THE HORSEPOWER OF HADOOP: FAST AND FLEXIBLE INSIGHT WITH RESULTS

May 2016

➔ **Michael Lock,** Vice President & Principal Analyst, Analytics & Business Intelligence

Report Highlights

As a foundation for creating and delivering business insight, a company's data infrastructure is a vital organ impacting the health of the business. In order to exploit the diversity of data available and modernize their data architecture, many organizations explore a Hadoop-based data environment for its flexibility and scalability in managing big data. This report investigates the impact of Hadoop on the data, people, and performance of today's companies.

ABERDEEN GROUP

2

Those using Hadoop, or kicking its tires, are most driven to action out of a need to make better use of unstructured and semi-structured data.

Definitions

Known by its iconic yellow elephant logo, **Hadoop®** is an open source project from the Apache ™ Software Foundation (ASF). Allowing for distributed processing of large data sets of multiple disparate varieties, Hadoop is designed for scalability and flexibility in managing Big Data. The core modules in a Hadoop environment include:

Hadoop Distributed File System (HDFS™): Provides high-throughput access to multiple types of application data

Hadoop YARN: A framework for job scheduling and cluster resource management

Hadoop MapReduce: A YARN-based system for parallel processing of large data sets

Big Data for a Little Elephant

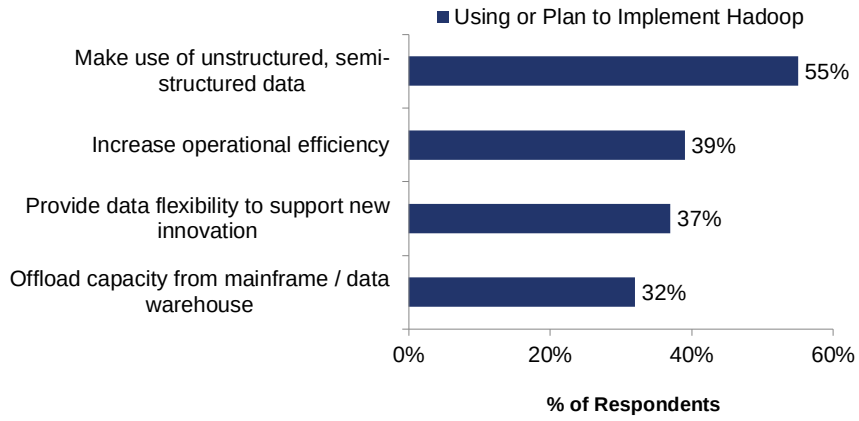
Big data used to be a big problem. As more of our working lives moved to the digital realm, the typical organization was flooded with information that needed to be captured, moved, stored, and archived. Moreover, these were merely baseline activities that fell in line with best practices or compliance with Sarbanes-Oxley and other regulations. Most companies didn't have the time or capability to exploit that data beyond a few static reports.

The past several years, however, have seen a marked change in attitude. Companies find insight hidden in data beyond just traditional structured, transactional information. Emails, social data, rich media, geospatial information, and sensor-generated data (among others) have emerged, along with an evolved suite of technologies to help manage and utilize that information. This has given way to a shift in attitude toward big data from problem to opportunity.

At the technology forefront of this opportunity is the data architecture known as Hadoop (see sidebar). In-line with the trend toward big data as an opportunity, Aberdeen Group's research demonstrates that those using Hadoop, or kicking its tires, are most driven to action out of a need to make better use of unstructured and semi-structured data (Figure 1).

3

Figure 1: Top Drivers of Hadoop Implementation



n = 42, Source: Aberdeen Group, May 2016

Another major trend is the shift in mindset among a wider variety of organizations toward viewing data as a strategic asset and a critical part of their core business — both operationally and strategically. First, regardless of the industry sector or long-term focus of the company, the use of data has become an important part of boosting operational efficiency. Secondly, and from a more strategic perspective, flexibility in the data environment is now seen as vital to a company's ability to innovate, develop new products, and bring them to market more efficiently. Unstructured data from social channels enables companies to tap into customer sentiment and get ahead of any product issues. Using geospatial information to track customer behavior enables companies to target customers better, message to them more effectively, and produce long-term revenue gains. Companies look to a Hadoop-based infrastructure to help support that efficiency, both tactically and strategically.

Masking Complexity, Empowering Users

Any technology deployment, regardless of size or technical sophistication, needs to be viewed through the lens of those using it or indirectly impacted by it. The growth in data disparity

Fast Facts

Companies using unstructured data on a frequent basis are:

- **53%** more likely to have policies and / or tools in place for governing / controlling end-user data access
- **55%** more likely to formally train and develop analytical talent in-house
- **66%** more likely to have an executive-level sponsor or champion for analytics

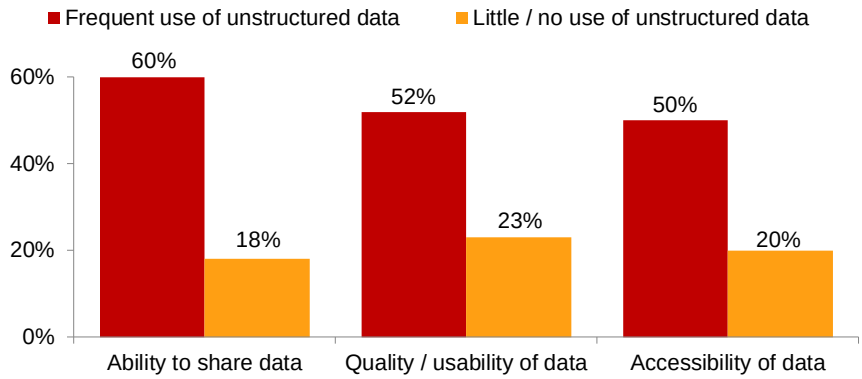
4

- Related Research, "Running Lean Analytics with a Cloud or Hybrid Approach"
- Related Research, "The Visual Edge: Interactive Discovery vs. Traditional BI"

and complexity certainly carries opportunity for game-changing insights to be extracted from its analysis, but only for organizations that have this user focus. A marketing director doesn't think in terms of star schemas or node clustering, but probably understands the financial impact of social media clout scores or location-based insights. Traditional structured data from business applications typically amounts to the bulk of data consumed by most business users, but there is an argument to be made that unstructured data fits their minds just as well – if not better.

In addition to the bevy of use cases for unstructured data, there is another less obvious benefit that impacts the internal organization. Aberdeen's research shows that companies using unstructured data frequently experienced a higher degree of user satisfaction with several key aspects of the data environment (Figure 2).

Figure 2: User Satisfaction with Unstructured Data



% of Respondents "Satisfied" or "Very Satisfied" with above

n = 65, Source: Aberdeen Group, May 2016

If the analytical process and the generation of insight is to succeed, companies need to think closely about factors like this. When searching for answers (or new questions to ask), users need broader access to data, quality and trustworthy

information, and to be able to socialize and share insights with others in the organization.

This need for user empowerment and user satisfaction is largely why companies explore Hadoop today. Among other factors, Hadoop-based data architectures allow for two key advantages related to the end-users themselves:

- **Data flexibility.** The architecture of Hadoop is purpose-built to house traditional structured data alongside text-based unstructured data, geospatial information, rich media files, and a variety of other data types. This gives users access to the data they need, all under one roof, and also allows for the centralized management and oversight of that data to help ensure higher quality and usability.
- **Data elasticity.** The business landscape is changing fast. Strategies change, metrics change, analyses change, and ultimately, people start asking different questions of the business. With a rigid and inflexible architecture for modelling data, the ability to adapt to changes in the business and modify the analytical process is compromised. Another benefit of Hadoop is its elasticity with respect to the modelling and expansion of the data. The architecture is significantly more flexible in allowing for the addition of new data sources and system nodes (clustering), as well as modifications to the underlying schema or data model. Known as data elasticity, this has the ultimate effect of empowering users with the ability to evolve their analytical strategy and their approach to understanding the business.

The connection between Hadoop and the ultimate end-user is highlighted in Figure 2 and underscores why companies are

The architecture is significantly more flexible in allowing for the addition of new data sources and system nodes (clustering), as well as modifications to the underlying schema or data model.

Key Terms

Data Lake: Supported by a Hadoop-based technology environment, a data lake refers to a repository of disparate data comingled together in their native formats, and including all relevant data attributes.

Apache Spark: An open source cluster computing framework designed for low-latency querying of large data sets. Spark is typically used in applications requiring real-time processing of advanced analytical algorithms and techniques.

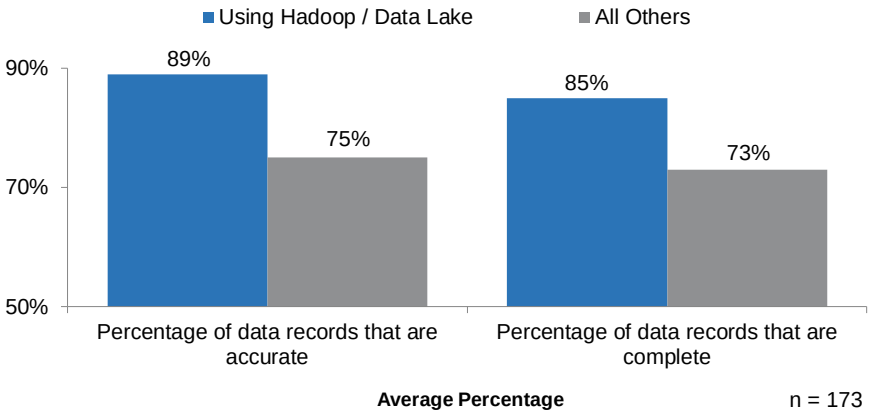
investigating Hadoop. Its ability to handle unstructured data alongside other types of information, along with its scalability and elasticity, makes it a powerful data environment for the evolved business world.

Paving the Way for Hadoop

From the user-facing side to the data-facing side, companies using or exploring Hadoop understand the need for technology and organizational maturity. That maturity starts with the right tools and processes in place to lay the foundation for Hadoop. Aberdeen's recent report, Data Preparation and the Evolution of Analytics, demonstrated how managing and improving the quality of data can lead to enhanced decisions and superior business performance as well.

Data preparation can include technologies or processes for data cleansing, profiling, enrichment, de-duplication, or a variety of other features. The research shows that Hadoop users are almost twice as likely than all other companies to have data preparation capabilities in place. These companies are also more likely to have policies and procedures in place for governing and overseeing their data environments. These aspects of maturity are instrumental in helping generate more complete and accurate data (Figure 3).

Figure 3: Enhanced Data Maturity Drives Quality



Source: Aberdeen Group, May 2016

Additionally, Hadoop users have a more mature portfolio of technologies in use to help improve the decision process and take advantage of their Hadoop-based infrastructure. According to the research, Hadoop users are:

- 3.3 times more likely to use real-time data integration technology
- 4.9 times more likely to have a high-speed/high-performance data warehouse in place
- 4.6 times more likely to use interactive data visualization
- 6.5 times more likely to use predictive analytics

Combining the breadth and sophistication of technology usage that Hadoop users have, with potential benefits to the end-user described before, it becomes easier to see how these companies can experience elevated performance. Users have more centralized access to a cleaner, more complete, and more consumable foundation of data, thus reducing wasted time in the analytical process. The flexibility of the data environment

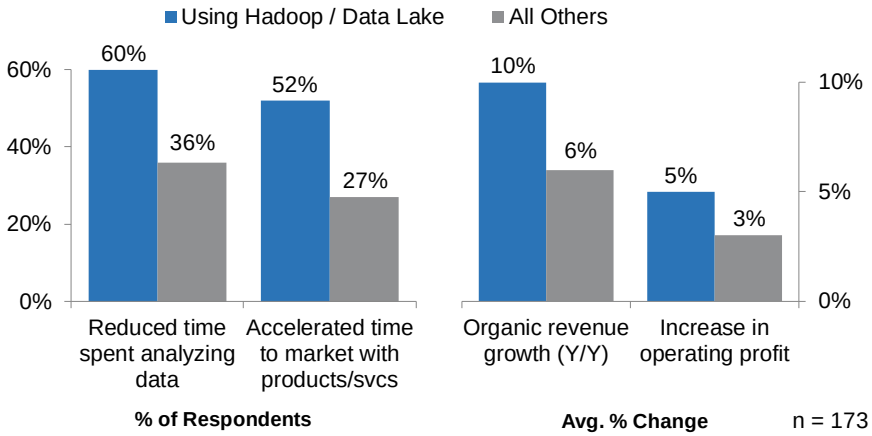
Fast Facts

Hadoop users have an enhanced focus on developing capabilities for **data preparation**. In comparison to other companies, Hadoop users are:

- **66%** more likely to have a standalone data preparation solution in place
- **43%** more likely to have data preparation capabilities integrated as part of a data discovery / visual analytics platform
- **3.3x** more likely to have data preparation capabilities integrated as part of a data science / advanced analytics platform

and the breadth of information at their disposal also contribute to a company's ability to develop products and services quicker and improve time to market. Drawing closer to tangible business performance, Hadoop users were able to improve these key metrics (Figure 4).

Figure 4: Decision Efficiency and Business Execution



Source: Aberdeen Group, May 2016

Done right, business analytics follow an effective process from raw data to business action. Companies that can improve efficiency in the data environment by creating accessibility, flexibility, and speed, are in an advantageous position to transform their data into usable business insight, and put it into the hands of users more quickly. Ultimately, these users can put that insight to work identifying and acting on opportunities for business expansion and growth, as well as operational efficiency. Companies using Hadoop were able to exploit their sophisticated data environments to generate more effective decisions and deliver tangible results in key areas like revenue growth and operating profit, as well.

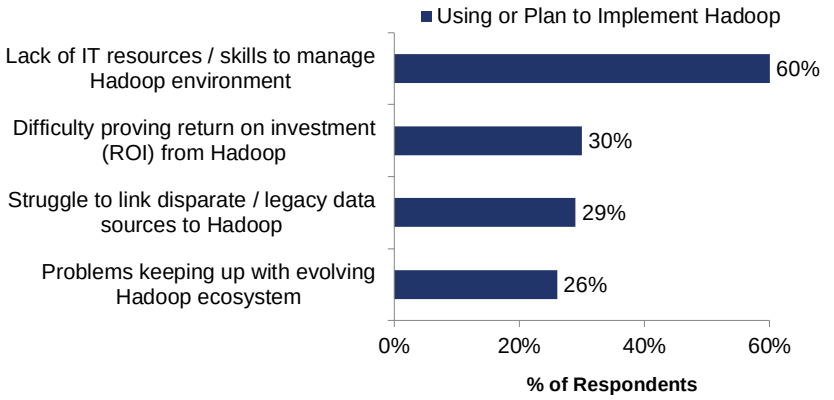
Navigating the Challenges

Some would argue that Hadoop adoption has been slower than the hype would suggest. Certainly there are companies out there for which Hadoop simply wouldn't be a good fit, particularly smaller organizations with less complexity or urgency in their data environments. However, there are tens of thousands of companies in the U.S. alone that could benefit from the efficient use of Hadoop. The fact is though, not many companies have the resources or technical firepower to deploy and manage a Hadoop environment solely in-house, regardless of size. This lack of resources and internal skill sets is in fact the most commonly cited challenge of Hadoop (Figure 5).

➔ Related Research, "Data Preparation and the Evolution of Analytics"

➔ Related Research, "A BI Blessing from Above: The Impact of the Executive Touch"

Figure 5: Key Roadblocks to Implementing Hadoop



n = 42, Source: Aberdeen Group, May 2016

At a distant second, but still important, is the difficulty that companies find in proving a tangible ROI from the implementation of this type of architecture. Not for the technically feint-of-heart, the complexities of Hadoop are typically managed by seasoned IT and data professionals, oftentimes significantly far removed from the executive-level budget holders. Convincing a CFO to invest in the adjacent technologies and support required to take full advantage of

10

The real ROI from Hadoop implementation goes far beyond just the cost savings on software licenses or server hardware. Like any other technology deployment, true ROI comes from adoption, engagement, and business results.

Hadoop would entail more than just a back-of-the-envelope business case. Moreover, the real ROI from Hadoop implementation goes far beyond just the cost savings on software licenses or server hardware. Like any other technology deployment, true ROI comes from adoption, engagement, and business results. Companies considering a Hadoop deployment have difficulty making this case to the powers that be.

Other challenges that make companies think twice about Hadoop deployment have to do with complexity. First, there is the complexity of their own data environments, as linking Hadoop to a legacy mainframe or other antiquated server hardware can be a struggle for companies not heavily staffed in IT. Second, there is complexity related to the Hadoop environment itself. As an open source platform, Hadoop is constantly evolving and taking on new revisions and extensions. Many companies are hesitant to rebuild their data infrastructure to fit a certain version of Hadoop out of concern for it changing multiple times, requiring costly and time-consuming reconfigurations of the architecture.

At the end of the day, these very challenges drive a growing and robust sector of Hadoop support services, from implementation, to ongoing maintenance and upgrade management. While the open-source Hadoop software is available for free, most companies lack the strength, sophistication, and sheer manpower in their IT organization to tackle all these challenges on their own, opting to engage a third-party support organization instead.

Key Takeaways

As mentioned, Hadoop isn't necessarily a great fit for every company. On the other hand, there are a multitude of companies out there that suffer from the type of data complexity and user

urgency that an effective Hadoop environment can address. With its scalability and flexibility, Hadoop offers a path toward excellence in the data environment, empowered end-users, and legitimate business results. For organizations contemplating a Hadoop-based data environment, the following key research takeaways should be taken into consideration:

- **There is power in data flexibility.** Because of the native functionality of Hadoop, allowing for a data lake architecture with commingled disparate data types, companies have the ability to exploit more of their data and make it available to a variety of different users. Masking the complexity of the underlying applications and platforms involved, Hadoop can provide a buffer for non-technical business users to access the data they need to find answers, and raise the analytical bar within their organizations.

- **Hadoop-enabled analytics delivers results.** Effective business analytics is a process of transforming raw data into consumable insight, and taking action based on that insight. At each stage of that process, a well-functioning Hadoop environment can have a major impact. With a back-end data architecture where information is simultaneously plentiful, diverse, and accessible, companies see efficiencies in the exploration and discovery of insight on the front-end. This analytical process accelerates the flow of relevant insight, creates more efficient decisions, and leads to performance benefits like accelerated time-to-market and revenue growth.

- **Help is out there.** Just looking at the core modules associated with Hadoop is enough to make anyone's head spin. To say nothing of the expansive suite of

complementary and adjacent technologies, third-party Hadoop service providers exist to help navigate the breadth of technology offering and help tailor a combined solution and environment appropriate for the company at hand. These services not only help organizations through the process of implementation and ongoing management, but also provide visibility into the direct and indirect impact of Hadoop to the organization. This visibility can be instrumental in estimating, and delivering on, an advantageous ROI from Hadoop deployment.

For more information on this or other research topics, please visit www.aberdeen.com.

Related Research	
Running Lean Analytics with a Cloud or Hybrid Approach; April 2016	*A BI Blessing from Above: The Impact of the Executive Touch*; March 2016
The Visual Edge: Interactive Discovery vs. Traditional BI; April 2016	*Three Levels of ROI from Data Quality Initiatives*; March 2016
Data Preparation and the Evolution of Analytics; March 2016	*BI Excels in the Mid-Market: A Nimbler Version of the Enterprise*; February 2016

Author: Michael Lock, Vice President & Principal Analyst, Analytics & Business Intelligence (michael.lock@aberdeen.com)

13

About Aberdeen Group

Since 1988, Aberdeen Group has published research that helps businesses worldwide improve their performance. Our analysts derive fact-based, vendor-agnostic insights from a proprietary analytical framework, which identifies Best-in-Class organizations from primary research conducted with industry practitioners. The resulting research content is used by hundreds of thousands of business professionals to drive smarter decision-making and improve business strategy. Aberdeen Group is headquartered in Boston, MA.