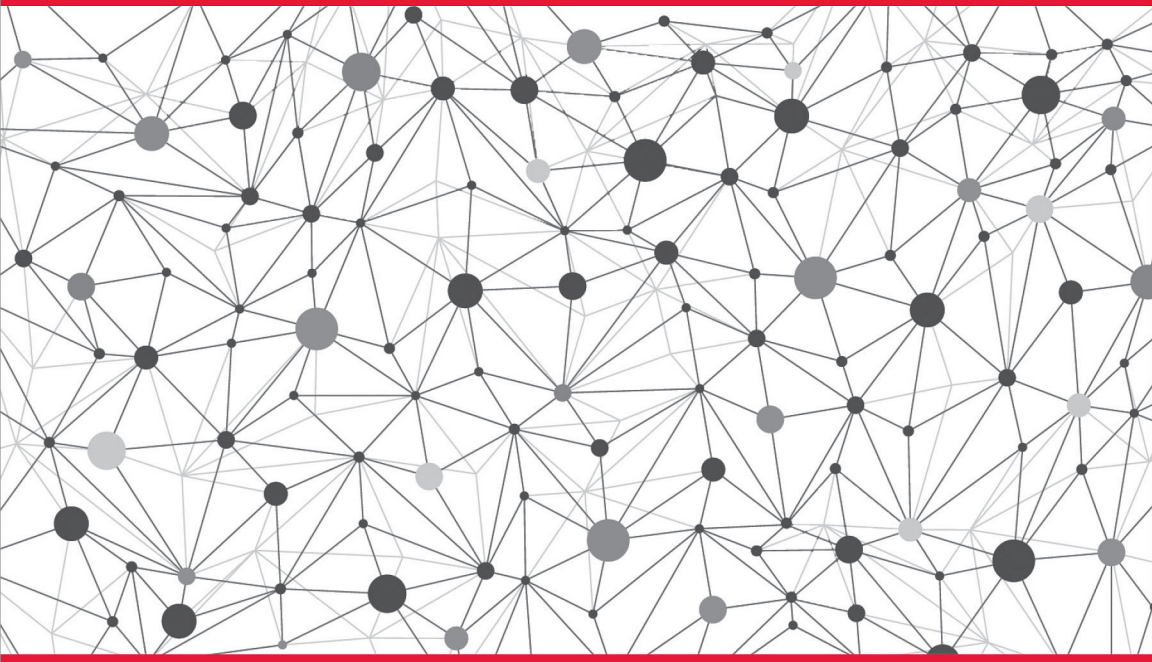
O'REILLY®

Evolving Data Infrastructure

Tools and Best Practices for Advanced Analytics and AI



Ben Lorica & Paco Nathan

Evolving Data Infrastructure

Tools and Best Practices for Advanced Analytics and AI

Ben Lorica and Paco Nathan

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Evolving Data Infrastructure
by Ben Lorica and Paco Nathan

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://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Mac Slocum
Production Editor: Katherine Tozer
Copyeditor: Octal Publishing, LLC
Proofreader: Sharon Wilkey

Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

January 2019: First Edition

Revision History for the First Edition
2018-12-14: First Release

978-1-492-05076-6

[LSI]

Table of Contents

Evolving Data Infrastructure

Introduction

We know that companies are moving key pieces of their data infrastructure to the cloud. However, the lack of data is a bottleneck for companies that want to take advantage of artificial intelligence (AI). In many instances, this is *literally* the case: they want to use machine learning models but haven't collected the data needed to train them.

We wanted to understand how companies are using and combining the ABC components (AI, big data, cloud) as they become more serious about analytics and automation. The means of collecting and storing data, processes for data preparation, tools for querying, and so on are table stakes for organizations that want to start evaluating AI use cases. Additional data infrastructure components are required for companies that have serious plans for production work.

On one hand, we wanted to see whether companies were building out key components. On the other hand, we wanted to measure the sophistication of their use of these components. In other words, could we see a roadmap for transitioning from legacy cases (perhaps some business intelligence) toward data science practices, and from there into the tooling required for more substantial AI adoption?

Here are some of the notable findings from the survey:

- Companies are serious about machine learning and AI. Fifty-eight percent of respondents indicated that they were either building or evaluating data science platform solutions. Data science (or machine learning) platforms are essential for com-

panies that are keen on growing their data science teams and machine learning capabilities.

- Companies are building or evaluating solutions in foundational technologies needed to sustain success in analytics and AI. These include *data integration and Extract, Transform, and Load (ETL)* (60% of respondents indicated they were building or evaluating solutions), *data preparation and cleaning* (52%), *data governance* (31%), *metadata analysis and management* (28%), and *data lineage management* (21%).
- Data scientists and data engineers are in demand. When asked which were the main skills related to data that their teams needed to strengthen, 44% chose data science and 41% chose data engineering.
- Companies are building data infrastructure in the cloud. Eighty-five percent indicated that they had data infrastructure in *at least one* of the seven cloud providers we listed, with two-thirds (63%) using Amazon Web Services (AWS) for some portion of their data infrastructure. We found that users of AWS, Microsoft Azure, and Google Cloud Platform (GCP) tended to use multiple cloud providers.
- Companies used a variety of streaming and data processing technologies. We learned that half of the respondents (49%) used either Apache Spark or Spark Streaming. Other popular tools included open source projects (Apache Kafka, Apache Hadoop) and their related managed services in the cloud (Elastic MapReduce, AWS Kinesis).
- Business intelligence uses a mix of open source and managed services. When it comes to SQL, we found that respondents favored open source tools (Spark SQL, Apache Hive) and managed services in the cloud (AWS Redshift, Google BigQuery).
- Use of durable cloud storage is prevalent, and 62% of all respondents indicated they used at least one of the following: Amazon S3 or Glacier, Azure Storage, or Google Cloud Storage.
- Although a majority (60%) aren't using serverless technologies, one-third (30%) are already using AWS Lambda. In fact, 38% indicated that they were using *at least one* of the serverless technologies we listed. We found this pattern was consistent across geographic regions.

Survey Respondents

The survey ran for a few weeks in late October 2018, and we received more than 3,200 responses. There were more than 1,400 respondents from North America, close to 900 from Western Europe, and more than 350 from Asia (South and East Asia). Figure 1-1 presents the complete breakdown.

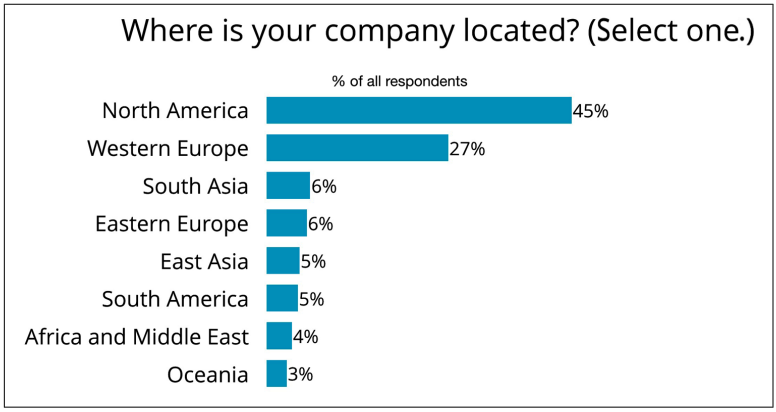


Figure 1-1. Geographic distribution of survey respondents

For the remainder of this report, we've adopted the following terminology to describe these cohorts from our survey:

Exploring
: Respondents who work for organizations that are just beginning to use cloud-based data infrastructure.

Early adopter
: Respondents who work for organizations that have been using cloud-based data infrastructure in production for one to three years.

Sophisticated
: Respondents who work for organizations that have been using cloud-based data infrastructure in production for more than four years.

About one-third (31%) of respondents are still in the early stages of using cloud-based data infrastructure. It's interesting to note the geographic distribution of respondents versus the maturity of their cloud adoption. As Figure 1-2 illustrates, North America has a

higher proportion of sophisticated respondents, whereas Eastern Europe and East Asia have a higher rate who are exploring.

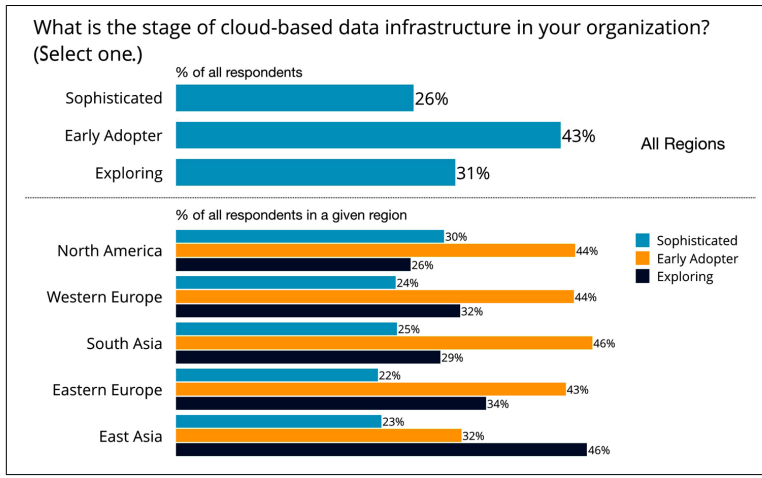


Figure 1-2. Stage of cloud-based data infrastructure

Toward AI: Foundational Data Technologies

For most companies, the road toward machine learning often initially involves work with simpler analytic applications. This isn't surprising given that machine learning requires data, and many simpler analytic tools that precede machine learning require data infrastructure to be in place already.

The growing interest in machine learning will spur companies to continue investing in the foundational data technologies that are required to scale and sustain their AI initiatives. Technologies for collecting, cleaning, storing, and making data available are critical.

In the past 6 to 12 months, we've also been hearing more companies voice interest in solutions for managing data lineage and metadata, both critical to organizations that want to use machine learning and AI across products and systems. For example, we found that one-fifth (21%) of all respondents were either currently building or evaluating solutions to help them manage data lineage, and a majority of companies are interested in solutions for data integration and data preparation, as shown in Figure 1-3.

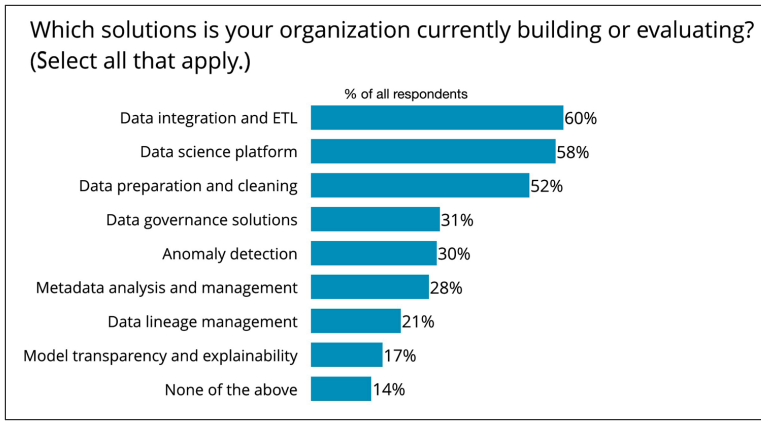


Figure 1-3. Priorities for solutions needed

Companies that employ teams of data scientists have a growing interest in data science or machine learning platforms. Data science platforms typically support collaboration, multiple machine learning libraries, notebooks, and other features. See, for example, recent descriptions of internal data science platforms from Uber, Facebook, Netflix, and Twitter.

We found that 58% of respondents were interested in data science platforms, spread across companies of various stages of cloud adoption: one-quarter (26%) of early adopters of cloud technologies were either building or evaluating a data science platform solution. That said, as Figure 1-4 demonstrates, compared with other organizations, the early adopters are still relatively focused on the more basic stages of data pipelines: data integration and ETL, data preparation and cleaning, and data science platform.

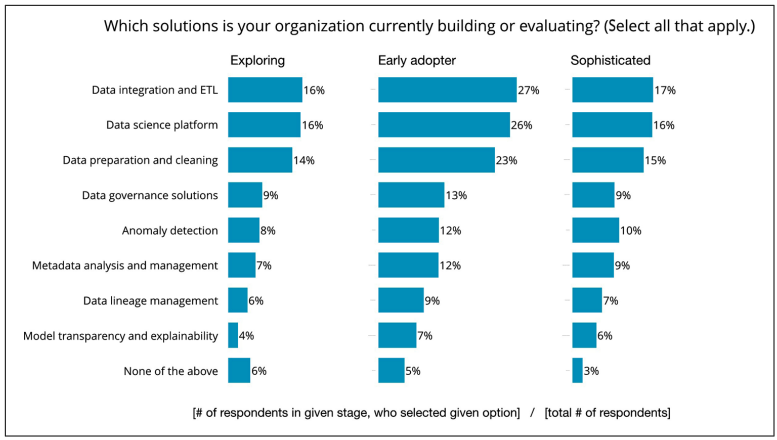


Figure 1-4. Priorities for solutions, by stage of maturity

As an alternative to Figure 1-4, we also compared the distributions for each stage. In Figure 1-5, note that percentages for a given stage don't add to 100%; instead, they show the percentage of respondents at a given stage who selected that option. For example, 63% of respondents from both the early adopter and the sophisticated organizations selected *data integration and ETL* as a current focus. That's significantly higher than the 51% from organizations that are still exploring data infrastructure in the cloud.

We see how this differentiation continues across other solutions: *data science platform*, *data preparation and cleaning*, *anomaly detection*, *metadata analysis and management*, and *model transparency and explainability*. A plausible interpretation would be that as soon as companies begin to build out data infrastructure in the cloud, these kinds of solutions become higher priorities. That is useful advice for organizations that haven't developed their cloud infrastructure yet: consider adopting these priorities, as well, sooner rather than later.

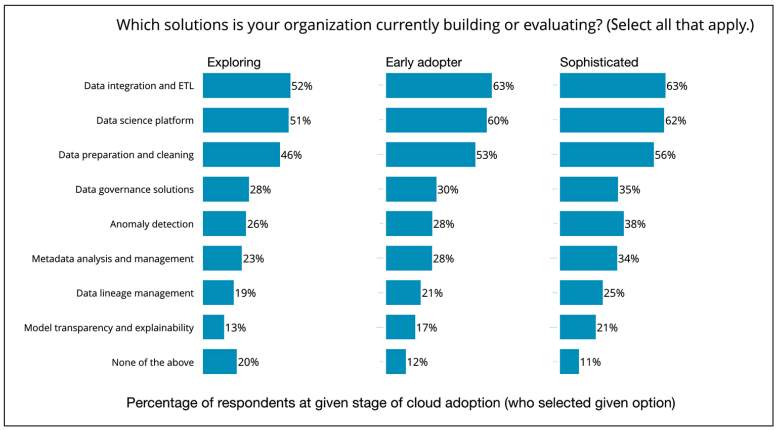


Figure 1-5. Priorities for solutions, by stage of maturity (percentage of respondents)

We found interest in foundational data technologies to be strong across geographic regions. Figure 1-6 shows that one-quarter (25%) of respondents based in North America were interested in solutions for managing data lineage, and a majority of respondents in North America, Western Europe, and Asia were addressing needs in data integration and data preparation.

Figure 1-6. Priorities for solutions, by geographic region

Skills and Roles

Specialized roles for managing cloud services and deployments are well established. As Figure 1-7 illustrates, respondents noted DevOps (47%), platform engineer (18%), and site reliability engineer (10%) as specialized roles related to cloud use, versus one-fifth (21%) of the teams that self-serve for their cloud needs. There's also growing interest in DataOps (14%), although its definition is not as clear yet.

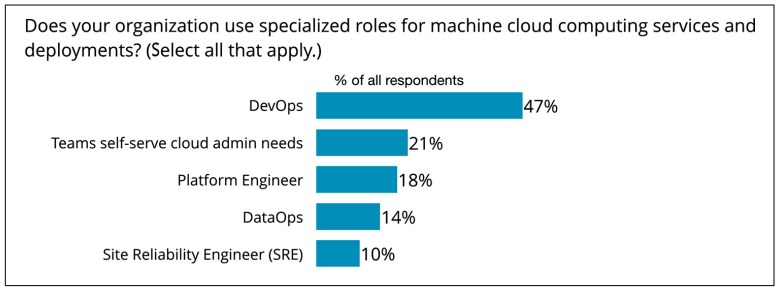


Figure 1-7. Specialized roles

The geographic distribution for those specialized roles was very similar across North America, Western Europe, and Asia. However, note in Figure 1-8 that DevOps gets a bump among early adopters. Given that some cloud practices have been in place for several years, the more specialized roles might be more frequent among early adopters who will have structured their organizations more recently.

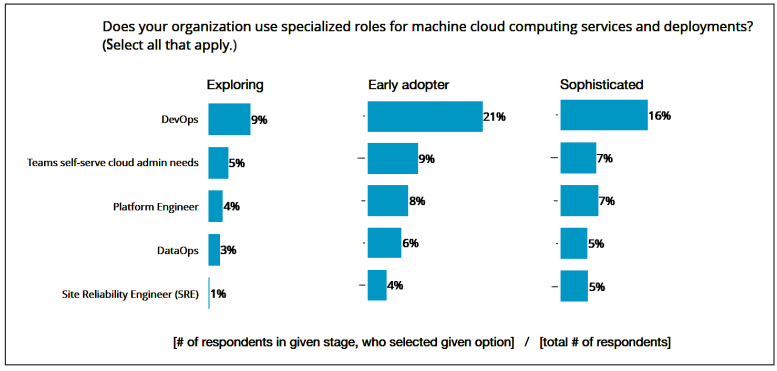


Figure 1-8. Specialized roles, by stage of maturity

Looking at that point about DevOps, is this true if you factor in the size of each group? Again, we looked at these distributions with a different tallying approach to show the percentage of respondents at each given stage who selected each option. See in Figure 1-9 how these specialized roles are amplified among the early adopter and sophisticated organizations. That's telling, and it provides good advice for organizations that follow.

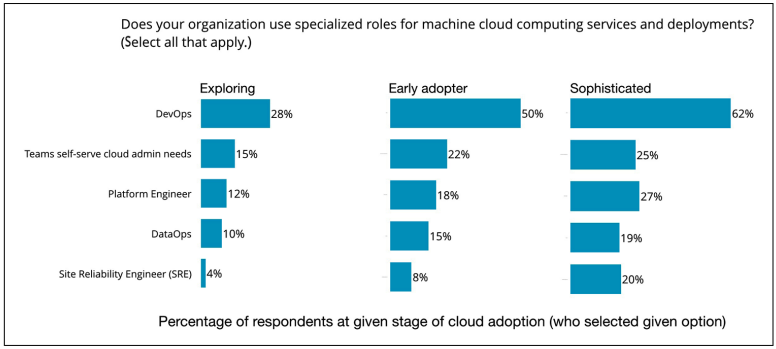


Figure 1-9. Specialized roles, by stage of maturity (percentage of respondents)

In a previous survey, we found that a skills gap (lack of skilled people) remains one of the key factors holding back the adoption of machine learning. Respondents in our current survey expressed the need to strengthen many key roles, including data science (44%) and data engineering (41%), as presented in Figure 1-10.

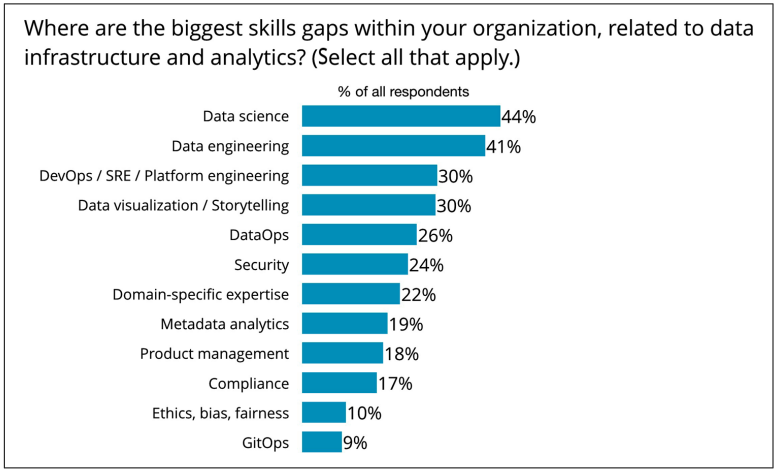


Figure 1-10. Biggest skills gaps

Along the same lines, LinkedIn found that within the United States, demand for data scientists is "off the charts." We found demand for data science and data engineering talent to be strong across all regions. For example, as Figure 1-11 demonstrates, more than half (52%) of respondents based in Asia expressed the need to strengthen their data science teams.

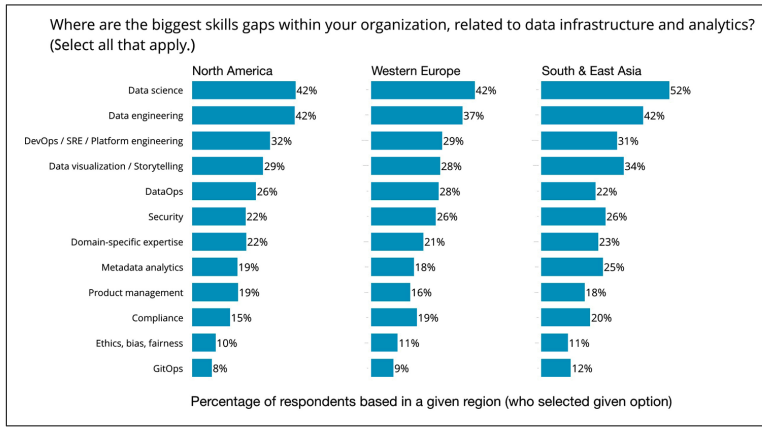


Figure 1-11. Biggest skills gaps, by geographic region

We wanted to try to differentiate between adoption rates for the basic components needed for data science work and some of the more advanced practices required for machine learning in production. For example, repairing metadata and tracking data lineage are needed for serious machine learning work that is subject to regulatory compliance and other accountability.

Although a majority of companies are paying attention to the most foundational work (e.g., ETL, data prep, analytics platforms) required for data science, a larger portion than expected are building or evaluating more sophisticated practices—again, required for work on ethics, bias, compliance, and so on—such as data governance (31%), metadata analysis and management (28%), and data lineage management (21%), as noted in Figure 1-3. Using the skills gap as a measure of demand, similarly more than one-third (35%) of respondents included at least one of the following: compliance, metadata analytics, or ethics/bias/fairness.

Data Infrastructure Technologies

Recent surveys of CIOs suggest that many are planning significant investments in cloud, AI, and automation technologies. Are companies embarking on data infrastructure projects on public cloud platforms? If so, which technologies are being used most?

Cloud Platforms

We provided our respondents with a list of seven major cloud providers and asked whether they were planning to use them for data infrastructure: 85% picked *at least one* of the seven providers we listed, with two-thirds (63%) indicating that they were using AWS for some portion of their data infrastructure (Figure 1-12).

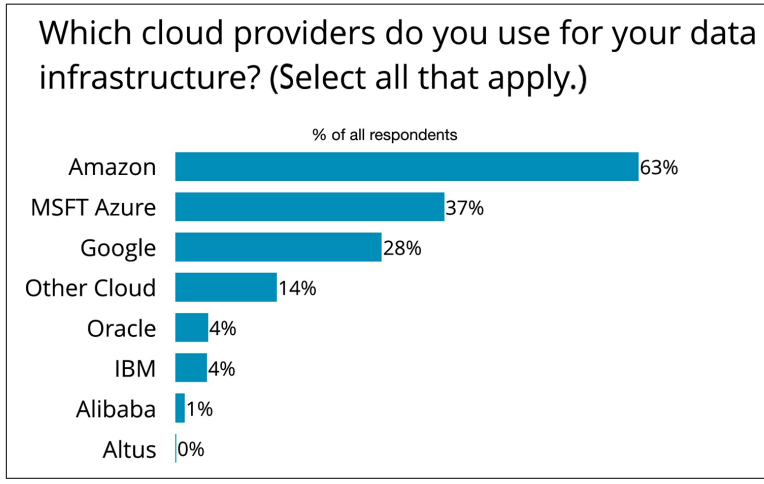


Figure 1-12. Cloud providers used for data infrastructure

Interest in using cloud platforms for data infrastructure held across geographic regions: the percentage of respondents who picked *at least one* of the seven providers we listed was 89% for North America, 83% for Western Europe, and 87% for Asia. Amazon was the favorite cloud platform across regions, as shown in Figure 1-13.

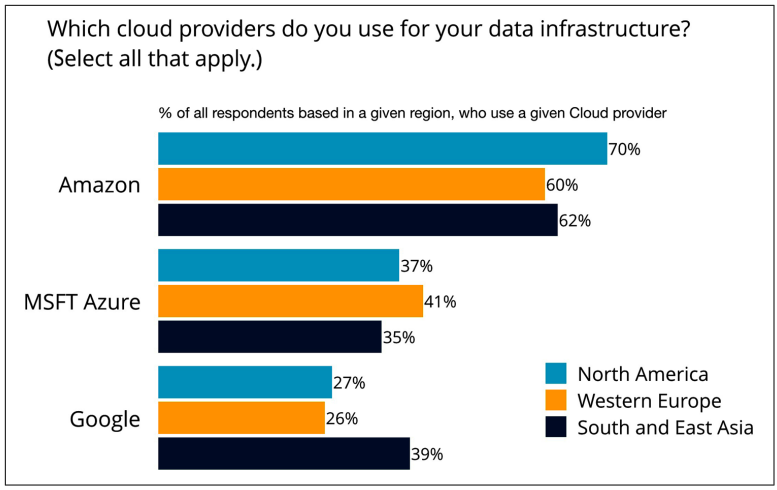


Figure 1-13. Cloud providers, by geographic region

Many companies use more than one cloud provider. Of the 63% of respondents who use AWS for some part of their data infrastructure, only 29% did so to the exclusion of Azure or GCP. In fact, as shown in Figure 1-14, close to 1 in 10 respondents (8%) indicated that they used *all three* major cloud providers (Amazon, Google, Azure) for some of their data infrastructure.

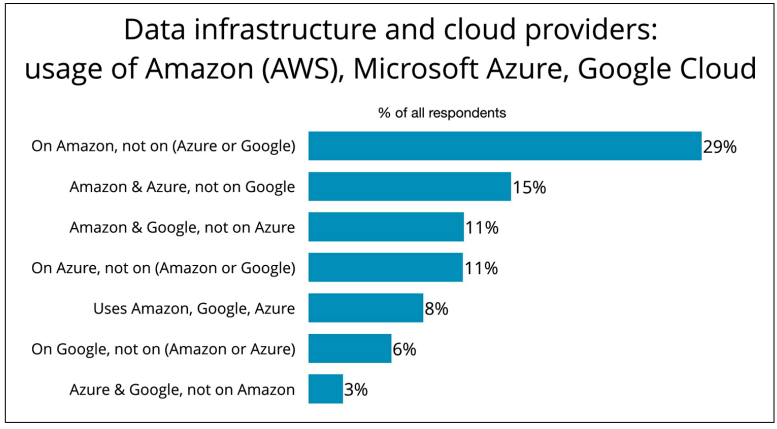


Figure 1-14. Use of multiple cloud providers

Technologies for Streaming and Data Processing

Given the importance of data for training models, companies that are serious about machine learning and AI need strategies and tech-

nologies to collect and store data. Streaming and data processing tools and frameworks are core components of many data platforms. Figure 1-15 shows that half of the respondents (49%) used either Apache Spark or Spark Streaming. Other popular tools included open source projects (Apache Kafka, Apache Hadoop) and related managed services in the cloud (Elastic MapReduce, AWS Kinesis).

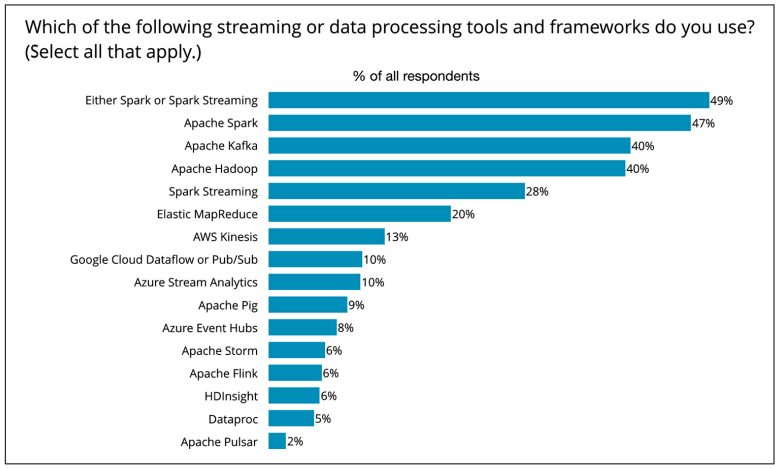


Figure 1-15. Technologies used for data processing and streaming

Usage for specific streaming and data processing technologies was high across regions, with higher usage rates in Asia for Spark, Kafka, and Hadoop, as depicted in Figure 1-16.

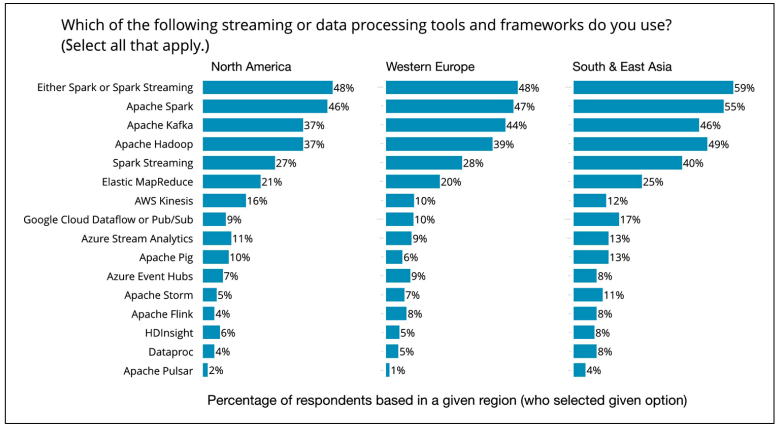


Figure 1-16. Technologies used for data processing and streaming, by geographic region

Note that streaming and more "real-time" infrastructure will need to become increasingly prevalent as reinforcement learning use cases move into production.

NoSQL and SQL

One-quarter of respondents (23%) used Cassandra, and one-fifth used HBase (19%) or DynamoDB (18%) (Figure 1-17):

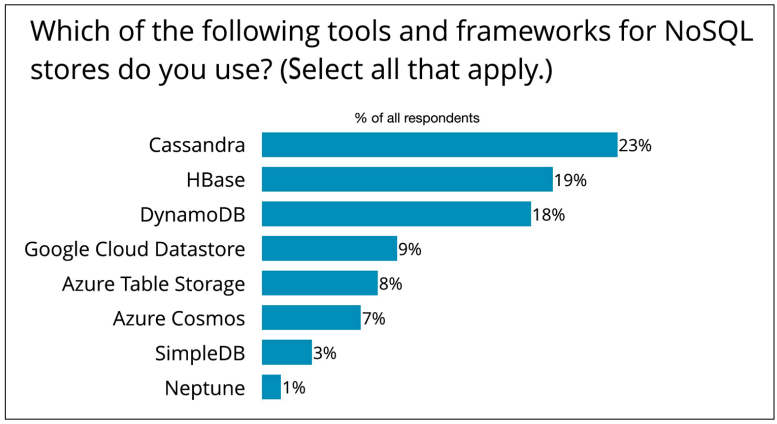


Figure 1-17. NoSQL frameworks

Although the data for this survey did not provide details, it would be interesting to examine the split between SQL and NoSQL. Also, where is the momentum here? Are the sophisticated organizations moving toward particular frameworks?

For many companies, the road to machine learning and AI begins with simpler analytics, which, in many instances, involves the use of SQL tools. Much of the business intelligence (BI) world falls into this category. We found that respondents favored both open source tools (Spark SQL, Apache Hive) and managed SQL services in the cloud (AWS Redshift, Google BigQuery), as illustrated in Figure 1-18.

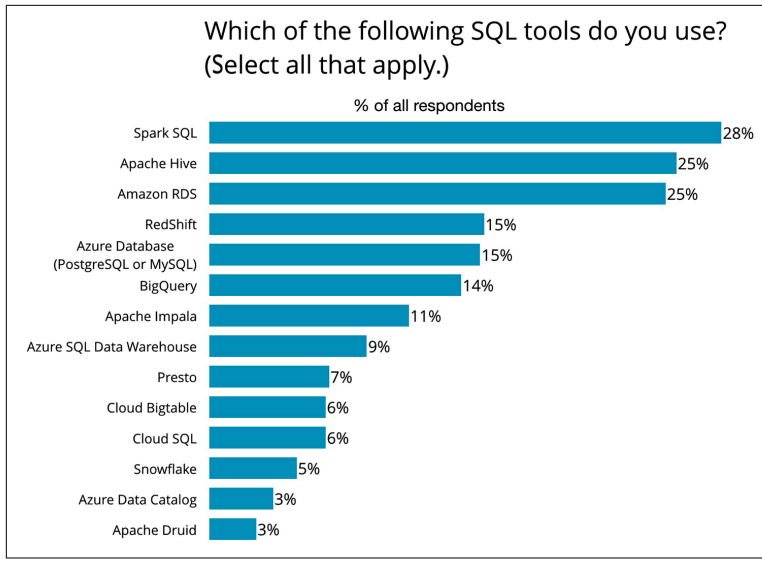


Figure 1-18. SQL frameworks

Storage, Search, and Cache

Long-term object storage in the cloud, such as Amazon Simple Storage Service (Amazon S3), is often described in terms of its *durable* characteristics; for example, 99.999999999% reliability. This precludes files becoming corrupted over time. We found that use of durable cloud storage is prevalent, with 62% of all respondents indicating that they used *at least one* of the following: Amazon S3 or Glacier, Azure Storage, or Google Cloud Storage, as shown in Figure 1-19.

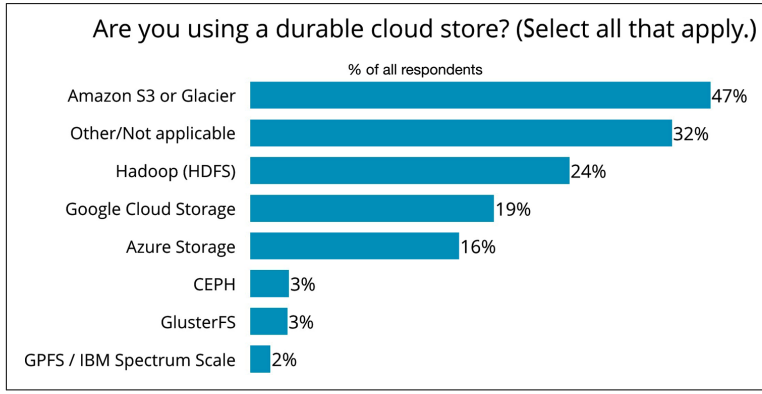


Figure 1-19. Durable cloud storage

Among the regions, the percentage of all respondents who indicated use of *at least one* of those cloud storage options was as follows: North America, 68%; Western Europe, 60%; and Asia, 64%.

Looking at a caching layer for data infrastructure—for example, used to store analytics results for later lookup—Elastic and Redis lead among open source solutions, along with their related managed services in the cloud; for example, ElastiCache and Azure Redis Cache. As demonstrated in Figure 1-20, there's approximately a 5:1 ratio between the use of open source and managed services.

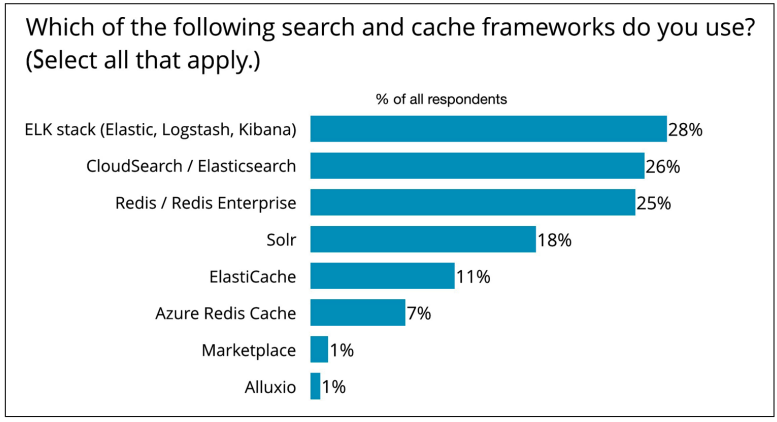


Figure 1-20. Search and cache technologies

Serverless Technologies

Although serverless technologies in the cloud have been growing in popularity among web developers, we were interested in how these are being used for data science and machine learning. In a recent podcast interview, Eric Jonas, from UC Berkeley RISELab, described their research into the performance at scale for both serverless and cloud storage used in analytics. Jonas mentioned that durable cloud storage has increased in speed over the years, which begins to eliminate the need for some frameworks that previously acted to buffer storage access. Jonas also is the author of Pywren and the upcoming NumPywren, which use both AWS Lambda and Amazon S3 to run existing Python code at massive scale. The economics work particularly well for ad hoc queries (too simple to require a pipeline) plus some kinds of machine learning work.

In a February 2017 paper, Jonas and coauthors suggest that "stateless functions are a natural fit for data processing in future computing

environments" as a way to simply distribute computing and reduce the need for complex cluster management and configuration. More recently, the Pywren team has conducted side-by-side comparisons of AWS Lambda against other serverless technology: Google Cloud Functions and Azure Functions. Implications are that although we see much cloud use today for popular data frameworks such as Spark, Kafka, and Hadoop, it's likely that many analytics functions could migrate to serverless at scale as a way to simplify operations and reduce costs.

We found that organizations are still in the early stages of adoption of serverless technologies (Figure 1-21): a majority (60%) aren't using them yet. With that said, one-third (30%) are already using AWS Lambda. In fact, 38% indicated that they are using *at least one* of the five serverless technologies we listed (Apache Pulsar Functions, AWS Lambda, Azure Cloud Functions, GCP Functions, and Nuclio).

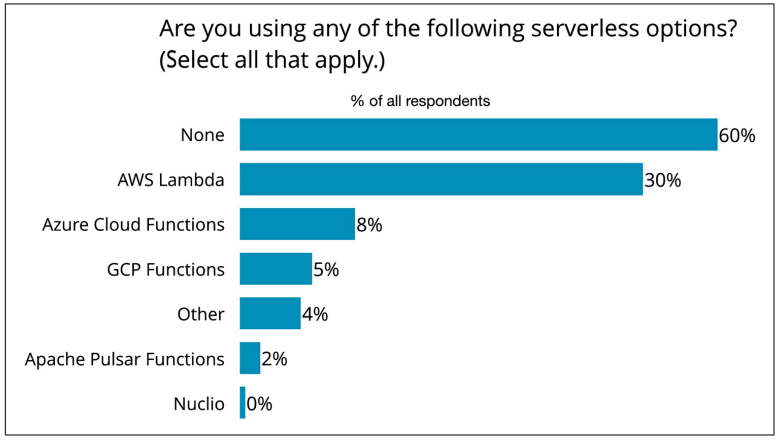


Figure 1-21. Serverless technologies

Even though the geographic distribution for serverless usage was virtually the same across North America, Western Europe, and Asia, it's interesting to see how that 38% adoption spreads across the stages of maturity. Figure 1-22 shows the share of the total number of respondents in each stage who are using one or more of the five serverless technologies we listed as options. Tallying these in a another way (not shown), we can say that 6% of respondents in the exploring stage, 18% in early adopter, and 14% in sophisticated are using *at least one* serverless option.

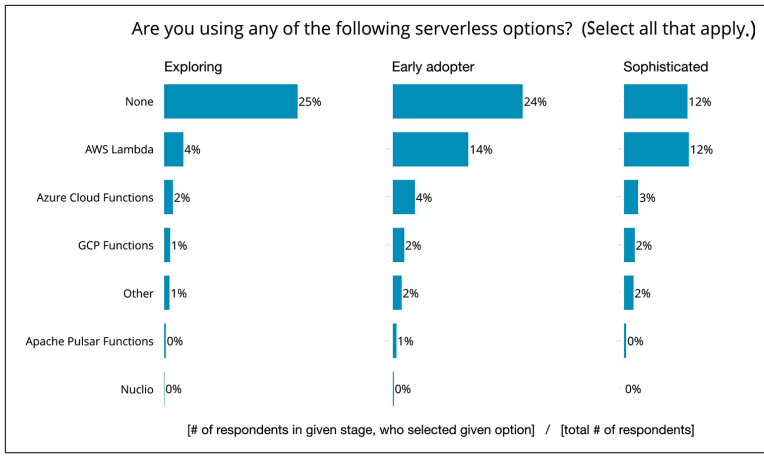


Figure 1-22. Serverless technologies, by stage of maturity

Again, tallying this data to show the percentage of respondents at each stage for each selected option, Figure 1-23 illustrates how half (53%) of those who chose at least one of the serverless options are sophisticated organizations.

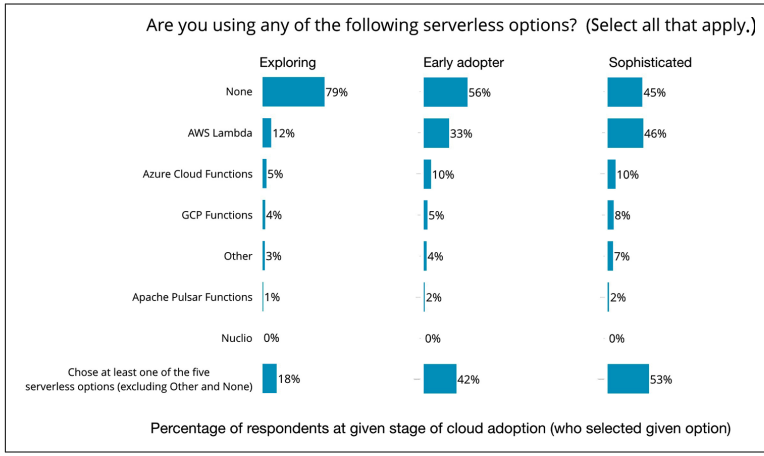


Figure 1-23. Serverless technologies, by stage of maturity (percentage of respondents)

Closing Thoughts

How are companies using the cloud for their data infrastructure? Digging deeper, are they putting the necessary foundations in place to support serious AI adoption? What can we infer about the state of data infrastructure readiness for machine learning in production—beyond the basics required for reporting?

Overall, North America has a higher proportion of sophisticated respondents, whereas Eastern Europe and East Asia have a higher rate of those who are exploring. More than half (58%) indicated that they were either building or evaluating data science platform solutions. Those are table stakes. Digging deeper, one-third are putting the required tooling into place for AI adoption—perhaps to address concerns about metadata, bias, fairness, ethics, compliance, and so on. We may assume that at least 40% are still transitioning from legacy infrastructure; for example, perhaps some BI work but not moving beyond that.

In terms of cloud use for data infrastructure, 85% use at least one cloud, 35% use two clouds, and 8% combine all three major cloud providers. Spark, Kafka, and Hadoop are the most popular data processing tools, along with equivalent managed services in the cloud, with roughly a 2:1 ratio between open source and managed services. For SQL, perhaps the most essential common denominator needed for analytics work, respondents favored both open source tools (Spark SQL, Apache Hive) and corresponding managed services in the cloud (Redshift, BigQuery). Companies use a variety of streaming and data processing technologies: one-half use either Apache Spark or Spark Streaming. Other popular tools included open source projects (Apache Kafka, Apache Hadoop) and related managed services in the cloud (Elastic MapReduce, AWS Kinesis).

In terms of skills and roles related to cloud data infrastructure, data scientists and data engineers are in demand: 44% chose data science and 41% chose data engineering as important skills that their teams needed to strengthen. Specialized roles for managing cloud services and deployments are well established: DevOps is used among one-half for cloud management, along with similar roles, versus one-fifth of the companies using self-serve.

The early-adopter and sophisticated organizations show different priorities than the exploring stage for building and evaluating solutions in the cloud. Those priorities were higher for *data integration and ETL*, *data science platform*, *data preparation and cleaning*, *anomaly detection*, *metadata analysis and management*, and *model transparency and explainability*. Organizations that haven't developed their cloud infrastructure yet may consider adopting these priorities, as well, sooner rather than later.

Uses of storage and processing are undergoing transformation as hardware options for these become richer. Two-thirds of companies use durable storage from at least one of the major cloud providers. Although serverless uses are still early for data analytics, more than one-third use at least one of the five serverless technologies we listed as options, with slightly more use among the sophisticated and early adopters. Similarly, more than one-half of those that use serverless are among the sophisticated organizations.

Current research indicates that wider use of serverless plus durable storage might help simplify distributed computing for data analytics at scale. Even though we see much cloud use today for popular data frameworks such as Spark, Kafka, and Hadoop, it's likely that many analytics functions could eventually migrate to serverless at scale to simplify operations and reduce costs.

About the Authors

Ben Lorica is the chief data scientist at O'Reilly Media and is the program director of both the Strata Data Conference and the Artificial Intelligence Conference. He has applied business intelligence, data mining, machine learning, and statistical analysis in a variety of settings including direct marketing, consumer and market research, targeted advertising, text mining, and financial engineering. His background includes stints with an investment management company, internet start-ups, and financial services.

Paco Nathan is known as a "player/coach" and has core expertise in data science, natural language processing, machine learning, and cloud computing. He has more than 35 years of tech industry experience, ranging from Bell Labs to early stage start-ups. He is co-chair of the Rev summit and an advisor for Amplify Partners, Deep Learning Analytics, Recognai, and DataSpartan. Recent roles include director of the Learning Group at O'Reilly Media, and director of the Community Evangelism at Databricks and Apache Spark. Innovation Enterprise named him one of the Top 30 People in Big Data and Analytics in 2015.