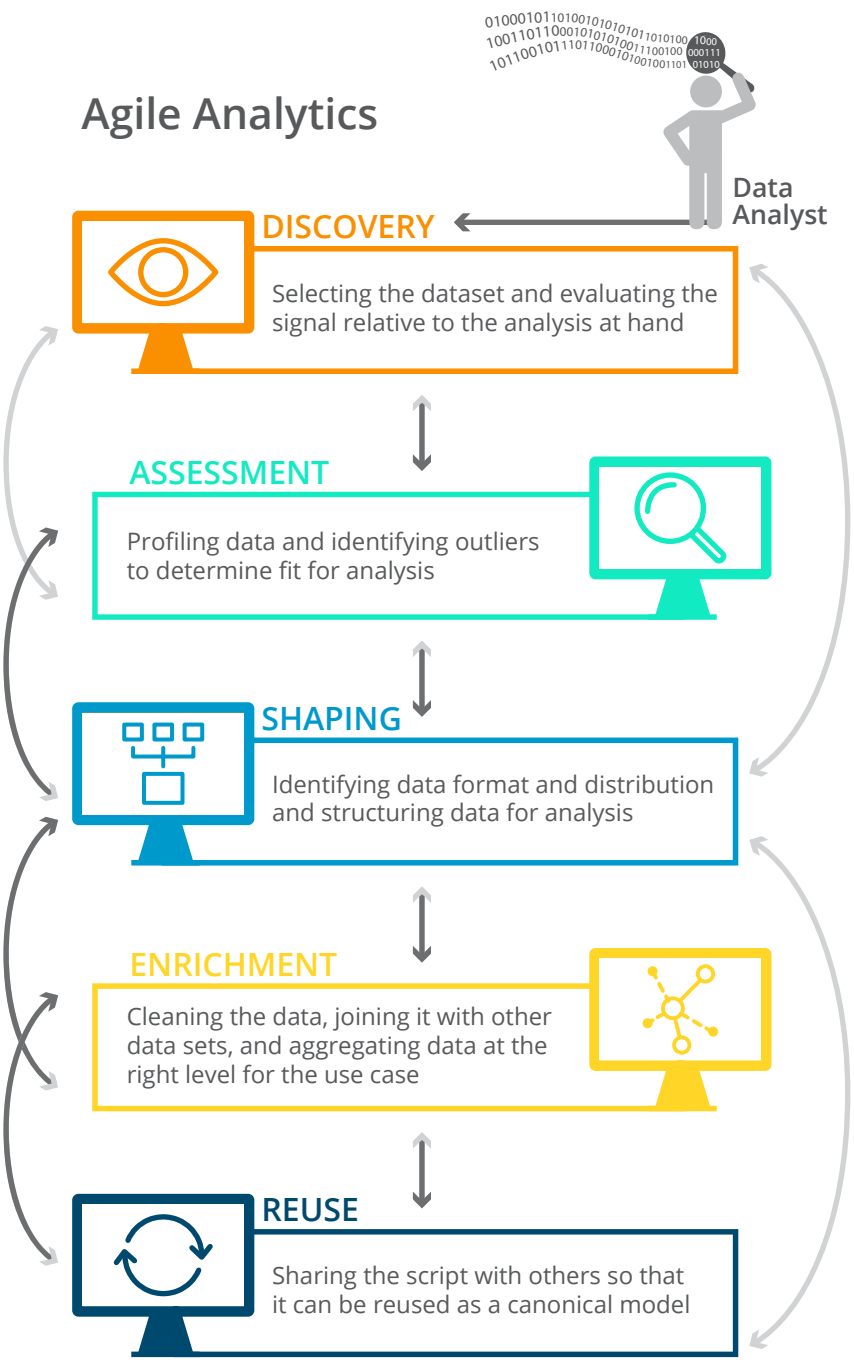
Same dataset, different stakeholders. Different stakeholders frequently use the same data in different ways. Consider logs of usage data from a personal fitness bracelet. Product development wants to correlate log data with support tickets. They wonder which features cause users to contact support and whether the product design could be tweaked to make it more intuitive.

Marketing looks at the same logs from a completely different angle. The marketing department might be more interested in the correlation between application usage trends, customer demographics, and engagement on the product forums.

Each group consults the product usage logs, but uses them in entirely different ways. Further, their use of the data will evolve over time across groups.

Agile analytics gives you the ability to explore, try new things, and then change your mind

[3] http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0



Retransform the data to find new signals. Suppose product development found out that female customers use the bracelet's sleep monitor feature more than men do. They would like to design a product geared for women and want the data transformed with finer grained temporal information. What days of the week do women use the feature most? Workdays? Weekends? This information is in the raw logs, but it was transformed away during the first iteration. Agility demands the ability to go back to raw data and retransform it to support additional use cases.

These scenarios cover just one data source and only two lines of business. If you multiply those needs by the variety of sources of big data and the number of stakeholders who want to make data-driven decisions, it's easy to see why data transformation work dominates the time of data engineers and data scientists.

Figure 1. Agile analytics is an iterative and flexible process that supports changing business needs and requirements

Leveraging Business Context

Human interaction is vital to the process of preparing data for analysis. Users need to identify features in the data that are interesting to them and provide feedback on the representation and meaning of the data they're attempting to use. Otherwise, the data is never tied to organizational change or corrective actions.

A leader in the design software industry recently said in a CITO Research interview, "Everybody at some level needs to be a data analyst. Anybody who's working at the company who is trying to improve internal processes, external processes, customer behavior, has to have willingness to ask questions and work from the context of data. I've been advocating that each product team have at least one or two people who are the mavens of their team's data." In other words, business experts supply business context, imbuing the data with domain knowledge. Even if more people are hired and tasked with data transformation, if they don't have the domain knowledge of their data, their work on behalf of business stakeholders will be less efficient than a method that empowers domain experts to transform data themselves.

Democratizing Data Access

Companies looking to get the most out of Hadoop must overcome the shortage of Hadoop experts. They need to tap into the broader community of analysts that already exist in their organizations and implement platforms that bridge the divide between the business user and the data. These platforms will not replace human participation in the data extraction process. Quite the contrary, in fact.

The more employees who can use big data, the more powerful it can become. Users in marketing and operations may ask very different questions that together yield new and powerful insights.

The more employees who can use big data, the more powerful it can become

Adopting new technology can be an intimidating process for employees at all levels of an organization. Users are accustomed to the way existing tools work and frequently develop their own workarounds to address limitations. But with Hadoop, without new tools specifically designed for the vast quantities of data from varying sources, users only see a limited picture of the insights their data could be providing to them.

The best add-on platforms for Hadoop allow for a democratization of data that ignites data-driven decision-making across an entire organization. Data is no longer strictly in the province of the data scientist; business users can find value from interacting with data themselves. Users can start with smaller datasets and once they see that the data transformation platform gives them the results they need, their trust builds and they can move to larger datasets. This becomes a virtuous feedback cycle; giving more people useful access to data drives demand for more data. Suddenly, data becomes integral to the company and insights come from unexpected places.

Using Machine Learning

The only way to provide skill-neutral access to data is to adopt platforms on top of Hadoop that utilize an interactive interface, rather than code, to simplify the complexity of the data investigation process. These easy-to-use interfaces allow all users to see how the data will be transformed, and make data easier to read and understand. Point and click tools make even the trickiest data as simple to manipulate as perusing a spreadsheet.

Beneath that intuitive exterior, however, these revolutionary data transformation tools are using a predictive interaction approach that leverages machine learning to anticipate users' needs and speed them through the process of understanding and manipulating their data. By anticipating what users need, the platform allows people to scale their abilities rapidly. As one user stated, "It's the symbiosis between the machine and human that really makes the analyst feel superhuman" in handling big data velocity and variety.

Predictive interaction interposes machine learning techniques between human users and the data that they see. By browsing the data, a user's behavior effectively teaches the machine what to find in a given dataset. The machine builds its knowledge with every piece of feedback. Users guide the process, but the machine does the detailed work, meaning experts are no longer spending 80% of their time preparing the data. They're able to get up to their elbows in the data much faster, regardless of the amount of data they're looking to analyze.

Users guide the process, but the machine does the detailed work, meaning experts are no longer spending 80% of their time preparing the data

It's About Repeatable Speed and Scale

Hadoop requires tools that are fit for schema on read. Companies often do not budget for these add-ons. Without the add-ons, they cannot get the cohesive data manipulation that drives true analytic capacity from the data they store in Hadoop.

Once Hadoop adopters recognize their need, many opt for makeshift solutions. They either try writing scripts by hand to transform data (not a scalable solution) or try to retrofit existing tools that were designed for traditional data warehouses and traditional monolithic data structures.

Another approach is to rely heavily on services. While the initial investment in Hadoop may run in the tens of thousands, it may take a couple of million dollars in services to make it useful. Opting to use a data transformation platform not only saves money initially but also gives the organization a repeatable approach to agile analytics. Repeatability is key. As the VP of Data from a data distribution company told CITO Research, "We are in the business of data transformation and data management, so of course everything we do has to be repeatable."

Investment in Data Transformation for Hadoop Delivers 10x Productivity Gains

Businesses that have used data transformation tools have been astonished at how going from a dozen people using data to a few hundred has impacted the decision making process. The returns could be even greater if thousands within a business were using data.

CITO Research has found that data transformation platforms like Trifacta offer productivity gains with a factor of 10 for data scientists and data engineers as well as business users.

With tools like Trifacta, no coding is required. Users:

- Have an interface that guides them in transforming data
- Get immediate visual feedback
- Can detect any problems with the output
- Obtain a sharable and repeatable history of steps taken to transform raw data into analysis-ready data

Perhaps most importantly, Trifacta provides greater visibility into datasets, ensuring wider use of big data.

Trifacta is a prime example of how effective tooling can be for non-expert users. Using a Google-like "auto-complete" approach, users guide Trifacta through a predictive interaction process in which a portion of their big data is visualized to see whether it is correct. The tool can predictively highlight information the user will find relevant. The visualization aspect of a tool is crucial, but so too is the ease of use with which it incorporates user feedback and learns from these inputs.

With data transformation tools in place, the job satisfaction and productivity of not only analysts, but also data scientists and data engineers will increase. They will no longer have to spend the bulk of their time wrangling data for others' use. Instead, with data transformation tools like Trifacta, data scientists and data engineers will be able to collaborate more efficiently with the business by sharing the same toolset. Business analysts can participate in transforming data in Hadoop, collaborating more efficiently with data scientists and experts, rendering it accessible via end-user BI tools like Tableau and QlikView.

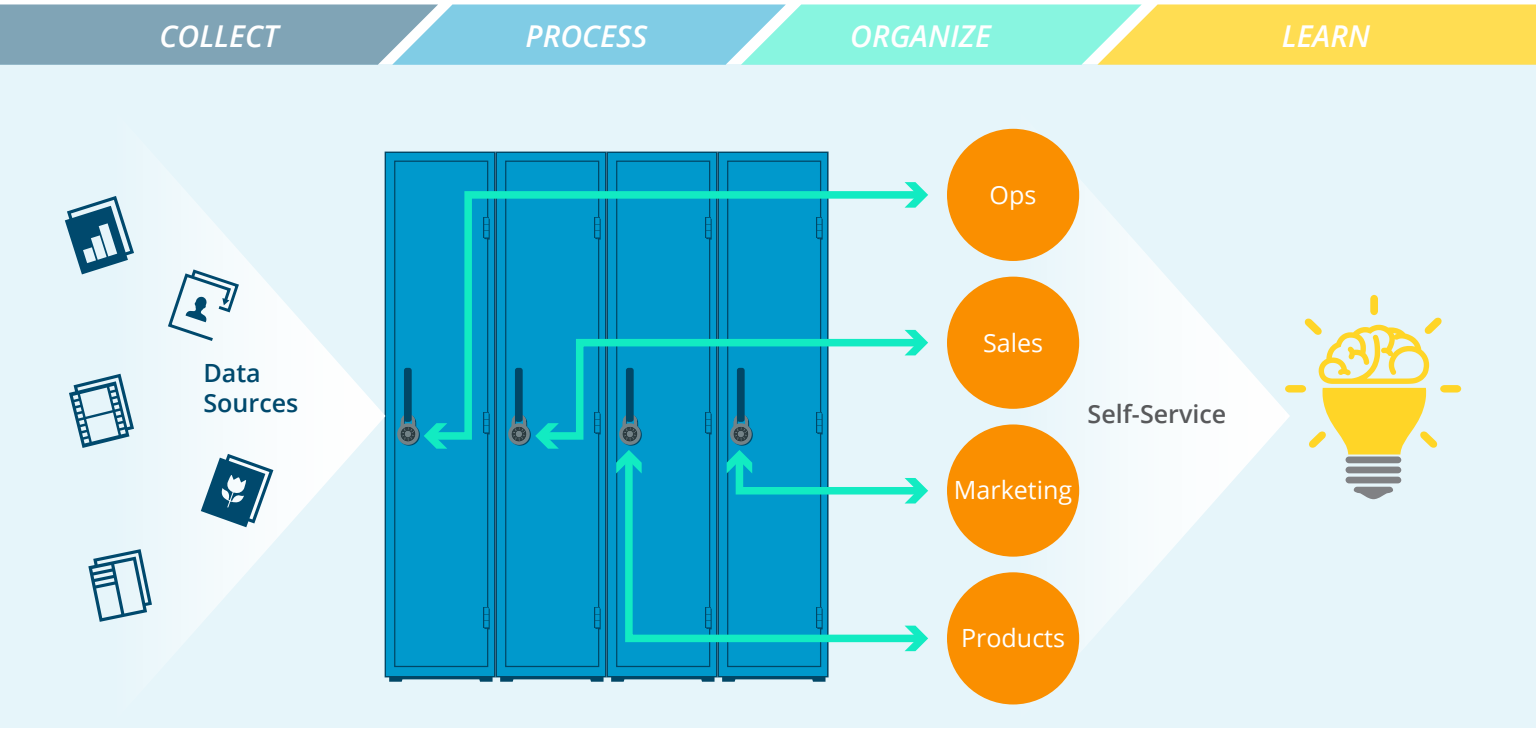


Figure 2. With data transformation tools such as Trifacta, Hadoop becomes the basis for agile analytics, actively used by all lines of business.

Conclusion

Unquestionably, Hadoop has the potential to change the way companies work with data. Hadoop allows businesses to store, collect, and process the ever-increasing quantities of big data that are being generated every second – data from customers, from business partners, from devices, in all formats, whether that data is in tabular, textual, or other machine-generated formats. Yet, before jumping off the diving board into the Hadoop deep end, companies should recognize that the advantages of Hadoop are only fully realized with additional tooling.

Hadoop is not a data transformation product. It is a data repository. It expects schema on read. Someone must supply that schema in order for the data to be usable. By enabling business users to provide that schema interactively, without having to understand even so much as that term, Trifacta brings about data transformation and frees the data in Hadoop for wide business use. CITO Research believes that without such tools, Hadoop will remain a Hadumping ground, where more and more data is stored and gathers dust. Trifacta, relying on machine learning and user input, liberates Hadoop data. Regardless of the origin of the data, regardless of its format, Trifacta enables the process of standardizing data stored in Hadoop and making it ready to use far more efficient.

When it comes to data preparation and transformation, Trifacta is a prime example of how effective new platforms can be in accelerating Hadoop's time to value. Data preparation has been a huge barrier for early adopters of Hadoop, meaning the time for actual analytics and using data to inform decisions has been limited. Trifacta's ability to balance machine learning and human input for productive data transformation can change that.

Learn more about Trifacta ▸

This paper was created by CITO Research and sponsored by Trifacta

CITO Research

CITO Research is a source of news, analysis, research and knowledge for CIOs, CTOs and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com