O'REILLY®

Migrating Big Data Analytics into the Cloud

Mike Barlow

SAN JOSE

LONDON

NEW YORK

SINGAPORE

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

Migrating Big Data Analytics into the Cloud

Mike Barlow

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Migrating Big Data Analytics into the Cloud

by Mike Barlow

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: Holly Bauer **Illustrator:** Rebecca Demarest

October 2014: First Edition

Revision History for the First Edition:

2014-10-01: First release

ISBN: 978-1-491-91698-8

[LSI]

Table of Contents

Survey Reveals Perceived Challenges and Benefits of "As-a-Service" Models for Analytics

Editor's Note: This report contains proprietary statistical, anecdotal, and observational research on the current state of big data analytics in the cloud. We are sharing the information for the benefit of users, decision-makers, and suppliers operating within the big data analytics community. For the purposes of this paper, we are including all frameworks for managing big data (e.g., relational, non-relational, NoSQL), regardless of the underlying architecture.

Into the Cloud

Despite the steady migration of numerous IT capabilities into the cloud, many organizations have been reluctant to embrace the idea of "big data-as-a-service." On one hand, cloud-based big data analytics squarely address ongoing issues of scale, speed, and cost. On the other hand, they also create new issues around privacy, latency, and veracity.

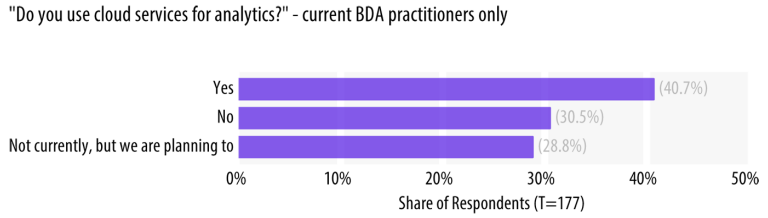
Oftentimes, the best way to gain insight into a complex technology problem is by digging beneath the surface and surveying the perceptions of the user community. John King, a data analyst at O'Reilly Media, designed and conducted a survey of the O'Reilly community, which typically includes software developers, systems architects, engineers, data scientists, and data analysts. The survey was conducted from July 9 through August 3, 2014. There were 312 respondents from various industries including technology, healthcare, finance, and telecom.

One of the main takeaways from the survey is that after an organization has gained some experience using big data in the cloud, it is more likely to expand its use of similar big data services. In other words, once they've tested the waters, they're more likely to jump into the pool.

That insight is neither surprising nor illogical. Humans are hard-wired to be suspicious of novelty, and many executives still regard the cloud as something new and largely unexplored. For suppliers of big data in the cloud solutions, the primary challenge is helping customers take the first steps.

According to our survey, the market seems ready for that kind of approach. The survey shows that roughly 40% of respondents who identify themselves as big data practitioners currently use cloud services for analytics, slightly more than 30% are not, and slightly less than 30% are planning to in the future.

Moreover, the survey shows that roughly 55% of respondents who plan to become big data practitioners also plan to use cloud services for analytics, compared to roughly 45% who said they would not.

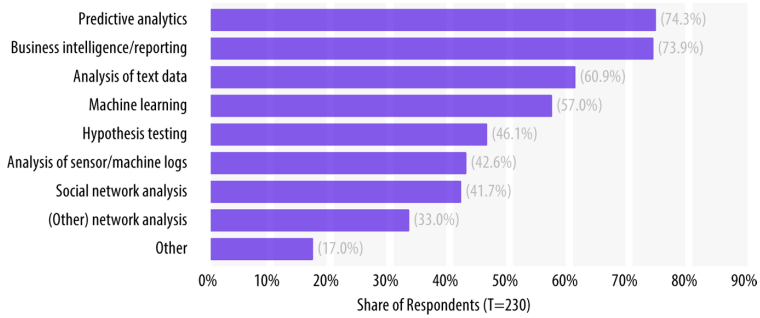


Current and Planned Use of Analytics

The survey also showed that more than 70% of respondents currently use or plan to use predictive analytics and business intelligence/reporting capabilities. Roughly 60% currently use or plan to use big data analytics for text mining or machine learning, and slightly less than 50% currently use or plan to use big data analytics for hypothesis testing. The poll results correspond with anecdotal research suggesting that big data is perceived mainly as a platform for advanced analytics and enhanced BI.

Interestingly, only about 40% of respondents indicated they currently use or plan to use big data for social network analysis, which seems counterintuitive based on the sheer volume of media coverage around social media topics.

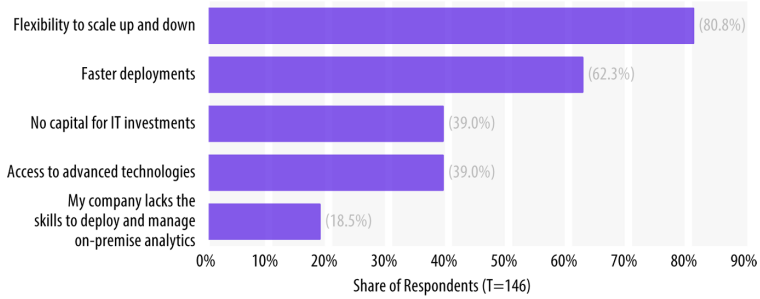
Types of Analytics, current or planned



The survey also revealed that more than 80% of respondents who currently use or intend to deploy cloud-based analytics solutions cite "flexibility to scale up or down" as a reason for choosing a service model over an on-premises delivery model. The next most popular reason cited was faster deployments, followed by reduced capital expenditures, access to advanced technologies, and lack of skills required for managing on-premises analytics.

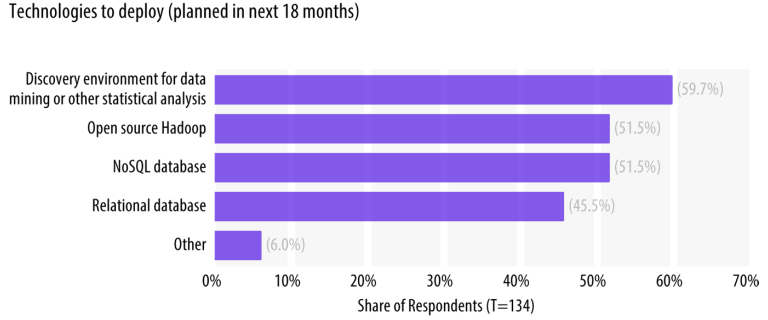
Those findings correspond with the expressed needs of IT executives and other corporate-level decision-makers, who are often quoted as saying the cloud offers the freedom to test new technologies quickly and inexpensively. It also plays to the seasonality of industries such as retail and energy, which often handle vastly different amounts of data at different times of the year.

Reasons to use the cloud for analytics



About 60% of respondents said that within the next 18 months, they planned to deploy a discovery environment for data mining or other

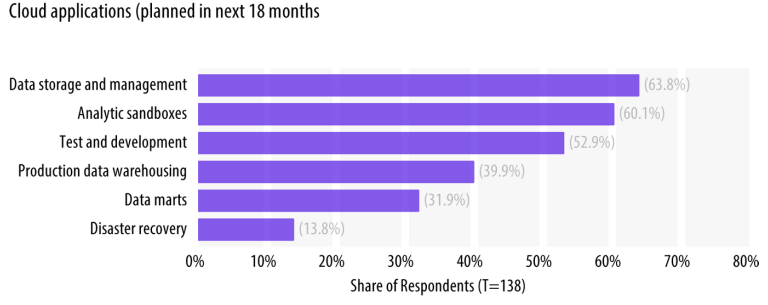
forms of statistical analysis. Slightly more than 50% said they would deploy an open source Hadoop framework or NoSQL database, while slightly less than 50% indicated they were planning to deploy a relational database. Clearly, users and decision-makers are still hedging their bets.



Cloud Applications

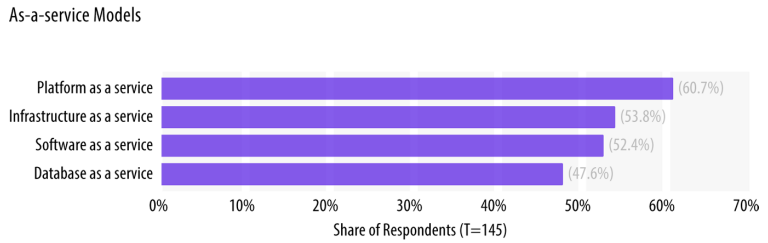
Drilling down into the big data stack, slightly more than 65% of respondents indicated they were currently using or planned to use cloud-based data storage and management applications. Slightly more than 60% indicated they are using or plan to use analytic sandboxes, while slightly more than 50% said they would use cloud-based services for testing and development.

About 40% of respondents said they use or plan to use the cloud for production data warehousing, while slightly more than 30% said they would use the cloud for data marts. Only about 12% said they would use the cloud for disaster recovery, which is somewhat surprising and signals a potential opportunity for cloud vendors.



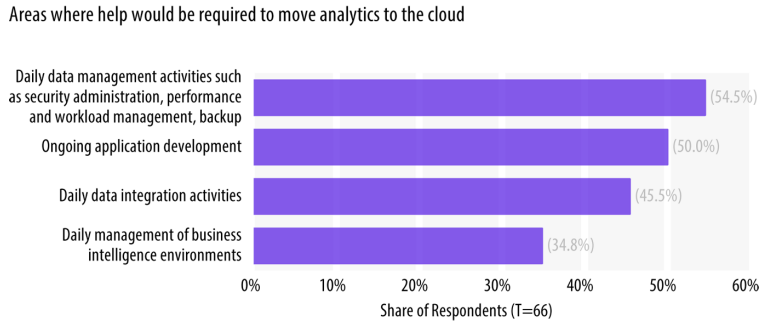
As-a-Service Models

In a ranking of cloud-based "as-a service" models, respondents cited platform-as-a-service, followed by infrastructure-as-a-service, software-as-a-service, and database-as-a-service. Again, this seems to follow larger IT trends, in which organizations follow paths of least resistance to achieve the highest perceived value at any given time.



Areas Where Additional Help Might be Necessary

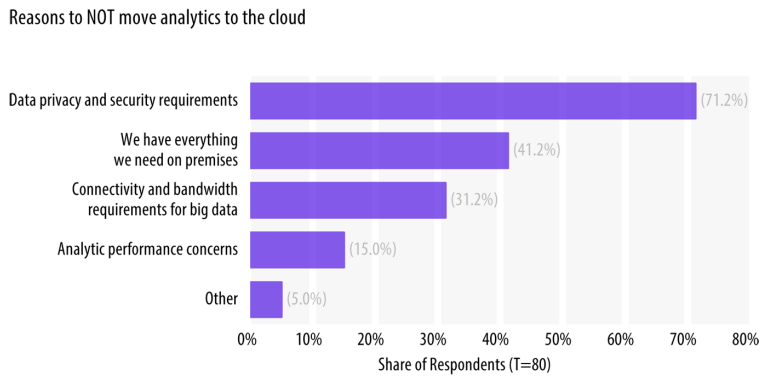
Respondents who indicated they were planning to use cloud-based data analytics in the future were also asked to rank areas in which they would require help moving analytics into the cloud. The survey results showed concern around daily data management activities (e.g., security administration, performance and workload management, and backup); ongoing application development; daily data integration issues; and daily management of business intelligence environments.



Reasons for Reluctance

Respondents who indicated they do not use cloud-based analytics were asked to choose the main reasons for their reluctance. Most chose

data and privacy requirements as the primary reason, followed by existing on-premises capabilities, connectivity and bandwidth issues, and performance concerns.



Those findings correspond with anecdotal and observational research. It seems clear that balancing the benefits and risks of a big data cloud strategy can be a daunting task. Based on our interviews with users and subject matter experts, advantages include scalability, elasticity, agility, lower cost, and rapid innovation. The areas of most concern are performance, security, bandwidth, and data accuracy.

Table 1. Benefits and challenges of moving big data into the cloud.

Benefits/Advantages	Obstacles/Concerns
Scalability/elasticity	Performance
Agility	Security
Cost	Bandwidth
Time to market	Data accuracy/veracity

"With the cloud, you're always going to be cutting edge with the push of a button or the swipe of a credit card," said Marc Clark, director of cloud strategy and deployment at Teradata. "Cloud vendors will continue adding new features to keep ahead of their competitors, which means that even smaller companies can use the latest technology. You just cannot do that with on-premises solutions. I know of many companies with on-premises technology that is two or three generations behind what's currently available through the cloud."

Flexibility is a main driver of cloud adoption, according to Clark. "The cloud lets you weigh the advantages and disadvantages of a system without committing the resources that would be required if you were

going to buy it or enter into a multi-year licensing deal with a vendor," he said. "That's the beauty of the cloud. You can test something for two or three weeks and see if it's right for you, without having to give up your right arm."

The ability to test systems in the cloud, in much the same way that a consumer test drives a new car before buying it, helps companies overcome their reluctance to adopt cloud-based data warehouse services. "Some companies worry that performance levels will be compromised in the cloud. The best way to find out is by trying out a cloud-based service for a couple of weeks. Then you get to test your assumptions and discover for yourself if the cloud makes sense for your organization," said Clark.

Some companies might discover they have bandwidth issues that would prevent them from taking advantage of a cloud-based service. Some might decide to upgrade their network connectivity, while others might decide to stick with their on-premises solution. "What's important is that you find out in minutes or hours, instead of finding out in weeks or months," said Clark. "You will know very quickly whether the cloud is meeting your performance needs."

For some IT executives, moving workloads into the cloud represents a potential loss of control. "If you feel the need to tuck your servers in at night and tell them a bedtime story, then there might be a problem," said Clark. "Some people see a disadvantage in not owning the hardware and the software. They feel as though they are losing control. Or they hear about someone who moved workloads into the cloud, and then moved them back on premises. The cloud is not the right solution for everybody."

Unexpected Costs

Stories about organizations encountering unexpected costs when moving into the cloud are common. Unrealistic expectations generated by endless media hype about the unparalleled virtues of the cloud can set the stage for disappointment. "If your primary motivation is cost reduction, don't move into the cloud," said Clark. "Lots of people think cost is the number-one reason to choose the cloud, but that's the wrong way to look at it, especially if you're planning on moving your big data analytics into the cloud."

When you're building a business case for the cloud, your primary considerations should be speed, scalability, and flexibility. "The cloud enables business agility. The cloud is elastic, so if your business suddenly grows you can respond very quickly if you're in the cloud. The same holds true if your market shrinks. You can move in either direction much more rapidly and with much more freedom," said Clark.

Paul Barsch, a services marketing director at Teradata, recommends taking the time to perform due diligence reviews with potential cloud vendors before moving forward with a services plan. "Make sure the supplier provides basic cloud infrastructure functions and walk the supplier through an itemized checklist of your requirements," he said. A typical checklist would include:

1. Hardware/software monitoring and maintenance
2. Security administration
3. Resource provisioning
4. Networking
5. On-boarding
6. Data center management
7. Backup and recovery
8. System availability
9. DBA support
10. Daily operational management

Ideally, cloud-service suppliers should also provide high-level logical and physical data models, industry report templates, consulting (when necessary), data integration management, data migration, and other capabilities required for launching successful cloud implementations.

In a 2013 article coauthored with Ed White, general manager for Teradata Cloud Solutions, Barsch advised against focusing solely on "lowest cost per terabyte" solutions and noted that "it's important to recognize that not all cloud infrastructures are created equal." Some kinds of cloud infrastructures are better at handling data analytics than others, so make sure the supplier has the appropriate infrastructure for supporting your analytic workloads.

Since analytics tend to be CPU- and I/O-intensive, "it does not make sense to run analytic workloads on cloud infrastructures that are pro-

visioned for general-purpose computing one day and for data warehousing the next," according to Barsch and White. "Business users expect top-tier performance, which is why a cloud environment dedicated and engineered specifically for analytic workloads is imperative."

For CIOs and other IT leaders who are looking to forge closer ties to the business side of their companies, big data in the cloud offers the promise of compressed development cycles for IT and faster time-to-market for the business. By default, however, IT leaders also would have to run interference for the business, since moving data into the cloud would likely require sign-offs from the chief corporate counsel and the CFO, as well as various internal boards and committees of the corporation.

About the Author

Mike Barlow is an award-winning journalist, author, and communications strategy consultant. Since launching his own firm, Cumulus Partners, he has represented major organizations in numerous industries.

Mike is coauthor of *The Executive's Guide to Enterprise Social Media Strategy* (Wiley, 2011) and *Partnering with the CIO: The Future of IT Sales Seen Through the Eyes of Key Decision Makers* (Wiley, 2007). He is also the writer of many articles, reports, and white papers on marketing strategy, marketing automation, customer intelligence, business performance management, collaborative social networking, cloud computing, and big data analytics.

Over the course of a long career, Mike was a reporter and editor at several respected suburban daily newspapers, including *The Journal News* and the *Stamford Advocate*. His feature stories and columns appeared regularly in *The Los Angeles Times*, *Chicago Tribune*, *Miami Herald*, *Newsday*, and other major US dailies.

Mike is a graduate of Hamilton College. He is a licensed private pilot, an avid reader, and an enthusiastic ice hockey fan. Mike lives in Fairfield, Connecticut, with his wife and two children.