Immuta Data Engineering Survey: 2021 Impact Report

IMMUTA™

Introduction

The advent of 'big data' saw organizations treat data not as a byproduct of systems, but as a source of innovation and competitive advantage.

With the rise of AI and ML, the availability of cloud data platforms, and the digital transformation of nearly all markets, data is no longer just about making more informed decisions — it's now treated as a product in and of itself.

To maximize the value of their data, organizations are building large data teams with both supply-side functions that provide analytics-ready data from across the enterprise, and demand-side teams that analyze, transform, model, and commercialize data. Data engineering teams — the "supply side" of the data value chain — are rapidly adopting cloud data platforms to improve agility, reduce cost, and get more data to more users. As they shift to the cloud, they're rewriting decades of on-premise data pipelines, shifting ETL processes to ELT to enable end-user agility, and adopting multiple cloud analytics platforms and data science tools.

Just as DevOps became a critical function as organizations modernized software development, DataOps is emerging as a guiding principle and ethos for data engineering and architecture teams. Search LinkedIn for "DevOps" and you'll find more than one million people with the term in their job title. Search "DataOps," and you'll find just thousands. This suggests we're still in the early innings of the modernization of "data as product."

Given Immuta's mission and where we sit in the data value chain for our customers, we want to understand where data teams are in their journey to modern DataOps. So we've decided to field an annual survey of data teams with a focus on data engineering, data architecture, data platforms, and the executives that oversee these teams.

This 2021 Impact Report summarizes key findings from our inaugural survey and provides a glimpse into the current and future state of data engineering and DataOps. The report highlights some of the major trends uncovered in this year's survey including the adoption of cloud data platforms, what platforms are winning (and emerging), what data engineers find to be their biggest challenges, and how organizations are handling sensitive data.

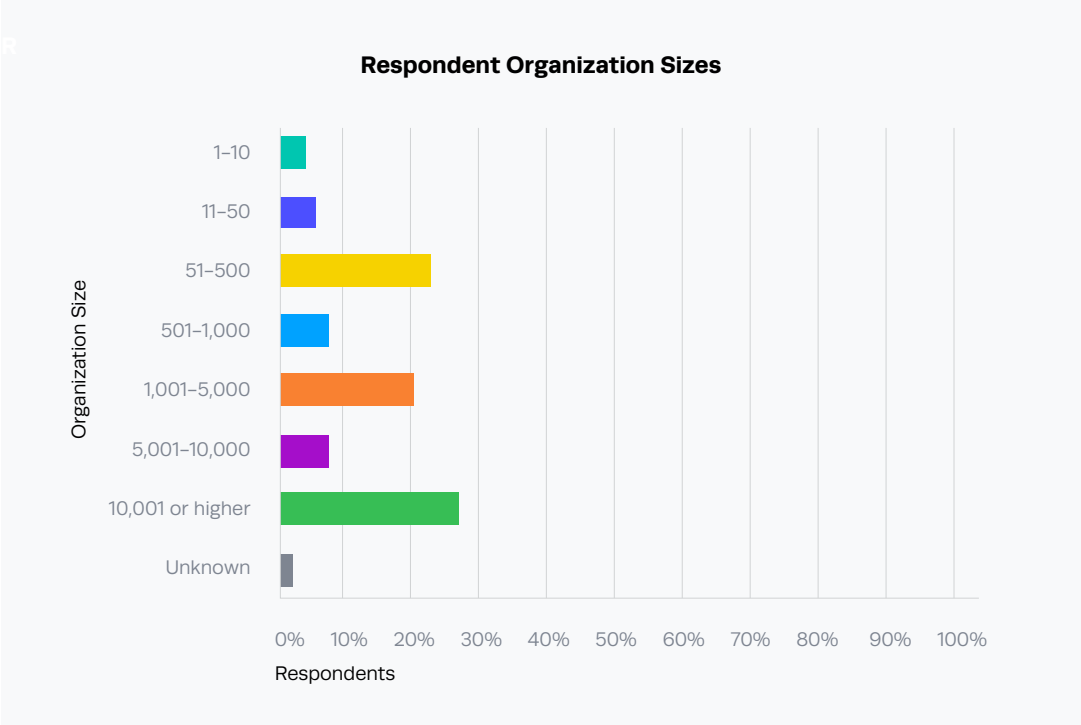
Have a question about the results?

Continue the discussion with us on LinkedIn or Twitter.

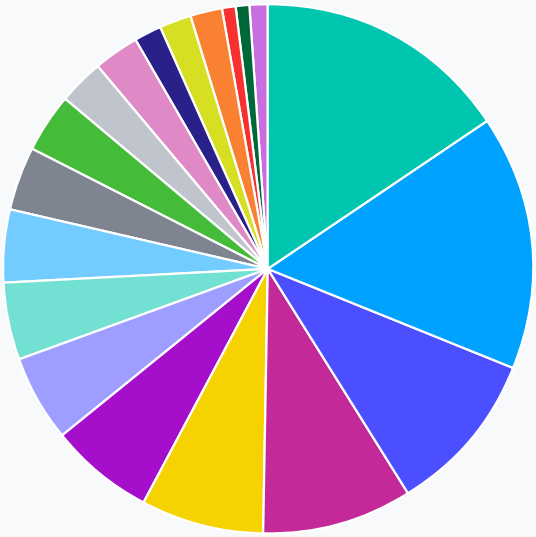
Survey Methodology

To gain insight about the state of Data Engineering and DataOps, we surveyed more than 130 data professionals to understand where they stand in the process of implementing agile data management and modernizing their data stacks.

Participants came from Immuta's database of organizations with some of the most complex data environments, which we've curated since the company's founding in 2015. Respondents represent a diverse group of roles and industries with a concentration in data engineering, data architecture and data governance. Approximately half came from larger organizations with more than 1,000 employees and half from smaller companies with fewer than 1,000. In exchange for participating, respondents receive a copy of the survey data and a small incentive.



Respondent Industry Representation



- Financial Services
- Technology
- Other
- Professional Services/ Consulting
- Insurance
- Retail
- Aerospace & Defense
- HealthTech
- Business Services
- Healthcare Providers
- Manufacturing, Industrial Machinery & Equipment
- Telecommunications
- Education
- Energy, Utilities & Waste
- Transportation
- EdTech
- FinTech
- Oil & Gas
- Restaurant

Adoption of Cloud Data Platforms

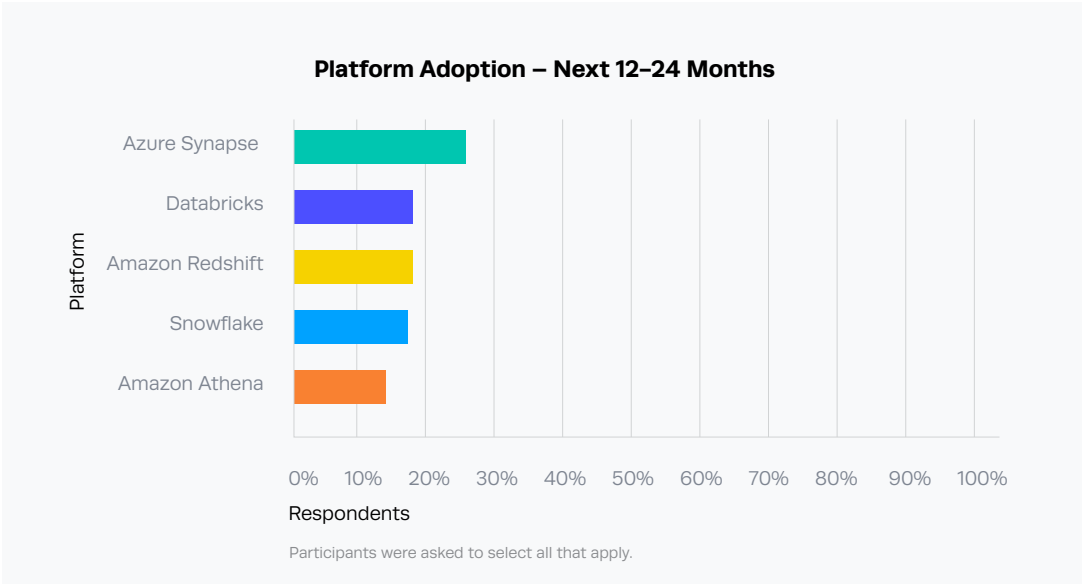
The first part of the survey asked respondents about where they're managing their data — in the cloud, on premise, and using which platforms.

We found that data engineering and platform teams are rapidly adopting cloud data platforms to execute faster and reduce cost, enabling data consumers to analyze and model data more quickly and efficiently.

> **75% of survey respondents expect to be "entirely" or "primarily" cloud-based within the next 24 months, and 52% plan to adopt two or more platforms in the same timeframe.**

Notable among the data was the plan to adopt multiple cloud compute platforms — not just one. Organizations are rapidly building a multi-platform cloud data stack to deliver best-of-breed functionality for different data use cases (e.g. BI, advanced analytics, data science, data sharing and marketplaces, and more).

These five core cloud compute and warehouse platforms were among the most cited by participants to be adopted within the next two years:



Other noteworthy cloud platforms cited include Google BigQuery (10%), Amazon EMR (3%), Cloudera Data Platform (2%), and Starburst Presto (2%).

The size and maturity of data teams, and the amount of data they must deliver, appears to impact the selection of their cloud platforms.

- Azure Synapse, Databricks, and Athena — powerful cloud compute platforms capable of handling extremely large data sets — are more likely to be adopted by enterprise organizations with over 1,000 employees.
- Cloud-native data warehouses like Redshift, Snowflake, and BigQuery are more likely to be adopted by organizations of fewer than 1,000 employees.

In aggregate, our findings suggest the formation of a new cloud data ecosystem — composed of many cloud data technology providers — is underway, and happening faster than many industry analysts have predicted. Not only will the next two years see more organizations shifting most data to the cloud, but the majority of organizations will leverage more than one cloud data platform. While this modern ecosystem stands to significantly accelerate data outcomes by leveraging the superior performance and lower cost of the cloud, it also introduces many new technology and process challenges that data engineering teams will have to overcome.

The formation of a new cloud data ecosystem — composed of many cloud data technology providers — is underway, and happening faster than many industry analysts have predicted.

The Cloud Data Management Process

The next section of the survey asked teams who are utilizing cloud data platforms to assess the level of challenge and satisfaction with existing tools in managing their cloud data.

Data engineering teams are tasked with turning raw data into analytics-ready data that can deliver valuable business insights as well as power machine learning algorithms and data-driven user experiences. The shift to cloud data platforms and warehouses, while providing cost and performance efficiencies, introduces new challenges for data teams.

We asked data teams about seven key steps in the modern cloud data management process:

1. Extracting and loading data
2. Classifying and cataloging data
3. Transforming and modeling data
4. Controlling user access to data
5. Masking or anonymizing sensitive data
6. Auditing data use to prove compliance
7. Sharing and publishing derivative data

We wanted to understand which steps are harder than others, which technologies are commonplace, and where the gaps lie. Let's explore the key findings.

1. Data Integration (Extraction, Loading, & Transformation)

Overall, our analysis suggests the processes and tools for integrating raw data with cloud data platforms are relatively mature. Teams have been handling cloud data integration for a number of years, and the tools available to them have become more refined.

Responses show that data integration — or the process of extracting, loading, and transforming raw data — is the least challenging stage of cloud data management. Sixty-six percent of respondents reported that extracting and loading data into the cloud is "not challenging" or only "moderately challenging." The majority of respondents are also satisfied with their ELT tools — only 19% say they are "unsatisfied" or "somewhat unsatisfied" with current data integration solutions.

> **66% of respondents reported that extracting and loading data into the cloud is "not challenging" or only "moderately challenging."**

One notable exception was the data transformation stage. We found that transformation and modeling of raw data in the cloud appears to be more challenging for most data teams. Almost half of respondents (48%) said that "transforming and modeling" data is "challenging," "very challenging" or "extremely challenging." This may represent the complexities inherent in shifting to the cloud and loading more raw data into data lakes prior to transforming or preparing the data for use.

2. Data Discovery and Access

Once data has been integrated with cloud platforms, data engineering teams must make it discoverable and accessible to data consumers: analysts, data scientists, and others who need to analyze or share it. According to our results, "classifying and cataloging data" and "controlling user access" are more challenging steps in the cloud data management process. Forty-nine percent and 42% of survey respondents, respectively, report these processes as "challenging" to "extremely challenging."

> **45% of data teams either use a homegrown data catalog or don't have a data catalog at all.**

Not only are these processes more challenging, but we found higher levels of dissatisfaction with the tools that enable data discovery and user access control. Forty-six percent of respondents are "unsatisfied" or "somewhat unsatisfied" with their Data Discovery or Data Catalog platforms, for example. Additionally, we found that 45% of data teams either use a homegrown data catalog or don't have a data catalog at all. In a modern cloud data stack — which our data suggests will be heterogeneous with multiple cloud data platforms — the lack of an enterprise-wide data catalog will make data much harder to locate and use securely.

3. Data Security and Access Control

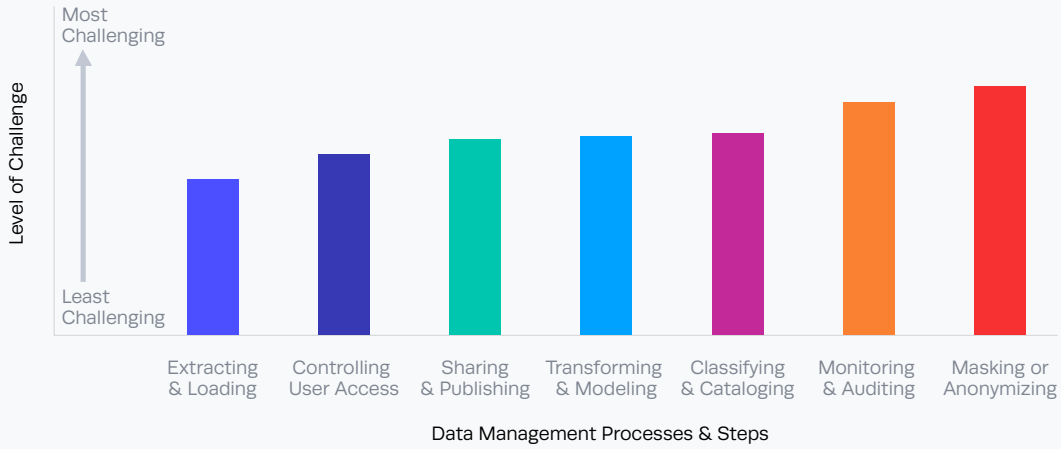
Once data has been integrated into cloud platforms and made discoverable and accessible, the final stages for data engineers are to provide secure access to data and to control who can access what types of data. According to survey respondents, these final steps in cloud data management proved most challenging. Sixty-four percent said masking or anonymizing data is "challenging" or "extremely challenging." Furthermore, 59% of respondents said data monitoring and auditing is "challenging" or "extremely challenging."

> **64% said masking or anonymizing data is "challenging" or "extremely challenging." Furthermore, 59% of respondents said data monitoring and auditing is "challenging" or "extremely challenging."**

Follow-up questions about the satisfaction levels with current data security, governance, and auditing tools corroborates these findings. Forty-one percent of survey respondents are either "unsatisfied," "somewhat unsatisfied" or "unsure" about satisfaction levels with their current data governance and access control tools. The high number of "unsure" responses suggests that many data professionals simply don't know if their data is effectively governed and compliant with all applicable rules and regulations. We'll explore this further in the next section.

Notably, the platforms data teams use impacted whether they ranked data security and access control as "challenging" or "extremely challenging." This may be due to the sheer volume of data sources and data consumers they must manage consistently and securely across the enterprise, especially when using multiple cloud data platforms.

Level of Challenge for Data Management Processes



Results represent a composite 'level of challenge score' across responses.

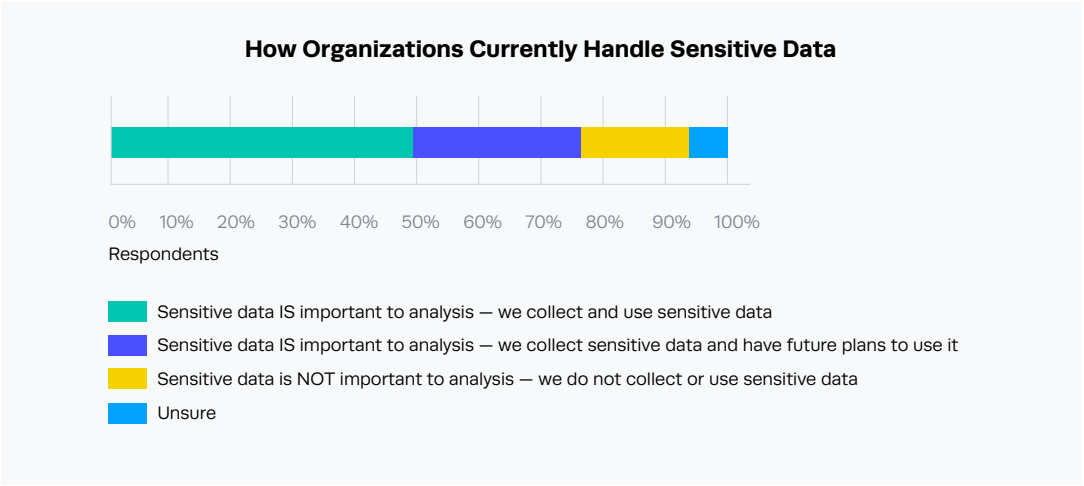
Data teams found "masking or anonymizing data" and "monitoring and auditing data use" to be the most challenging steps in the data management process.

Deep Dive: Handling Sensitive Data

The final section of this year's report focuses on how data teams are handling sensitive data that requires specialized protection or is subject to business rules or privacy regulations.

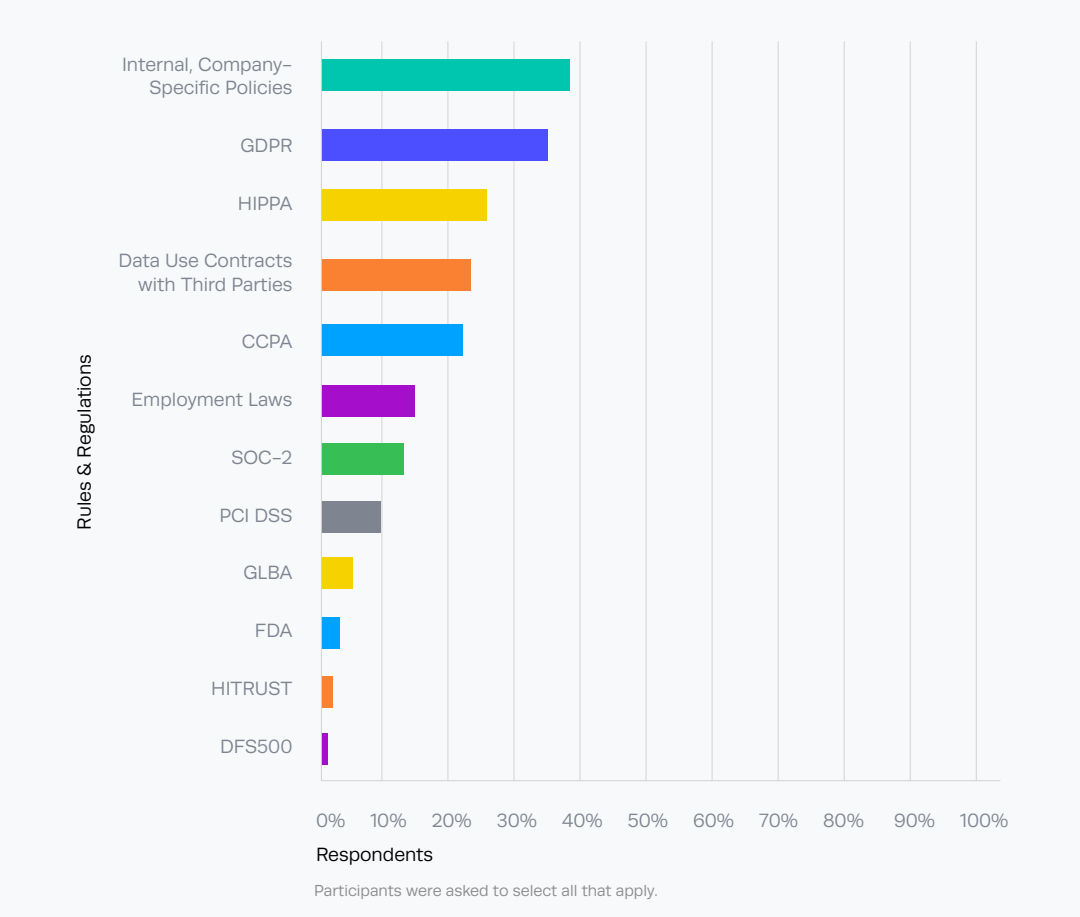
Given the expected rise in adoption of cloud data platforms, we wanted to understand how many organizations are using sensitive data in analytics or data science, and how they're handling access control given the increasingly complex regulatory landscape around sensitive data.

To start, we asked respondents how they handle sensitive data today. Notably, 75% of survey respondents report that sensitive data is important to analysis, and nearly half (49%) currently collect and use sensitive data in analytics.



> **75% of survey respondents reported that sensitive data is important to analysis. Furthermore, half (49%) of respondents are actively using sensitive data in analytics.**

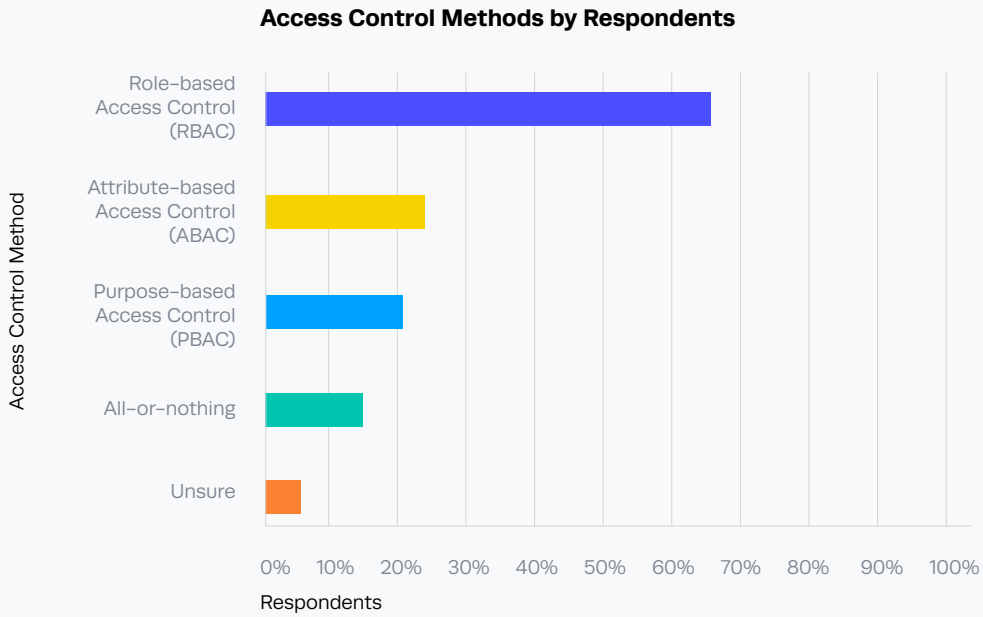
We also asked respondents what rules sensitive data is subject to within their organizations. The majority of survey respondents (92%) indicated their organizations — and their data — are subject to one or more data use rules or regulations that apply across industries and geographies. The most commonly cited are:



Additional data privacy rules respondents must comply with include FERPA, LGPD, APRA, Canada Protected B, ITAR, EAR, and State Insurance Regulation boards. These trends were the same in both larger corporate (>1,000 employees) and smaller (<1,000 employees) organizations, though larger organizations were much more likely to have to comply with SOC-2 and PCI DSS requirements.

Lastly, to understand how data teams are governing and controlling sensitive data access to comply with rules and regulations, we asked survey respondents what type of access control they're using to manage fine-grained data access across platforms. We found that more than 80% of survey respondents report using "role-based access control (RBAC)" or "all-or-nothing" policies.

More than 80% of survey respondents report using "role-based access control (RBAC)" or "all-or-nothing" data access control policies to manage data access.



While 44% indicated they are using attribute- or purpose-based access controls, it's not clear if these are simply policies that data consumers are expected to comply with, or if these are programmatically-enforced policies across cloud data platforms.

Our exploration of sensitive data management suggests nearly all organizations have some data that must be protected. In fact, only 8% of survey participants said they either do not process sensitive data or do not have to adhere to data protection rules and regulations. However, the continued reliance on role-based access control suggests most data teams lack a fine-grained solution to manage data access control and security. Data teams must develop new strategies and adopt new tools that allow data consumers to utilize protected data in accordance with each set of rules.

92% of respondents said their organizations either have sensitive data or must comply with at least one data rule or regulation. Yet the majority of data teams continue to rely on role-based or "all or nothing" access — lacking a fine-grained data access control solution.

Conclusion

The future of cloud data analytics and data science is already upon us. The next two years will present new practices and challenges related to the management of cloud data.

Based on the findings of our first survey of Data Engineers and DataOps, it's clear that data teams are facing a new set of challenges as they move to cloud data platforms and use sensitive data within analytics and data science. Several key trends emerged from our analysis including:

1. **Expanding Data Ecosystem** — Organizations are not just shifting to the cloud in the next two years, but many are adopting two or more cloud data platforms. DataOps teams will need automated, cross-platform tools and processes in order to make the most productive use of their data.

2. **Best-of-Breed Technology** — Data teams are responsible for a range of important steps in making raw data analytics-ready. To streamline this process — particularly in a multi-cloud data ecosystem — they need technology that enables easier and more comprehensive data governance, access control, and sensitive data discovery.

3. **Sensitive Data & Regulations** — The proliferation of sensitive data means that rules and regulations to protect its use will continue to be created, amended, and strictly enforced. Data engineers and architects need dynamic access control tools to help ensure consistent data security across platforms.

IMMUTA™

115 Broad Street, 6th Floor, Boston, MA 021100 | immuta.com | (800) 655-0982