REPORT

2022 State of Data Engineering

Emerging Challenges with Data Security & Quality

JESSE ANDERSON
Managing Partner,
Big Data Institute

BEN LORICA
Principal,
Gradient Flow

JENN WEBB
Managing Editor,
Gradient Flow

GRADIENT FLOW + IMMUTA™

Table of Contents

Executive Summary 3

Methodology 3

Key Insights 4

Introduction 5

Demographics and Key Segments 6

Cloud Computing for Data Processing and Analytics 7

Data Engineering Tools and Platforms 8

Criteria for Evaluating Databases and Data Platforms 8

Databases and Data Platforms 9

Key Data Engineering Challenges 12

Data Integration and Data Orchestration 13

BI and Analytics 15

Data Catalog and Data Discovery 16

Data Quality 17

Data Privacy and Security 18

Closing Thoughts 20

Acknowledgements 21

About Immuta 21

About Gradient Flow 21

Executive Summary

The modern data engineering technology market is dynamic, driven by the tectonic shift from on-premise databases and BI tools to modern, cloud-based data platforms built on lakehouse architectures.

More than the on-premises market that preceded it, the cloud data technology market is evolving rapidly, and spans a vast set of open source and commercial data technologies, tools, and products. At the same time, organizations are adopting multiple technologies to keep up with the scale, speed, and use cases that today's data environment demands.

To remain competitive and maximize the value of their data – including sensitive data – organizations are developing DataOps functions and frameworks to varying degrees. DataOps tools and processes enable continuous and automated delivery of data to power BI, analytics, data science, and data-powered products. The 2022 Data Engineering Survey examined the changing landscape of data engineering and operations challenges, tools, and opportunities.

Methodology

The global online survey ran for 61 days, from June 24 to August 23, 2021. There were 372 respondents. More than half of all respondents were Data Engineers or Data Architects. Respondents were recruited via social media, online advertising, the Gradient Flow Newsletter, the Big Data Institute Newsletter, and the Immuta Newsletter.

This report includes year-over-year comparisons to findings from Immuta's 2020 Data Engineering Survey and 2021 Impact Report, which took place during September and October of 2020.

Key Insights

Data Engineering Challenges

Overall, the most challenging areas cited by respondents were tasks that need to be primarily performed during the data curation and governance phases, once data has been extracted, loaded, and transformed. Top challenges included Data Quality and Validation, Data Monitoring and Auditing for Compliance, Data Masking and Anonymization, and Data Discovery.

The Shift to Cloud Computing for Data Processing and Analytics

Nearly two-thirds of respondents (65%) characterized their company as already either 100% cloud-based or primarily cloud-based. Looking ahead 12-24 months, 81% of respondents projected that they will be 100% cloud-based or primarily cloud-based. This projected shift to cloud-based analytics is 6% higher (compared to 75%) than the findings in the 2021 Impact Report, indicating an accelerating migration from on-premises to the cloud.

Top Cloud Data Platforms

Looking ahead 12-24 months, 62% of respondents signalled that they plan to adopt at least one of the following five cloud databases and platforms: Amazon Redshift, Amazon Athena, Google BigQuery, Databricks, and Snowflake. In the 2021 Impact Report, the top five systems projected for use in the next 12-24 months were SQL Server, Oracle, MySQL, Databricks, and Amazon Redshift. Amazon's rise in the top five rankings and the addition of Google Big Query are both notable, as is the continued popularity of Databricks and Snowflake.

The Need for Data Privacy and Security

Nearly two-thirds (64%) of survey respondents came from companies that already collect and store sensitive data. The vast majority of survey respondents (88%) indicated their organizations are subject to one or more data use rules or regulations, with GDPR, HIPAA, CCPA, and SOC 2 cited as the most common. Additionally, close to one-third of all respondents (30%) reported a need to comply with internal, company-specific rules.

Top BI and Analytics Tools

The most popular BI and Analytics solutions cited by survey respondents were Jupyter Notebooks, Tableau, Microsoft Power BI, Looker, and Google Colab.

The Challenge of Data Quality

Respondents cited Data Quality and Validation as the most challenging area they face today – a new finding in this year's survey. Notably, more than a quarter (27%) of respondents were unsure what (if any) Data Quality solution their organization is using. That percentage is higher (39%) for companies with low maturity DataOps practices.

The Rise of Data Catalogs and Data Discovery

The majority (60%) of organizations are now using Data Catalog and Data Discovery tools. Only 23% of respondents worked at organizations that do not have a data catalog or data discovery tool, while only 17% were unsure what (if any) solutions they had in these areas.

Introduction

The data engineering landscape is changing and maturing. Whereas years ago there were few, if any, tools to solve data challenges, a plethora of technologies – both commercial and open source – are now available.

In the past, the choice was simpler: adopt one of a few mature, off-the-shelf data technologies or build your custom solution. Now, data engineering and operations teams have a broader landscape of technologies to choose from, in addition to the option of outsourcing all data technologies to a third-party managed service. The challenge is making the right decisions for the long term amidst a growing number of data sources, users, platforms, and regulations.

Meanwhile, the move to the cloud is accelerating. Most organizations are either 100% cloud-based, or plan to be. The cloud's cost savings, performance, and extensibility benefits are simply too great for technical and business leaders to overlook. Another key reason for moving to the cloud is to remove the operational onus from data teams as they make technology decisions. Instead of having to train their operations team on new technologies, data teams are able to offload most, if not all, data operations to a cloud provider or vendor offering technology and solutions as cloud services. This dramatically lowers both time-to-value and time-to-production. Data architects and engineers no longer have to base decisions primarily on how difficult or expensive a new technology will be to implement and operate.

Databases, whether SQL or NoSQL, are using SQL as the lingua franca to inspire adoption and quick user uptake. As organizations look to expose data, SQL interfaces are becoming the preferred choice for data democratization. But teams that have only SQL skills miss out on advanced features or even entire technologies covered in this report. Programming languages like Python, Scala, and Java provide access to familiar abstractions (classes, functions) that let engineers decompose complex pipelines into manageable pieces with more complex logic. Open source engines like Apache Spark allow teams to handle the structured, semi-structured, and unstructured data that are common in many machine learning and AI applications.

We now have decades of data experience and deployments, yet complexity still exists as the data landscape evolves. While the technologies may change, the fundamental problems with data quality and other data issues remain the same. The one common denominator is the inherent complexity of data, particularly relative to data discovery, quality, and security. Though we have better tools to check and validate data, the notion of pristine "gold data zones' is still elusive for many data engineering teams.

The sheer number of technologies and regulations that modern data engineers must manage is a driver of increasing complexity. With more mature, advanced technologies, the onus falls on the data engineering team to choose and execute properly to achieve ROI on data initiatives and projects. As organizations adopt multiple cloud technologies and platforms, automation is growing in importance to streamline manual, risk-prone processes for many data engineers.

Yet, as the findings in this survey show, modern data engineering is also becoming easier in some ways. We have more mature technologies, for instance, with many miles on the codebase and clear, battle-tested use cases. These technologies are helping organizations leverage their sensitive data for real-time access and analytics, while protecting it in accordance with a growing body of regulatory requirements.

This report looks at areas of maturity and opportunity in the modern data landscape, to help data engineering and operations teams learn from their peers and make informed decisions about their data stacks to set them up for long-term success.

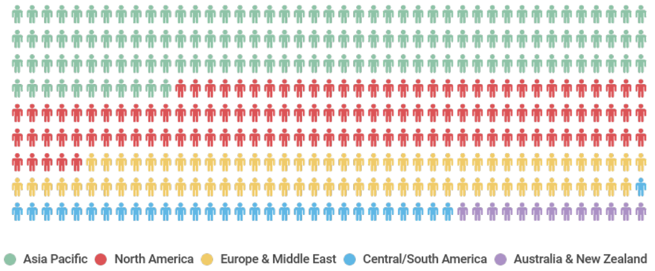
Demographics and Key Segments

The survey goal was to obtain a global perspective, with balanced participants across US, EMEA and APAC. Respondents primarily came from three regions: Asia-Pacific (37% of all respondents), North America (32%), and Europe & the Middle East (21%). About 40% of all respondents work in organizations with more than 1,000 employees.

Industry

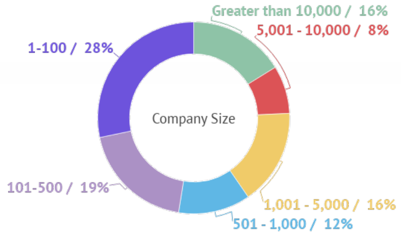
21%	Computers, Electronic, Technology
14%	Financial Services
10%	Education
8%	Healthcare
5%	Advertising
5%	Agriculture
4%	Automotive
3%	Manufacturing
3%	Telecommunications
3%	Food

Company Location



40%

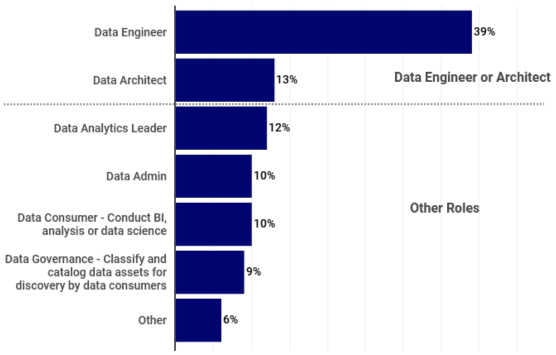
Respondents who work in organizations with more than 1,000 employees



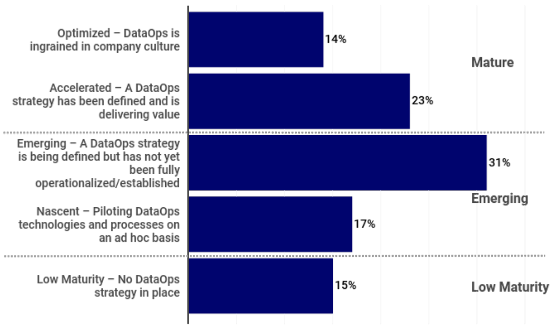
We asked respondents about their *organization's level of maturity with DataOps*, as well as their *primary role in using data for BI, analytics, and data science*. For the remainder of the report, we segmented findings using responses to these two questions:

1. **Job Role:** Alongside results for all respondents, we report results for respondents who are *Data Engineers or Data Architects* (52% of all respondents).
2. **Level of DataOps Maturity:** We report results for three unique segments: Mature, Emerging, and Low Maturity of DataOps.

Job Role



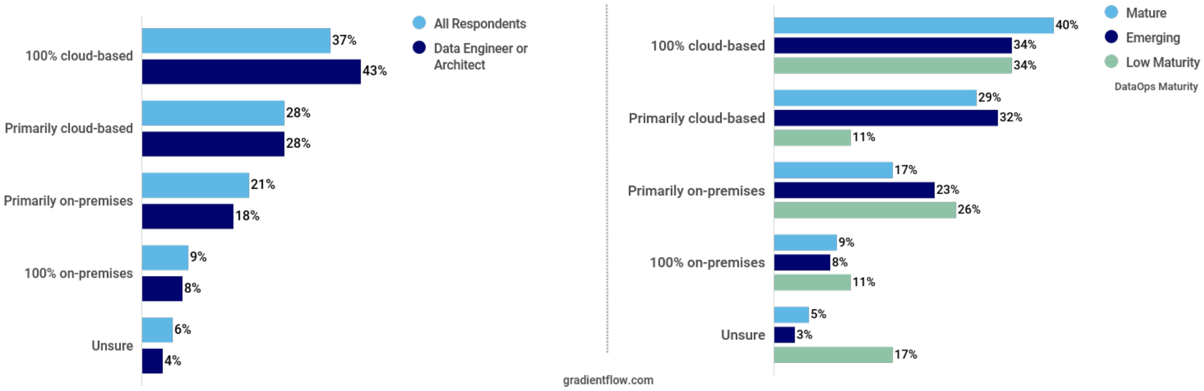
DataOps Maturity Level



Cloud Computing for Data Processing and Analytics

When it comes to data processing and storage for analytics and data science, approximately two-thirds of respondents (65%) characterized their company as already either *100% cloud-based* or *primarily cloud-based*. The share is even higher for *Data Engineers & Architects* (71%) and for respondents at companies with *Mature DataOps* capabilities (69%).

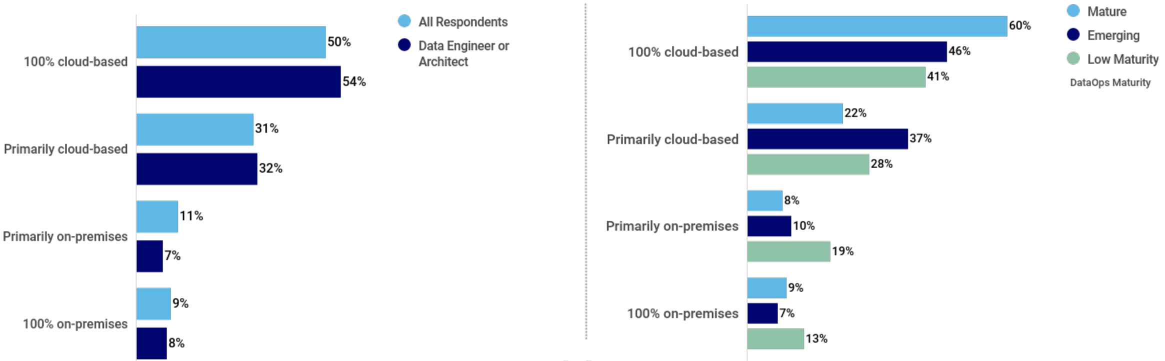
How would you describe how and where you process and store data today for BI, analytics and data science? (select one)



gradientflow.com

Looking ahead 12-24 months, respondents see their organizations moving faster toward cloud computing for data processing and storage: 81% of respondents project they will become 100% cloud-based or primarily cloud-based over that time frame. By comparison, the same 12-24 month projection in the 2021 Impact Report showed 71% of respondents becoming 100% cloud-based or primarily cloud-based, indicating an acceleration in the number of organizations planning to rely solely or primarily on the cloud in the coming months.

For BI, analytics and data science, what is your likely future state (within the next 12-24 months) with respect to using cloud-based software? (select one)



gradientflow.com

Data Engineering Tools and Platforms

The survey asked respondents to identify the tools they use, rate the criteria they used to evaluate data tools and platforms, and assess key challenges facing their data engineering teams.

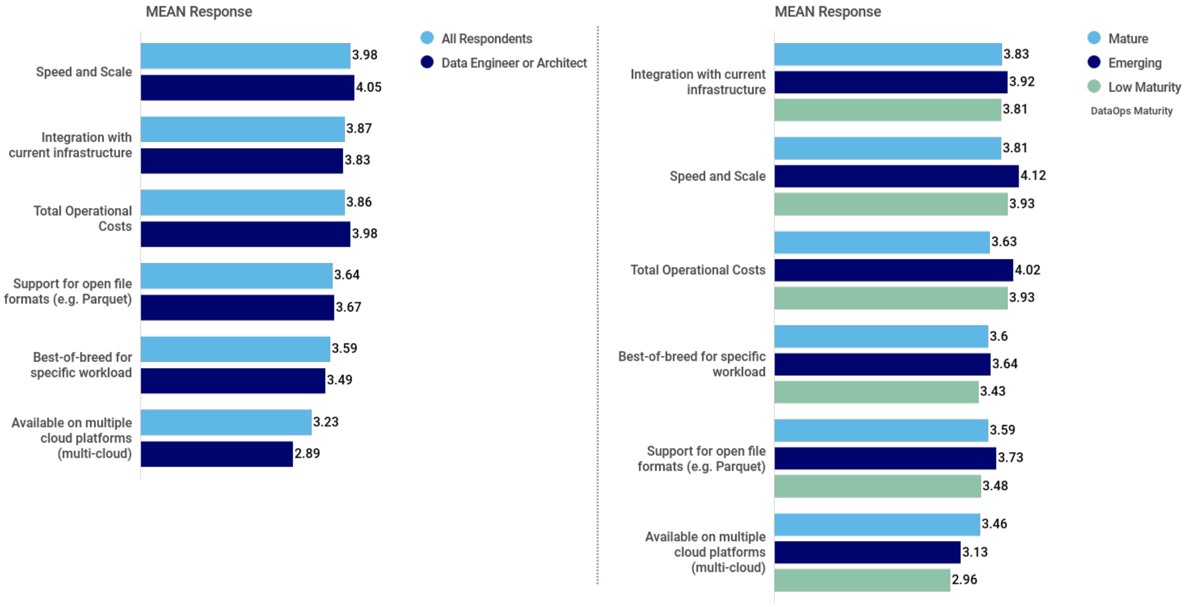
Criteria for Evaluating Databases and Data Platforms

Before we explore popular databases and data platforms, let's examine the importance of different criteria used to evaluate solutions. While companies are increasingly going multi-cloud, performance remains a key factor when choosing between cloud data platforms, warehouses, and data lakes. We found that respondents prioritized performance (speed and scale) over multi-cloud (*availability on multiple cloud platforms*).

We asked respondents to separately rate the importance of six key factors in choosing a platform. The following three factors emerged as the most important:

1. Speed and scale
2. Integration with current infrastructure
3. Total operational costs

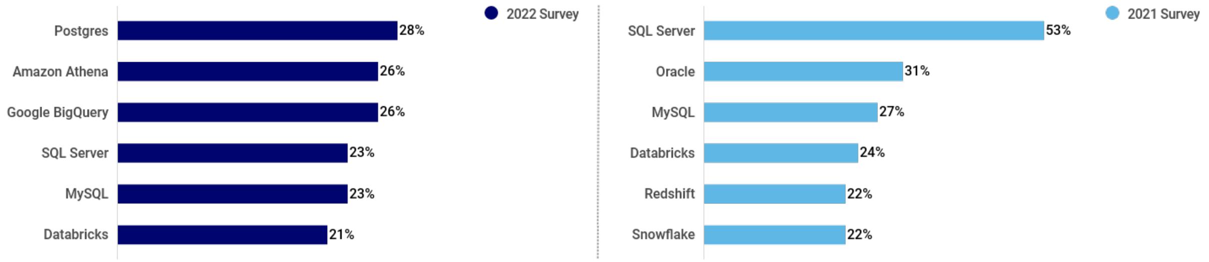
How important are the following when evaluating a database or cloud data platform? (1=Not important, 5=Very important)



Databases and Data Platforms

- **The top six databases and data platforms overall** were Postgres (28%), Amazon Athena (26%), Google BigQuery (26%), SQL Server (23%), MySQL (23%), and Databricks (21%). By comparison, the top six databases and data platforms overall in the 2021 Impact Report were SQL Server (53%), Oracle (31%), MySQL (27%), Databricks (24%,) Redshift (22%), and Snowflake (22%).

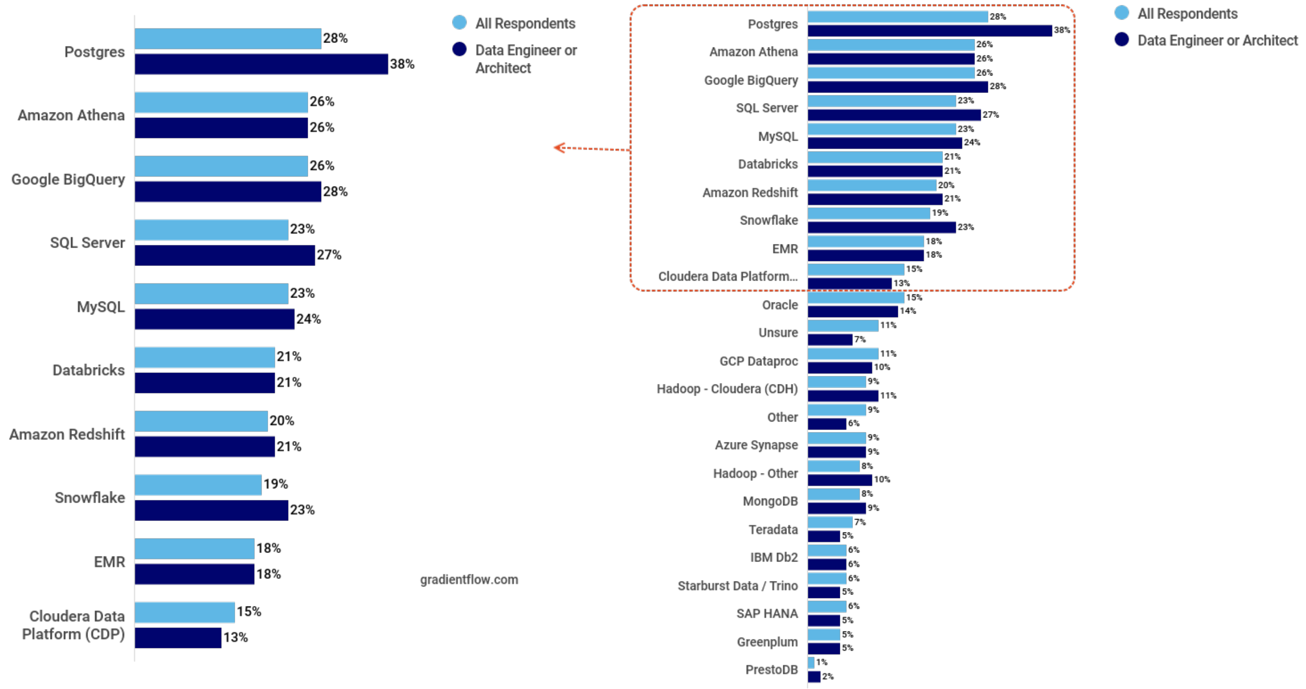
What database platforms or cloud data platforms are currently in production in your organization? (check all that apply)



gradientflow.com

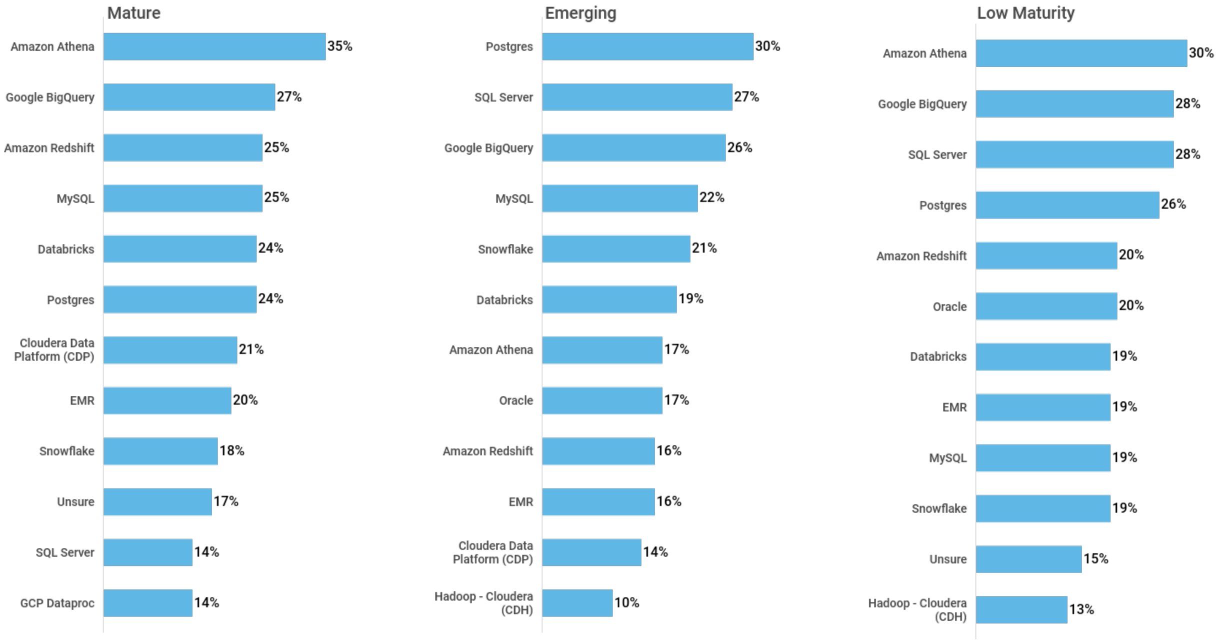
- **Among *Data Engineers or Architects*:** The top six databases and data platforms were Postgres, Google BigQuery, SQL Server, Amazon Athena, MySQL, and Snowflake.
- **Among companies with a Mature DataOps practice:** Cloud-managed services topped the list—Amazon Athena, Google BigQuery, and Amazon Redshift.

What database platforms or cloud data platforms are currently in production in your organization? (check all that apply)



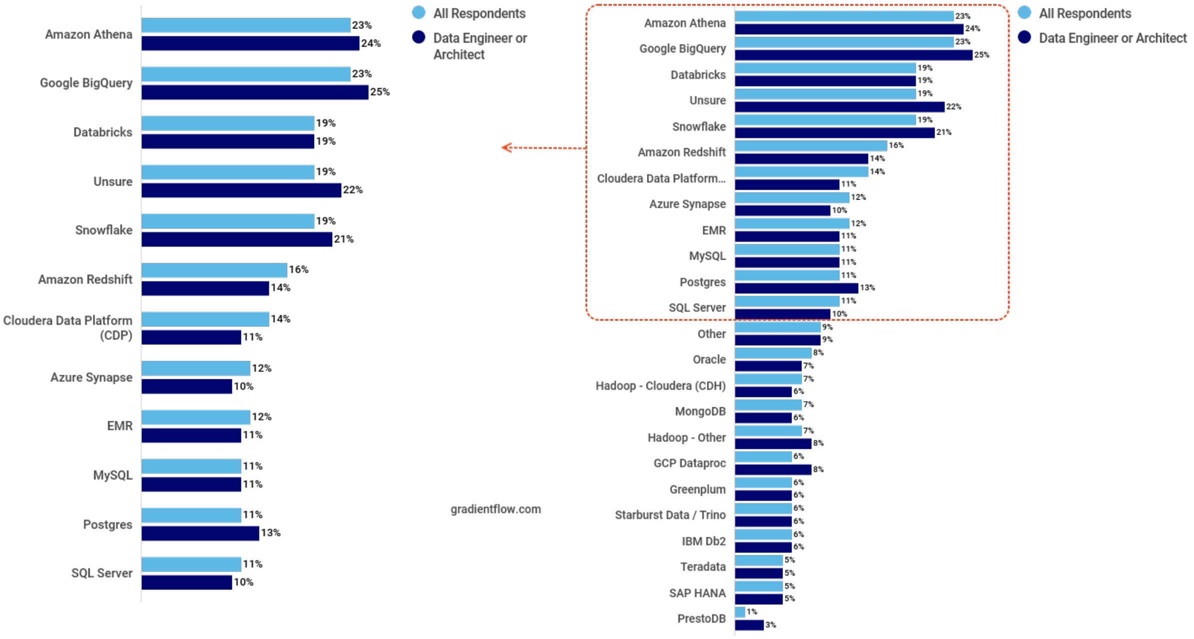
gradientflow.com

What database platforms or cloud data platforms are currently in production in your organization? (check all that apply)

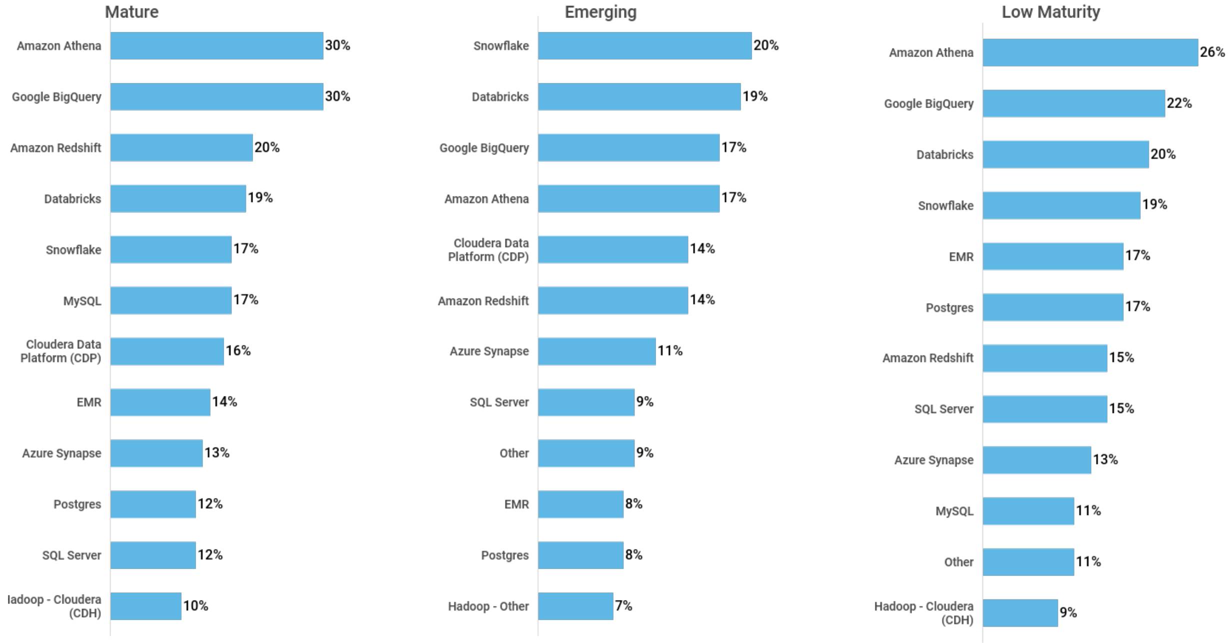


Looking ahead 12–24 months, 62% of respondents signalled they plan to adopt at least one of the following five cloud databases and platforms: Amazon Redshift, Amazon Athena, Google BigQuery, Databricks, and Snowflake. The previous year's survey also showed Amazon Athena, Databricks, and Snowflake as top platforms that respondents planned to adopt within the coming 12–24 months.

What database platforms or cloud data platforms are you likely to adopt within the next 12-24 months? (check all that apply)



What database platforms or cloud data platforms are you likely to adopt within the next 12-24 months? (check all that apply)



Key Data Engineering Challenges

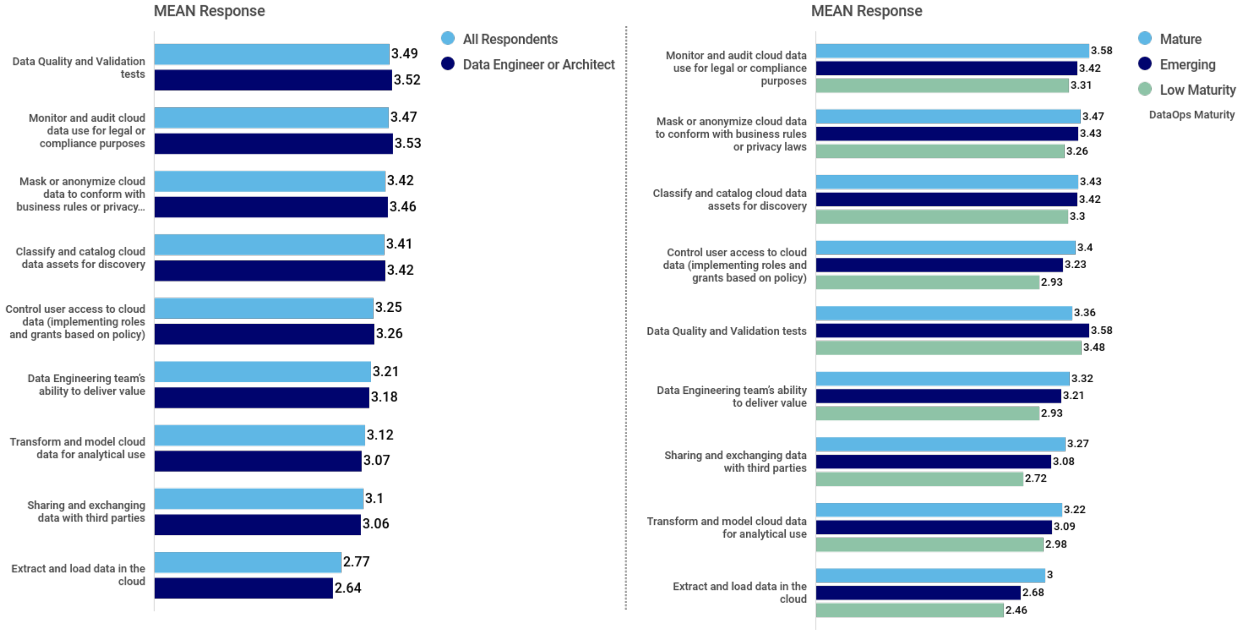
Before we dive further into adoption rates for tools and platforms, let's examine how respondents ranked key challenges faced by their data engineering and infrastructure teams.

We asked respondents to rate a range of items including Data Integration (ELT); Access Control, Security, and Privacy; Data Testing and Sharing; and their Data Engineering Teams.

Respondents cited these areas as the most challenging: Data Quality and Validation; Monitoring and Auditing for Compliance; Masking and Anonymization; and Data Discovery. This is relatively consistent with the 2021 Impact Report, which found that top challenges included Masking or Anonymizing Data, and Data Monitoring and Auditing.

Respondents—particularly those from companies with Mature DataOps practices—deemed tasks that are primarily performed after Data Integration (Extract and Load; Transform and Model) to be the least challenging item on the list. As we note in the next section, this may be because users now have access to many more data integrations tools and solutions. This is consistent with the previous year's findings.

To the best of your ability, rate the level of challenge involved in each of the following steps in using cloud platforms for analytics. (1=Not challenging, 5=Extremely challenging)



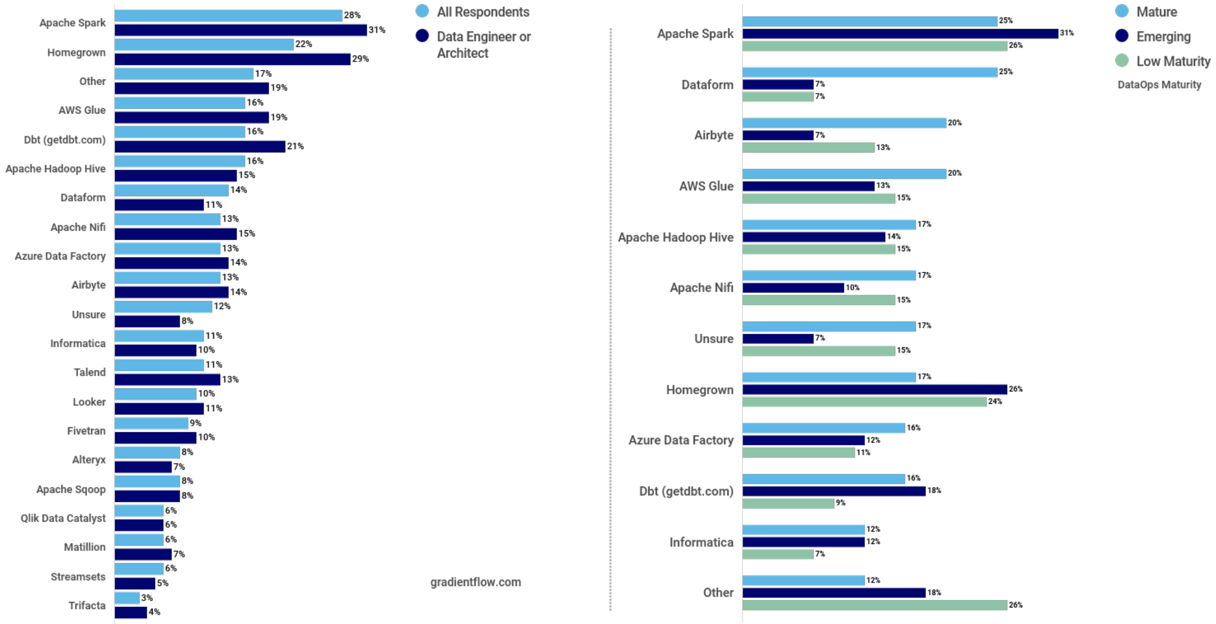
Data Integration and Data Orchestration

A new set of startups and open source projects have fueled a resurgence of interest in data integration within the data engineering community.

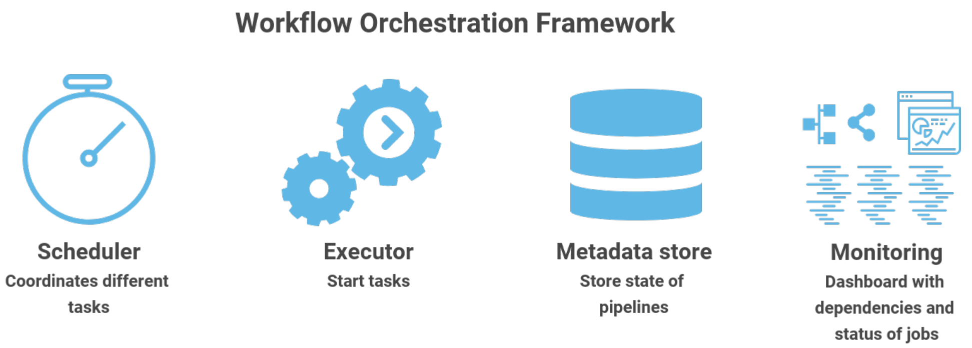
This new wave of solutions comes at a time when companies have to handle more data sources and data types, such as unstructured data consisting of text, images, audio, and video. Data now powers many companies' important analytic and AI products, as well as their applications. Modern data integration and data orchestration solutions are relatively mature and aim to help companies inject more engineering rigor into their vast array of data pipelines, in order to help manage the constant flow of data and keep pace with the demand for real-time data access and use.

- **Among all survey respondents:** Top data integration solutions include popular data engineering tools like Apache Spark, dbt, and Hive, as well as managed services like AWS Glue, Dataform, and Azure Data Factory.
- **Among companies with a Mature DataOps practice:** The new open source project Airbyte joins the top five data integration tools, along with Apache Spark, Dataform, AWS Glue, and Hive.
- **"Other"** tools mentioned: Prefect, Apache Kafka, Apache Airflow, and SQL Server Integration Services.

What tools does your organization use to integrate data (extraction, loading/ingestion, transformation, modeling, master data management) for BI, analytics and data science? (check all that apply)



Building, deploying, and managing pipelines are critical for companies that depend on an array of data and AI applications. A class of workflow management/orchestration solutions have emerged to help companies manage their growing collection of pipelines.

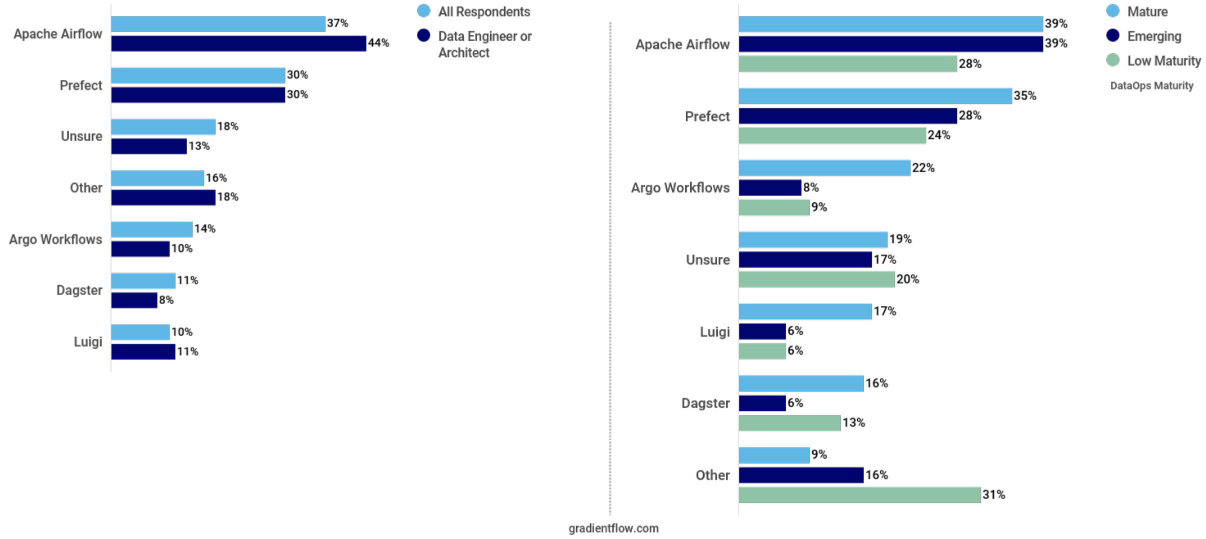

Workflow Orchestration Framework
Scheduler
Coordinates different tasks
Executor
Start tasks
Metadata store
Store state of pipelines
Monitoring
Dashboard with dependencies and status of jobs

Caption: Key components of workflow management and orchestration frameworks.

We asked respondents to select from a list of open source workflow orchestration solutions:

- Apache Airflow, an early implementation of “workflows-as-code,” was the most popular option among all respondents. A newer open source project called Prefect finished a strong second behind Airflow. Prefect was started by early contributors to Airflow, and designed to address its shortcomings for modern data applications.
- **Among companies with a Mature DataOps practice:** The third most popular option was Argo workflows, a Kubernetes-native framework.
- **“Other”** tools mentioned: Temporal.io, AWS Step Functions, and Azure Data Factory.

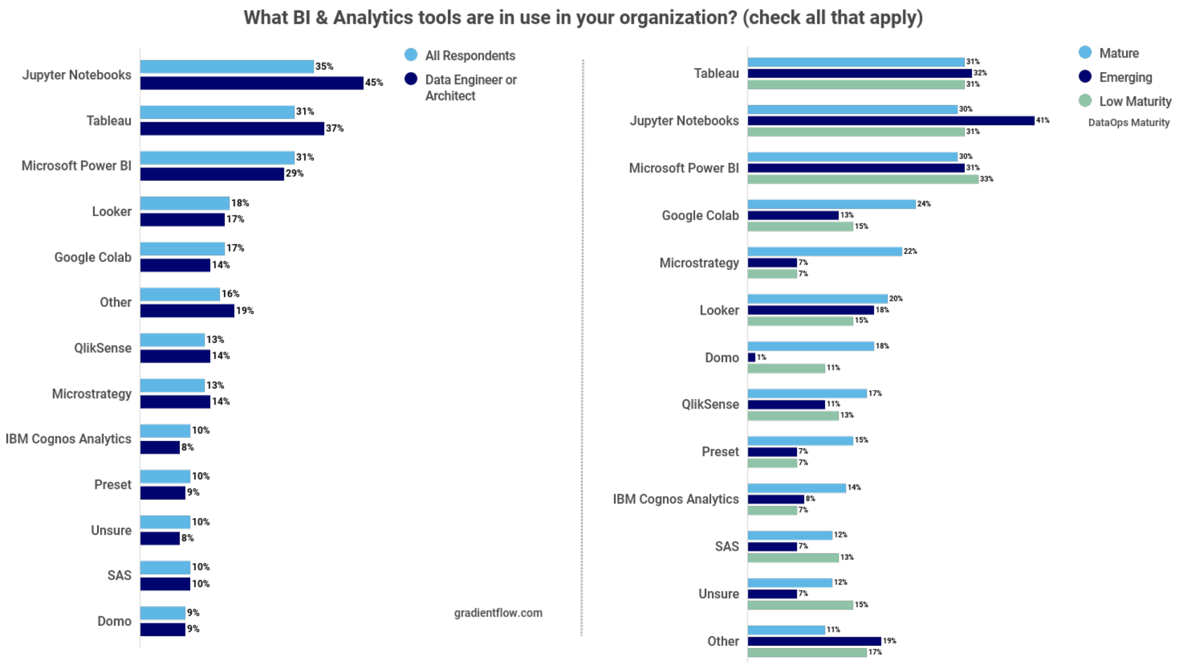
What tools does your organization use for workflow management and orchestration? (check all that apply)



BI and Analytics

Along with Data Integration and Orchestration, BI and Analytics is one the more mature categories covered in this survey. It's not hard to see why – many companies have had solutions in place to support BI and analytics initiatives for decades. This head start has allowed BI and Analytics solutions, such as tools for creating static charts, interactive dashboards, and advanced analytics, to become quite advanced and widely adopted.

- The top three most popular solutions across all respondents were Jupyter Notebooks, Tableau, and Microsoft Power BI.
- **Among companies with a Mature DataOps practice:** Nearly a quarter (24%) of respondents said they use Google Colab.
- **"Other"** tools mentioned: Metabase, TIBCO Spotfire, and Sisense.



These findings support the notion that tools designed for data pipelines' starting points (ELT) and end points (BI and Analytics) are relatively mature. Meanwhile, solutions meant for the middle of these pipelines – where data quality and security exist – are less mature, and the environment is becoming highly complex due to increasing numbers of data sources, users, and real-time apps, sensitive data use, and regulatory requirements. Thus, the challenges associated with data quality and security are becoming more acute for data engineers, while tasks related to ELT and BI/Analytics are perceived as less challenging.

Data Catalog and Data Discovery

As organizations grow the amount of raw and derived data they generate and store, users need tools to help them discover data resources. For example, when a data catalog is added to the data stack, data discovery is needed. Data catalogs and data discovery tools provide answers to many questions, including:

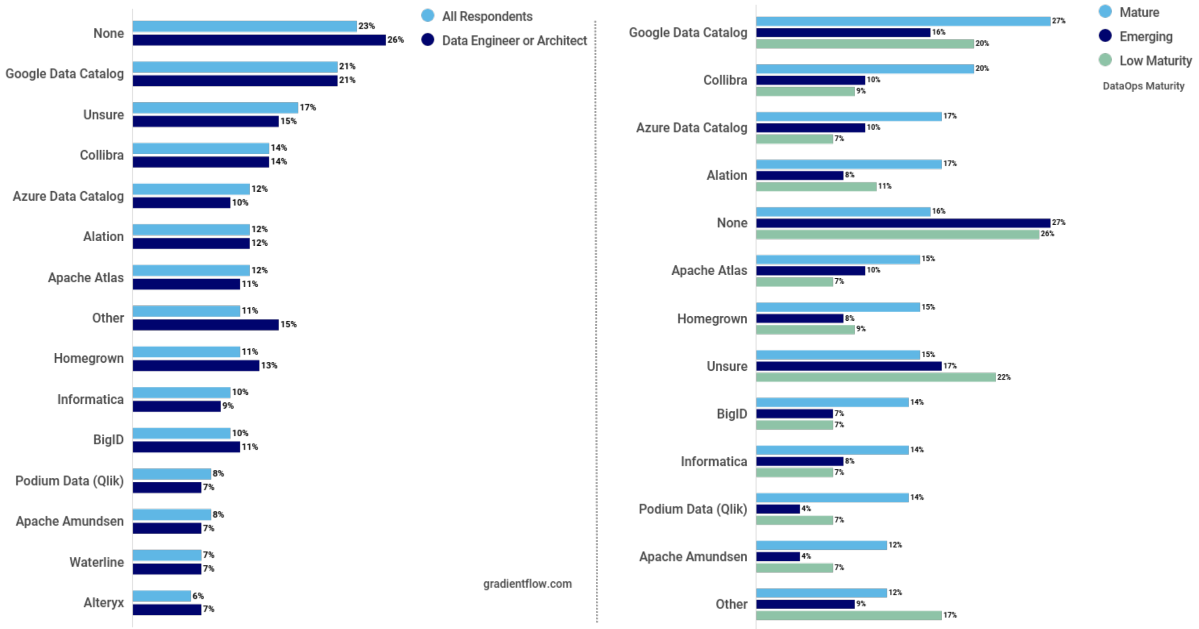
- Does the data needed to build this application already exist?
- If the data exists, who owns this data, who created it, and do we have access to it?

Having the proper tools in place can free up time for data scientists and other users. For example, Lyft estimated that prior to the rollout of their internal data discovery tool, their data scientists spent 25% of their time on data discovery.

Our survey indicates that data catalog and discovery tools are maturing and becoming more heavily adopted year-over-year. About a quarter (23%) of all respondents said their organizations don't have a data catalog or a data discovery tool. An additional 17% were Unsure what (if any) solutions they had in these areas—which means they are essentially not using data discovery solutions, even if they are available. Still, this means that well over half (60%) of respondents are investing in data catalog and discovery tools to enable self-service data use.

- The top three solutions across all user segments were Google Data Catalog, Collibra, and Azure Data Catalog.

What Data Catalog and Data Discovery tools are in use in your organization? (check all that apply)



gradientflow.com

Data Quality

As survey respondents cited Data Quality as their most challenging area, it's no surprise that it is an area that has seen an influx of startups in recent years.

A recent article identifies four key reasons why data quality has emerged as a top-level concern among companies:

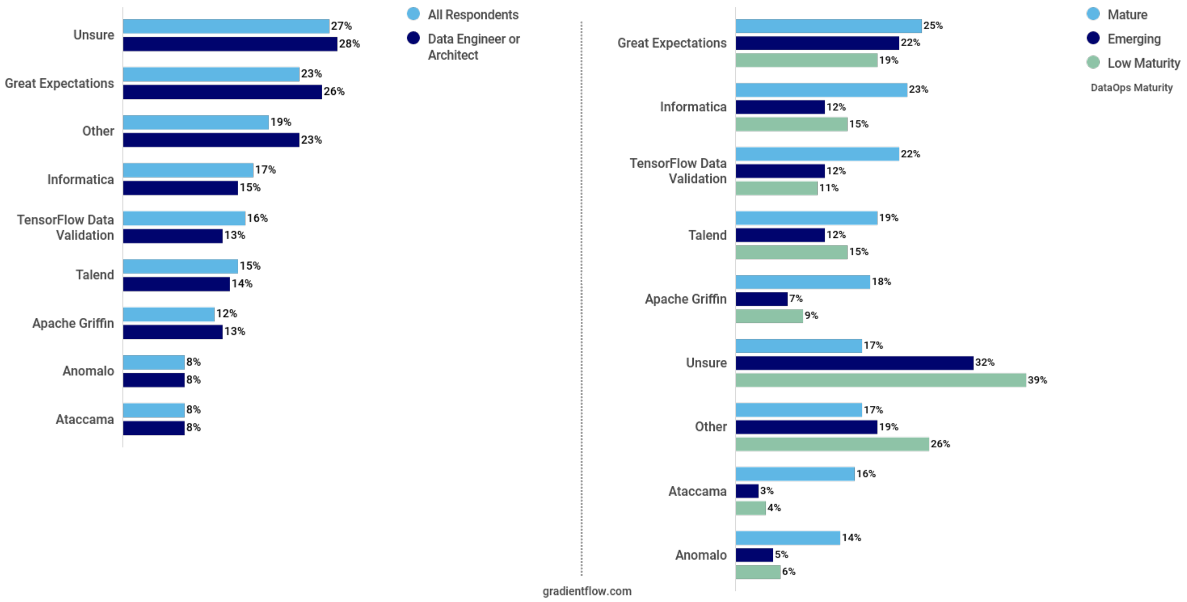
1. **More teams rely on data.** Users need to trust data products and services to facilitate adoption of data, analytics, and machine learning.
2. **Potential sources of error have increased.** The volume, variety, and velocity of data continue to increase, along with the number and types of data sources and providers.
3. **Data quality issues impact critical services and products.** A rapidly growing number of real-world applications for essential services and products depend on accurate data.
4. **Data architectures have become more complex.** The need for enhanced speed, advanced capabilities (ML and AI), and security in data-driven services and products is driving the growing complexity in data architecture systems.

Given the growing importance of data and AI products and services, companies need to tackle data quality systematically, holistically, and proactively. This means being able to address data quality issues before they impact critical products and services.

Yet, more than a quarter (27%) of respondents were Unsure what (if any) data quality solution their organization is using. This number is even higher (39%) for companies with Low Maturity DataOps practices.

- **Among companies with a Mature DataOps practice:** The top three tools were Great Expectations, Informatica, and TensorFlow Data Validation.
- **"Other"** tools mentioned: AWS Deequ and dbt.

What tools does your organization use to improve data quality? (check all that apply)

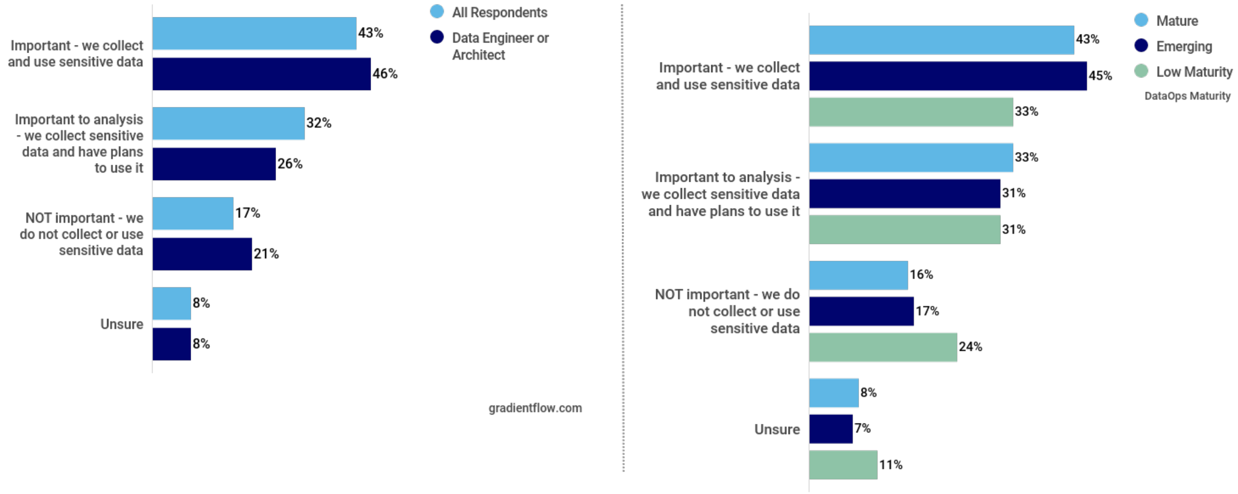


gradientflow.com

Data Privacy and Security

The final section of this report describes how survey respondents handle data security and privacy controls, amidst an increasing amount of sensitive data and more complex privacy regulations, with which they must comply. Three-quarters (75%) of survey respondents report that their organizations already collect and store sensitive data. The results from the 2021 Impact Report indicate a consistent YoY trend in the use of sensitive data.

What best describes how your organization handles sensitive data - defined as any data that cannot or should not be seen by all analysts or data scientists - within BI/analytics and data science today? (select one)

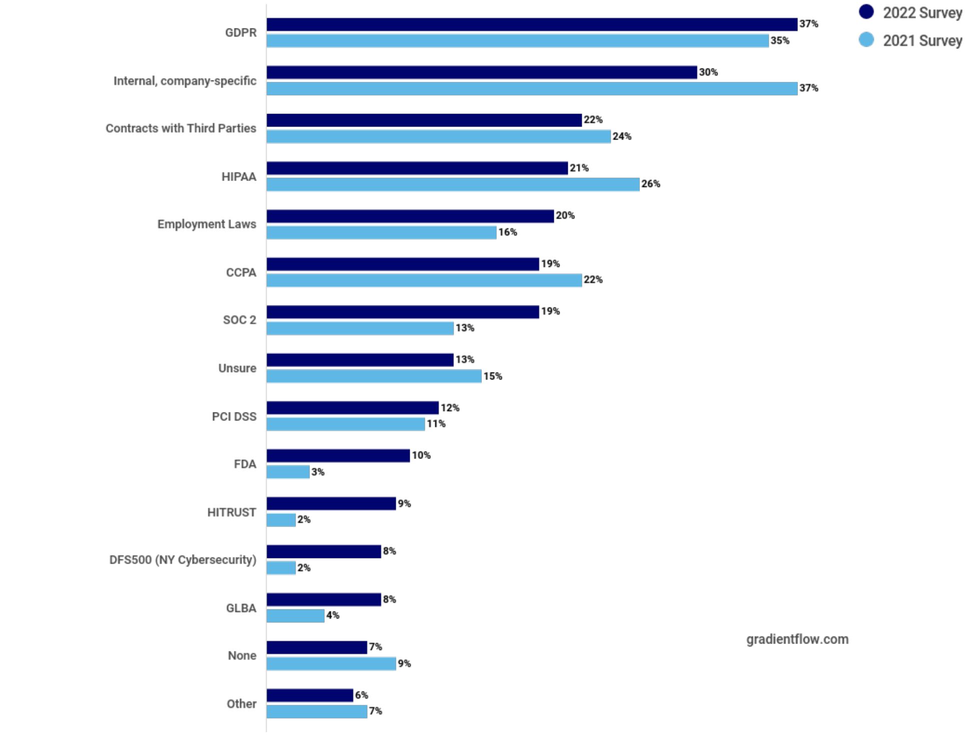


We also asked respondents which privacy rules or regulations for sensitive data they must adhere to. A vast majority of survey respondents (88%) indicated their organizations are subject to *one or more* data use rules or regulations.

- The top four regulations cited were GDPR (37%), HIPAA (21%), CCPA (19%), and SOC 2 (19%). Close to one-third of all respondents (30%) cited a need to comply with *"Internal, company-specific"* rules.

In the 2021 Impact Report, the top four external regulations cited were the same: GDPR, HIPAA, CCPA, and SOC 2. Notably, in the 2021 survey 37% of respondents reported a need to comply with internal, company-specific rules, vs. just 30% in the 2020 survey – the largest increase among any data use rules cited.

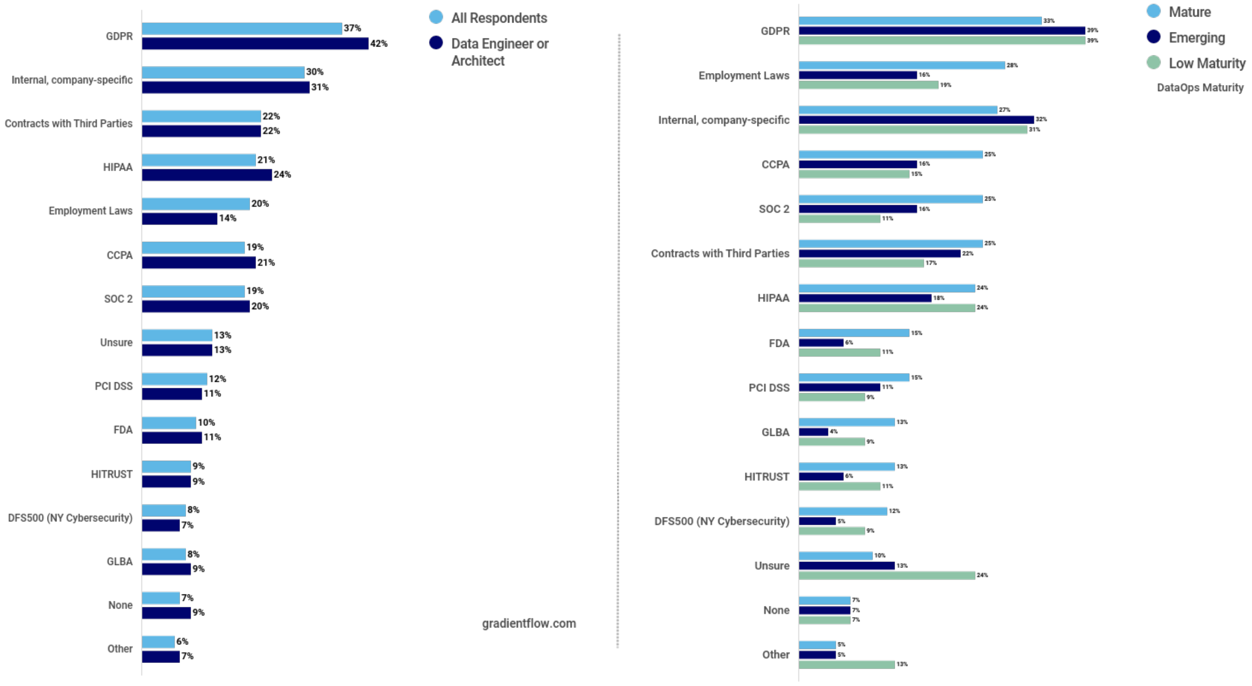
What rules must your organization comply with when sensitive or personal data is used for analytics or data science? (check all that apply)



- **Among companies with a Mature DataOps practice:** The second most popular option cited was *"Employment Laws"*.

- **"Other"** rules/regulations mentioned: *LGPD* (Brazilian General Data Protection Law).

What rules must your organization comply with when sensitive or personal data is used for analytics or data science? (check all that apply)



Closing Thoughts

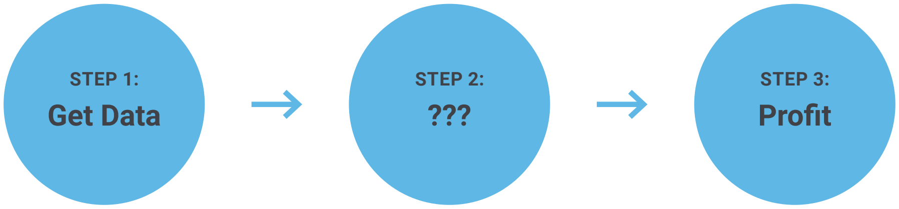
This report aims to help data teams understand and learn from the choices others are making, and the challenges they face as the data landscape continues its rapid evolution.

It can be easy for data engineering teams to make critical technology decisions based not on peer insights, but on factors like product marketing direction or familiarity with certain platforms or technologies. This guide should help inform those decisions based on specific organizational needs and data goals, and how peer organizations are building their next-generation data and analytics stacks.

As we look over the list of what is really causing issues in the data supply chain, we see problems are both people- and technology-based. Therefore, the solution should take both into account; data engineering teams should have a holistic approach to data product creation, as well as adequate human and technological resources to accomplish their goals.

As two years' worth of data engineering survey data now shows, the start and end points of the data spectrum are the least challenging, primarily because ELT and BI/Analytics tools are relatively mature and widely adopted. Meanwhile, the "in-between" processes – specifically, data cataloging, security, and quality – are becoming more complex amidst a rapidly evolving data landscape, and the tools to streamline and scale these processes remain comparatively immature.

This image illustrates a common approach to data strategy that falls in line with this phenomenon:



Organizations often focus on Steps 1 and 3, and assume that the "in-between" Step 2 can be solved through the latest technology, like ML and AI. Instead, they need a clear, actionable data strategy that is communicated to all stakeholders and takes into account the realities of today's complex, fast-paced data environment. Without a plan, data teams find themselves in a data purgatory of: "We have all of this data. Now what?" The "now what" usually translates into a lack of progress and value creation.

Our suggestion to data teams on executing their strategy is: "Eat your elephant one bite at a time, instead of trying to eat it in one bite." This will create manageable scale and velocity for data teams, providing both quick data wins and a longer-term roadmap for data-driven innovation. Prioritizing the aspects of the data strategy and pipeline that are underdeveloped is one way to impact results of DataOps initiatives.

The continuing maturity of the data engineering landscape is clear. Vendors and organizations today are realizing that the demands of scale, speed, and varied use cases may require multiple databases. Data teams that invest in and take full advantage of the right resources and solutions will be better able to outperform competitors with data, and will be prepared for the ever-changing future of data use.

Acknowledgements

Thanks to Immuta for sponsoring the Data Engineering Survey. Thanks to Kathy Yu and the Big Data Institute for providing critical assistance. This survey was conducted by **Gradient Flow**; see our Statement of Editorial Independence.

About Immuta

Immuta is the universal cloud data access control platform, providing data engineering and operations teams one platform to control access to analytical data sets in the cloud. Only Immuta can automate access control for any data, on any cloud service, across all compute infrastructure. Data-driven organizations around the world rely on Immuta to speed time to data, safely share more data with more users, and mitigate the risk of data leaks and breaches. Founded in 2015, Immuta is headquartered in Boston, MA. Learn more at www.immuta.com.

About Gradient Flow

Gradient Flow presents a rich array of high quality content on data, technology, and business, with a focus on machine learning and AI. Named by Coursera as one of the Top 10 Sites for Data Scientists, Gradient Flow helps you stay ahead on the latest technology trends and tools with in-depth coverage, analysis, and insights."

GRADIENT
FLOW

gradientflow.com

IMMUTA™

immuta.com | (800) 655-0982