# MAXIMIZING YOUR DATA LAKE WITH A CLOUD OR HYBRID APPROACH

May 2016

Companies today increasingly look for ways to house multiple disparate forms of data under the same roof, maintaining original integrity and attributes. Enter the Hadoop-based data lake. While a traditional on-premise data lake might address the immediate needs for scalability and flexibility, research suggests that it may fall short in supporting key aspects of the user experience. This Knowledge Brief investigates the impact of a data lake maintained in a cloud or hybrid infrastructure.

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

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

Base. Warehouse. Mart. Store. There is certainly no shortage of metaphor terminology used to describe the way we manage and organize data. Aside from their likeness to aspects of a retail value chain, these technologies have something else in common.

Some are set up for large volumes of data force-fit into a pre-defined schema, and others provide a more summarized view of data and their attributes, but none of them have the ability to store multiple types of data (structured and unstructured) in their native, natural formats. Commonly associated with open-source Hadoop technology, a data lake on the other hand (see sidebar) was conceived for this very purpose.

Aberdeen's recent report, The Horsepower of Hadoop: Fast and Flexible Insight with Results, demonstrated that companies utilizing a Hadoop / data lake environment enjoyed several benefits above and beyond their peers such as accelerated time-to-market, and operational efficiencies. However, the research

also demonstrated how taking this data lake approach in a cloud or hybrid cloud infrastructure drives extra value in the decision process, including:

- **Greater access to data across the company**. Because of the un-tethered nature of cloud-based applications and infrastructure (i.e. not requiring on-premise or VPN access), companies are typically able to boost the accessibility of data in these environments. Through the process of analysis, data that resides in a cloud based application or a cloud-based repository such as a data lake becomes available for analysis. Companies then have the ability to enrich their insights with a broader base of information.

- **Accelerated data gathering and preparation**. On a similar note, a critical step in the analytical process is not just accessing data, but preparing it for analysis. Between data quality issues, latency, and overall inaccessibility, many data professionals complain about the time they spend preparing data for analysis as opposed to generating useful business insight from its analysis. Therefore, it follows logically that a cloud-based environment would not only increase the accessibility of data, but having removed or mitigated a major barrier in the process would also jumpstart the process of data preparation.

- **Enhanced accuracy in the analysis process.** In the same way that data quality and lateness wreak havoc on the ability to create timely insights, these issues also impact the accuracy of decisions. Inaccessible information delays the analytical process, which forces assumptions to be made and decisions to rely more on experience and gut instinct rather than data. By the same token, incorrect or corrupted information fed into analyses

➔ Read the full report, "The Horsepower of Hadoop: Fast and Flexible Insight with Results"

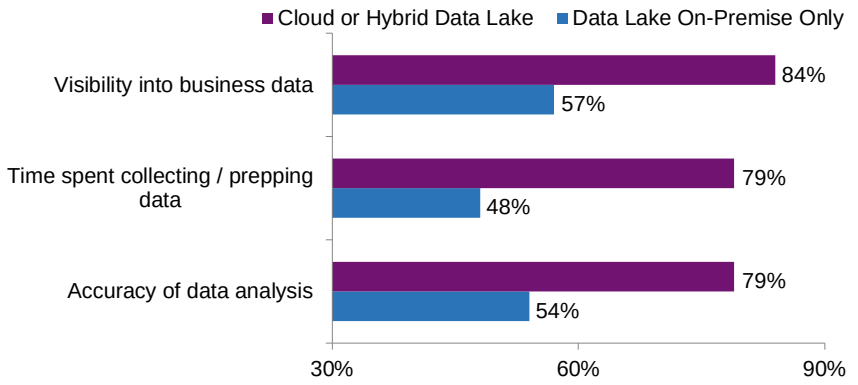
➔ Related Research "The Three Levels of ROI from Data Quality Initiatives"

**Those with a cloud-based data lake are twice as likely to be satisfied with the speed of information delivery, as well as the ease-of-use of their data systems.**

causes inaccuracy in the process and ill-fated decisions as well. The flexibility and accessibility of a cloud or hybrid environment help mitigate some of these issues and improve decision accuracy.

All of these concepts discussed above are critical factors in the analytical process of converting raw data into actionable insight. The research shows that companies with a cloud-based data lake in place were able to deliver improvements in metrics closely tied to these concepts (Figure 1).

**Figure 1: Decision Efficiency with a Cloud-Based Data Lake**




**% of Respondents that saw "improvement" or "substantial improvement"**

n = 71, Source: Aberdeen Group, May 2016

In addition to the improvement depicted above, companies with a cloud-based data lake environment experienced another set of positive side effects. Improving these key aspects of the decision process also brought about improvements in user satisfaction. These companies are twice as likely to be satisfied with the speed of information delivery, as well as the ease-of-use of their data systems.

## Conclusion

The ability to extract tangible business value from a Hadoop-based data environment rest upon far more than just the nature of their deployment and infrastructure. The research shows that

these organizations were using Hadoop as a catalyst for data and organizational maturity in a variety of ways. Hadoop users are more likely to have governance policies in place to oversee and ensure the proper use of information. They focus on hiring and training analytical talent in order to exploit the inherent opportunities in their data. They also surround Hadoop with a variety of complementary technologies that augment the technical sophistication of Hadoop and improve the speed and accuracy of decisions.

For those taking a cloud or hybrid approach however, the opportunities become even more enticing. These companies are able to create efficiency in the decision process, empower their users with cleaner and more usable data, and raise the bar for analytical activity within the organization. Time and time again, Aberdeen's research demonstrates how these factors contribute to reliable and repeatable business performance improvements.

For more information, explore the full research report here: The Horsepower of Hadoop: Fast and Flexible Insight with Results

**45%**

**of Hadoop users are also using cloud-based data visualization tools.**