February 28, 2023

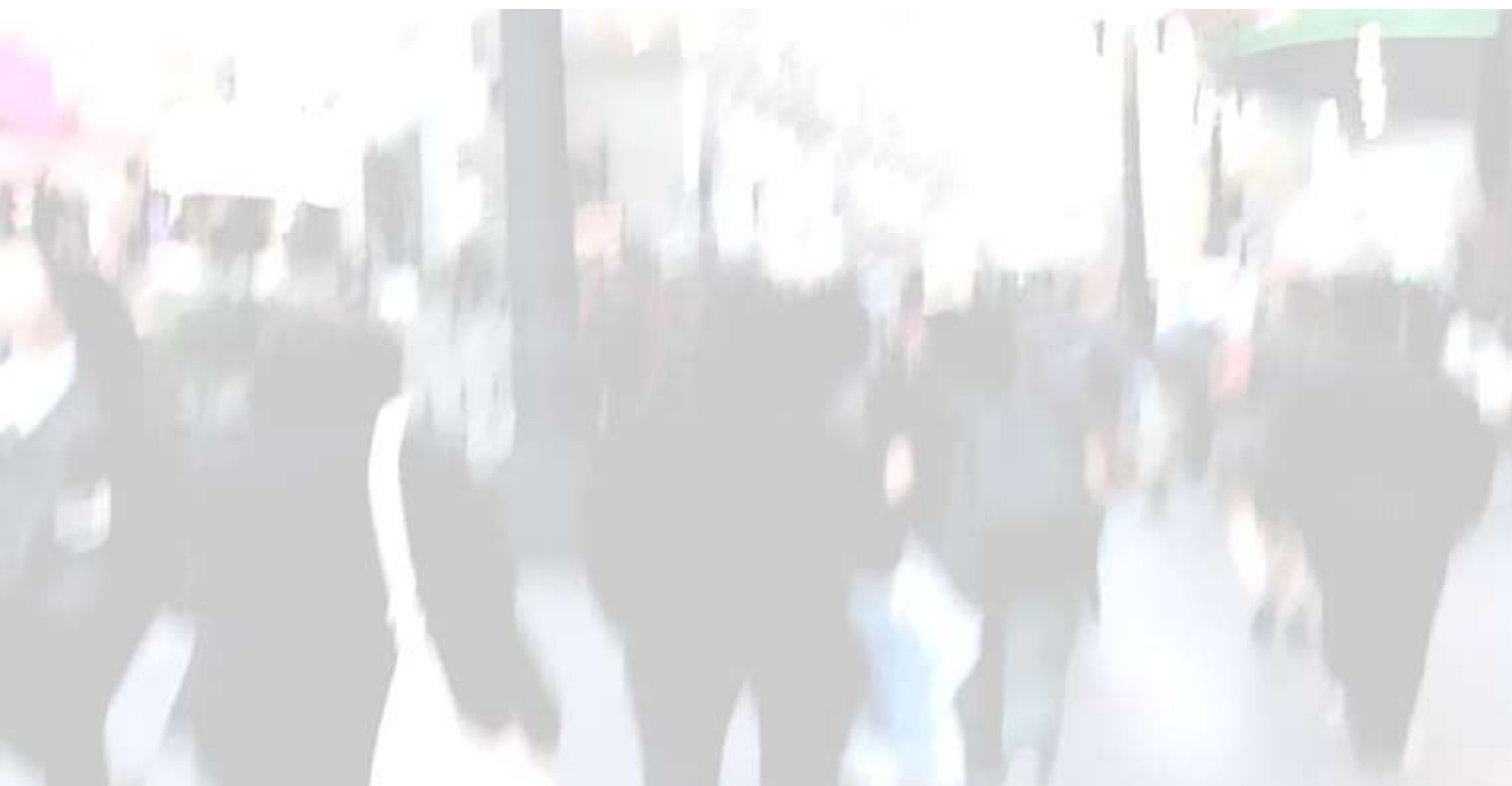
Dresner Advisory Services, LLC

2023 Edition

Data Engineering Market Study

Wisdom of Crowds® Series

Licensed to Alteryx



Disclaimer

This report should be used for informational purposes only. Vendor and product selections should be made based on multiple information sources, face-to-face meetings, customer reference checking, product demonstrations, and proof-of-concept applications.

The information contained in all Wisdom of Crowds® Market Study reports reflect the opinions expressed in the online responses of individuals who chose to respond to our online questionnaire and does not represent a scientific sampling of any kind. Dresner Advisory Services, LLC shall not be liable for the content of reports, study results, or for any damages incurred or alleged to be incurred by any of the companies included in the reports as a result of its content.

Definitions

Business Intelligence Defined

Business Intelligence (BI) is "knowledge gained through the access and analysis of business information."

Business Intelligence tools and technologies include query and reporting, OLAP (online analytical processing), data mining and advanced analytics, end-user tools for ad hoc query and analysis, and dashboards for performance monitoring.

Definition source: Howard Dresner*, The Performance Management Revolution: Business Results Through Insight and Action* (John Wiley & Sons, 2007)

Data Engineering Defined

Data engineering is the best practices and technology capabilities to develop engineered data workflows and pipelines to and between operational and analytic data management infrastructures. Data engineering includes the requirements and priorities for data orchestration, integration, and transformations including advanced analytics in the data engineering pipeline workflow.

Data engineering development and deployment capabilities are used to develop, debug, schedule, secure, govern, and run data workflows for BI / analytic use cases. Deployments may be use-case specific (i.e., small number of users doing data science projects with data sources/transformations required) or may be required to be fault tolerant, highly secure, and scalable to span large data volumes, multiple analytic steps, multiple analytic models, and support multiple BI use cases and tools.

Introduction

In 2023, we mark the 16th anniversary of Dresner Advisory Services. We are thankful for the support and encouragement of our clients and related communities. This has allowed us to build a stellar analyst organization and create world-class market research focused exclusively upon data, analytics, business intelligence, performance management, and associated topics.

Last year, in support of our members, we published over 3,500 pages of independent and objective primary research across 20 different Flagship and thematic market reports, 50 Research Insights (thought leadership articles), and 55 Vendor Insights reports. As in previous years, we remain committed to creating the most in-depth, and relevant research available for these domains.

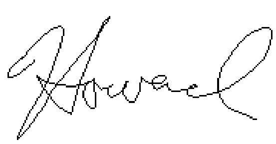
Recent global events and economic conditions created greater levels of uncertainty and seem to be delaying any perceived return to “normal” business conditions. Our data show that strategically leveraging data can help organizations manage through difficult times and be more successful.

As organizations face these business “headwinds” in 2023 and beyond, data engineering will be crucial to develop, manage, monitor, and govern data pipelines in support of a growing number of analytical use cases.

In its second year of publication, this report continues to explore market requirements and priorities for data orchestration, integration, and transformations, including advanced analytics in the data engineering pipeline workflow.

We hope you enjoy this report!

Best,



Chief Research Officer
Dresner Advisory Services

Contents

Benefits of the Study

The 2023 Dresner Advisory Services Data Engineering Market Study provides a wealth of information and analysis, offering value to consumers and producers of business intelligence technology and services.

Consumer Guide

As an objective source of industry research, consumers use the Dresner Advisory Data Engineering Market Study to understand how their peers leverage and invest in collaborative BI and related technologies.

Using our unique vendor performance measurement system, users glean key insights into BI software supplier performance, which enables:

- Comparisons of current vendor performance to industry norms
- Identification and selection of new vendors

Supplier Tool

Vendor licensees use the Dresner Advisory Services Data Engineering Market Study in several important ways:

External Awareness

- Build awareness for business intelligence markets and supplier brands, citing Dresner Advisory Services Data Engineering Market Study market trends and vendor performance.
- Gain lead and demand generation for supplier offerings through association with Dresner Advisory Services Data Engineering Market Study brand, findings, webinars, etc.

Internal Planning

- Refine internal product plans and align with market priorities and realities as identified in the Dresner Advisory Services Data Engineering Market Study
- Better understand customer priorities, concerns, and issues
- Identify competitive pressures and opportunities

About Howard Dresner and Dresner Advisory Services

The Dresner Advisory Services Data Engineering Market Study was conceived, designed, and executed by Dresner Advisory Services, LLC, an independent advisory firm, and Howard Dresner, its president, founder, and chief research officer.

Howard Dresner is one of the foremost thought leaders in business intelligence and performance management, having coined the term "Business Intelligence" in 1989. He has published two books on the subject, *The Performance Management Revolution – Business Results through Insight and Action* (John Wiley & Sons, Nov. 2007) and *Profiles in Performance – Business Intelligence Journeys and the Roadmap for Change* (John Wiley & Sons, Nov. 2009). He lectures at forums around the world and is often cited by the business and trade press.

Prior to Dresner Advisory Services, Howard served as chief strategy officer at Hyperion Solutions and was a research fellow at Gartner, where he led its business intelligence research practice for 13 years.

Howard has conducted and directed numerous in-depth primary research studies over the past two decades and is an expert in analyzing these markets.

Through the Wisdom of Crowds® Business Intelligence market research reports, we engage with a global community to redefine how research is created and shared. Other research reports include:

- Wisdom of Crowds® Flagship BI Market Study
- Analytical Data Infrastructure
- Cloud Computing and Business Intelligence
- Data Catalog
- Data Science and Machine Learning
- Embedded Business Intelligence
- Master Data Management
- ModelOps
- Self-Service BI

You can find more information about Dresner Advisory Services at www.dresneradvisory.com.

About Bill Hostmann

Bill Hostmann is a VP and Research Fellow with Dresner Advisory. His area of focus includes trends in Analytic Data Infrastructures (ADI)—integrating and managing the information and information models used by BI, Advanced Analytics, and CPM/PM applications.

Bill has more than 20 years of experience at the intersection of business intelligence / analytics and data analytics infrastructure, including positions in product and general management at Gemstone Systems, Informix, and Informatica.

He spent 14 years as a research analyst at Gartner, including several years as a VP and Distinguished Analyst for BI/Analytics. Bill's education includes a BS in Electrical Engineering, an MS in Computer Science and Engineering and an MBA.

The Dresner Team

About Elizabeth Espinoza

Elizabeth is Research Director at Dresner Advisory and is responsible for the data preparation, analysis, and creation of charts for Dresner Advisory reports.

About Kathleen Goolsby

Kathleen is Senior Editor at Dresner Advisory ensuring the quality and consistency of all research publications.

About Danielle Guinebertiere

Danielle is the Director of Client Services at Dresner Advisory. She supports the ongoing research process through her work with executives at companies included in Dresner market reports.

About Michelle Whitson-Lorenzi

Michelle is Client Services Manager and is responsible for managing software company survey activity and our internal market research data.

Survey Method and Data Collection

As with all our Wisdom of Crowds® Market Studies, we constructed a survey instrument to collect data and used social media and crowdsourcing techniques to recruit participants.

We include our research community of over 6,000 organizations as well as crowdsourcing and vendors' customer communities.

Data Quality

We carefully scrutinized and verified all respondent entries to ensure that only qualified participants are included in the study.

Executive Summary

Executive Summary

- Seventy-seven percent of our survey respondents indicate data engineering is critical (fig. 5) or very important. This is up from 61 percent in our 2022 report. This increase in importance of data engineering reflects the increased demand for access to analytical data as well as the increasing fragmentation of data sources and targets. Less than 2 percent indicate that data engineering is not important.
- Sixty-five percent of respondents say their organizations use data engineering capabilities today, compared to 63 percent in our 2022 report. Twenty-seven percent of respondents indicate plans to use data engineering tools within the next 12 or 24 months, indicating a growing demand for data engineering in 2023 (fig. 10).
- Organizations often purchase and use data engineering tools for multiple BI use cases (fig 18). Most respondents say they employe their data engineering capabilities for data integration, cleansing, and transformation workflows for data warehouses supporting dashboards and reporting. The lowest use case for data engineering is the re-platforming/replication of existing data warehouses.
- The top features for data engineering workflows include the *ability to aggregate and group data and ETL and ELT workflows* as well as the management of engineering workflow such as *alerting/job monitoring* and *execution plan, time, and event-based schedulers for jobs*. Least important is support for Kafka and Apache big data services (fig 30).
- Overall, respondents indicate that *relational databases*, *file systems*, and *applications* are the top data sources for data engineering workflows (fig 46). Of note is *object stores* (e.g., Amazon S3, Google Cloud Storage, etc.), which are the predominant targets for data engineering workflows. This indicates their role as an offline storage and staging repository for data engineering workflows. Most organizations (35 percent) *occasionally* enrich their data engineering workflows with third-party data. Eleven percent *constantly* enrich with third-party data, and 26 percent *frequently* enrich their data sets with third-party data as part of their data engineering workflows. Only seven percent of organizations never use third-party data for enrichment (fig 47).
- Twenty six percent of our survey respondents rate their current approach to data engineering as highly effective (fig. 52). Fifty-seven percent rate their current approach to data engineering as somewhat effective. Eighteen percent of respondents rate their current approach as somewhat or totally ineffective.
- The level of effectiveness with data engineering approaches can materially affect the overall success of a BI initiative (fig. 56).

Study Demographics

Study Demographics

Study participants provide a cross-section of responses across geographies, functions, organization sizes, and vertical industries. We constructed cross-tab analyses using the data to identify and illustrate important industry preference, priorities, and trends.

Geography

North America, which includes the United States, Canada, and Puerto Rico, represents the largest group with 60 percent of all respondents, followed by EMEA (23 percent). Asia Pacific and Latin America account for the balance (17 percent) of respondents (fig.1)

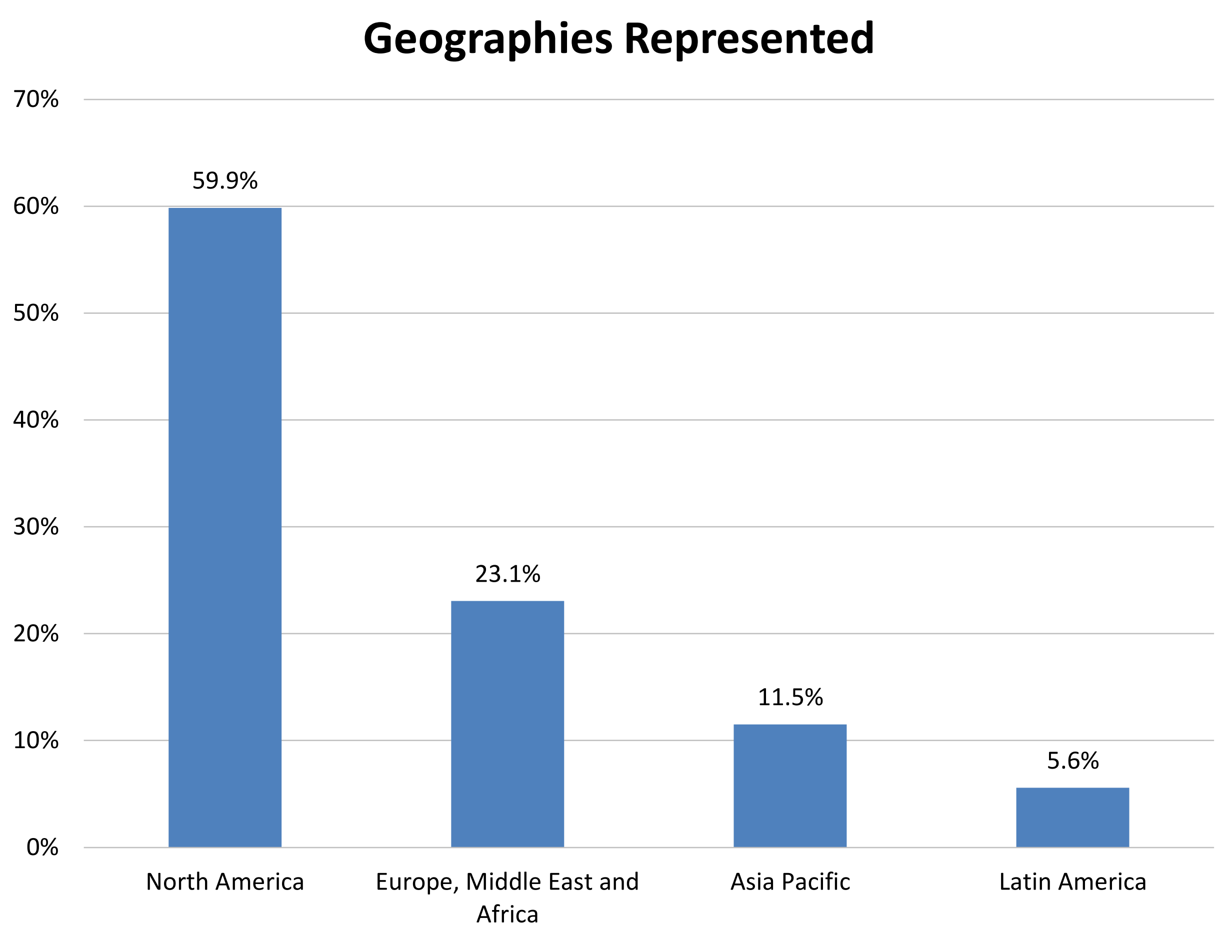


Figure 1 – Geographies represented

Functions

In 2023, respondents from IT functions account for 38 percent of our sample, followed by the Business Intelligence Competency Center (BICC) at 18 percent of the respondents (fig. 2). Participants from several business functions responded, including R&D, finance, operations, and executive management.

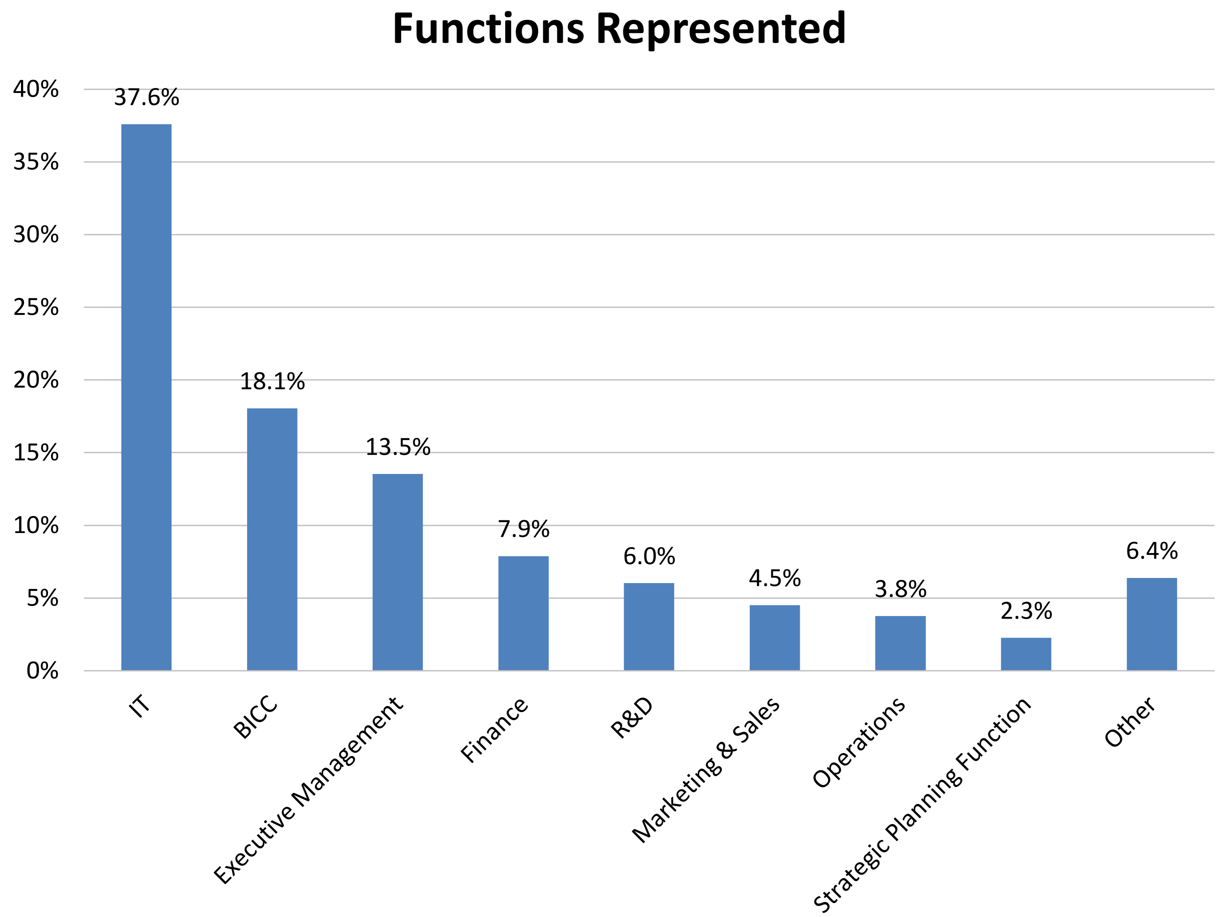


Figure 2 – Functions represented

Vertical Industries

In 2023, business services lead vertical industry distribution of respondents (fig. 3), followed by technology (17 percent), manufacturing (16 percent), financial services (15 percent), and consumer services (9 percent) as the most represented industries in the sample.

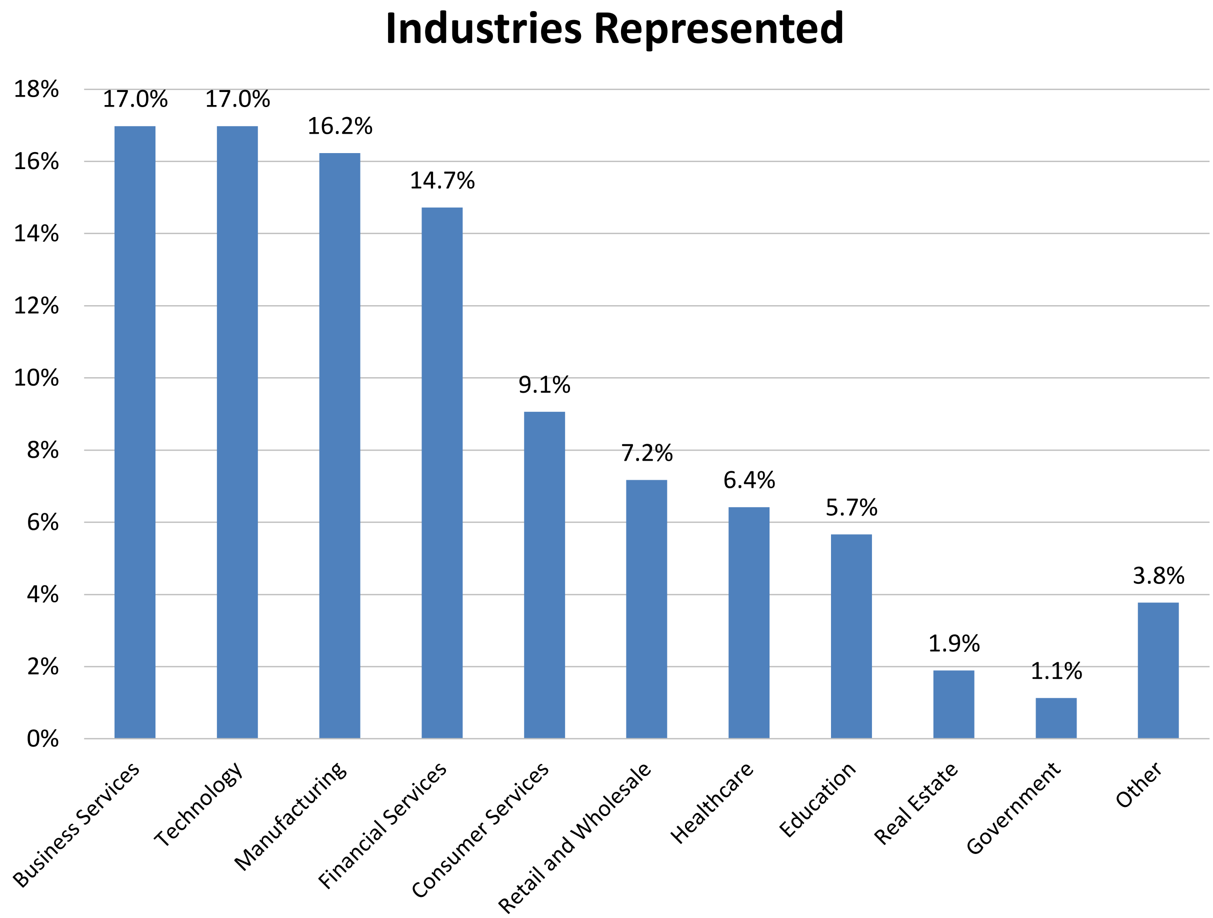


Figure 3 – Vertical industries represented

Organization Size

In 2023, our survey includes small organizations (1-100 employees), midsized organizations (101-1,000 employees), and large organizations (>1,000 employees) (fig. 4). This year, small organizations account for 27 percent of our sample, midsized organizations account for 27 percent, and 46 percent of respondents are from organizations larger than 1,000 employees.

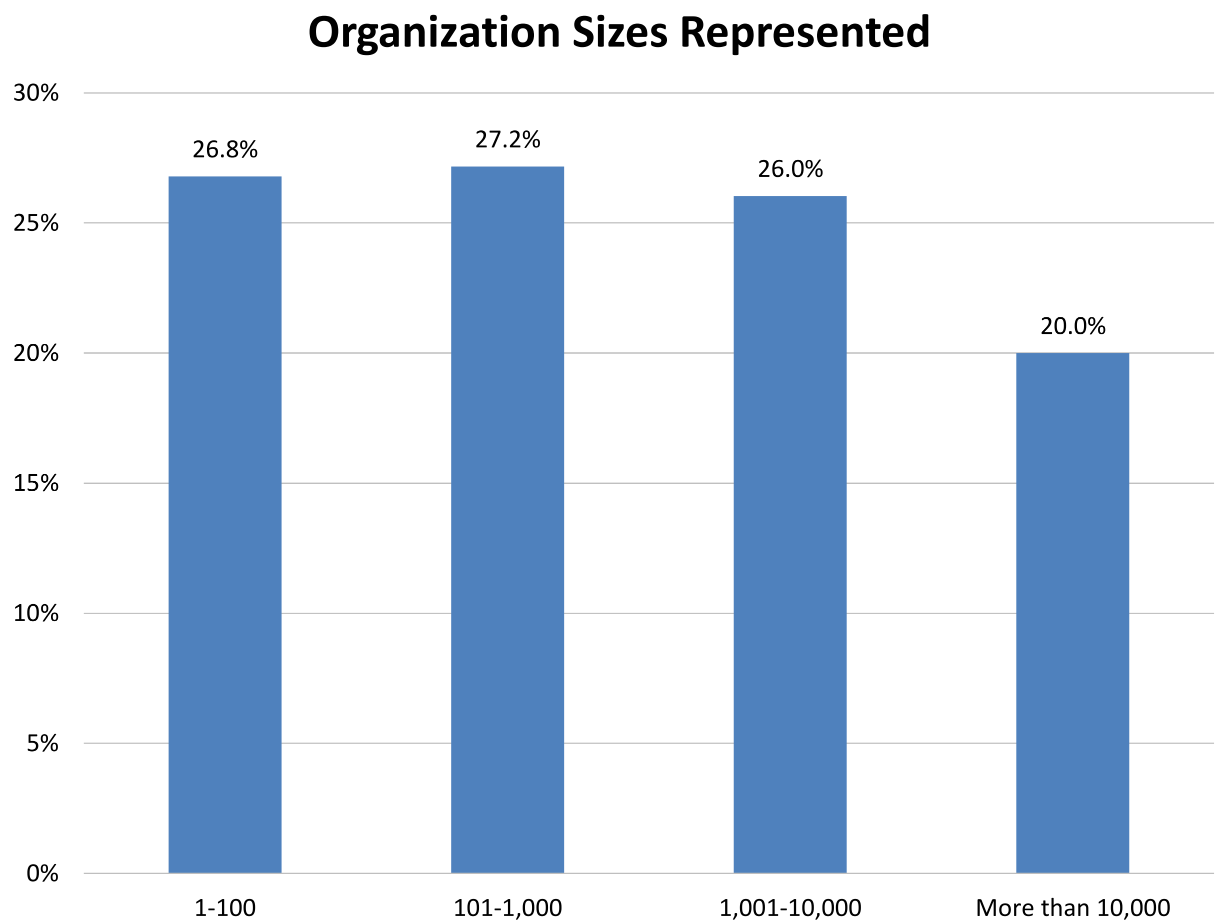


Figure 4 – Organization sizes represented

Analysis & Trends

Analysis and Trends

Importance of Data Engineering

We asked respondents about the importance of data engineering. In our 2023 survey, 39 percent of respondents indicate data engineering is of *critical importance* (fig. 5), up from 32 percent in 2022, while 38 percent indicate it is *very important*, up from 29 percent in 2022. Less than 2 percent indicate that data engineering is not important. Clearly, data engineering is a segment rapidly growing in importance.

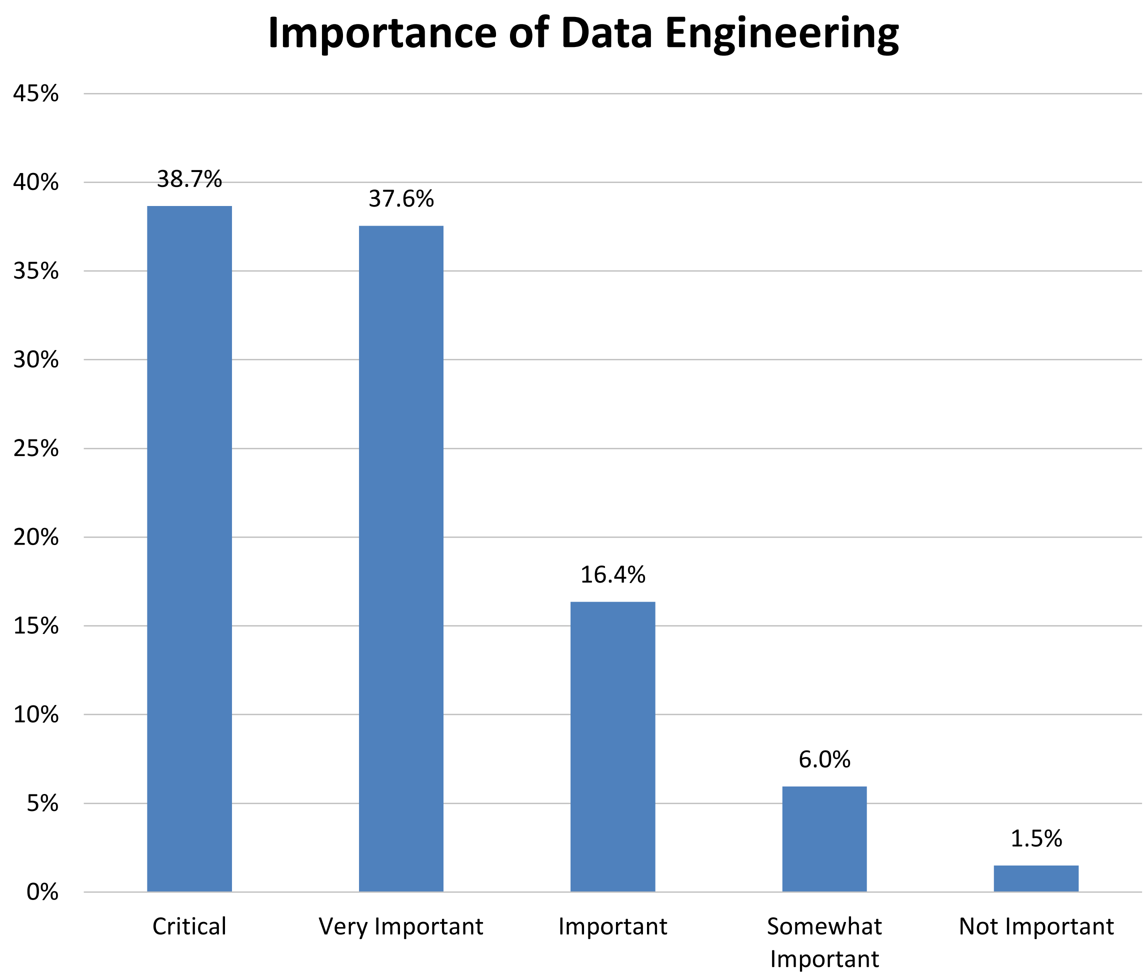


Figure 5 – Importance of data engineering

This year, respondents from *BICC* functions place the highest overall importance on data engineering. Over 45 percent of survey respondents from BICC functions indicate data engineering is of critical importance (fig. 6). Respondents from *operations* functions were more likely to place a slightly higher level of critical importance on data engineering capabilities than any other functions; but overall, they placed a lower level of importance on data engineering compared to other functions. Operations often span multiple business functions that each likely have their own data sets. Hence, the level of critical importance for operations respondents is not surprising, given the number of data and workflow integration points that operations respondents are often required to integrate and manage data across.

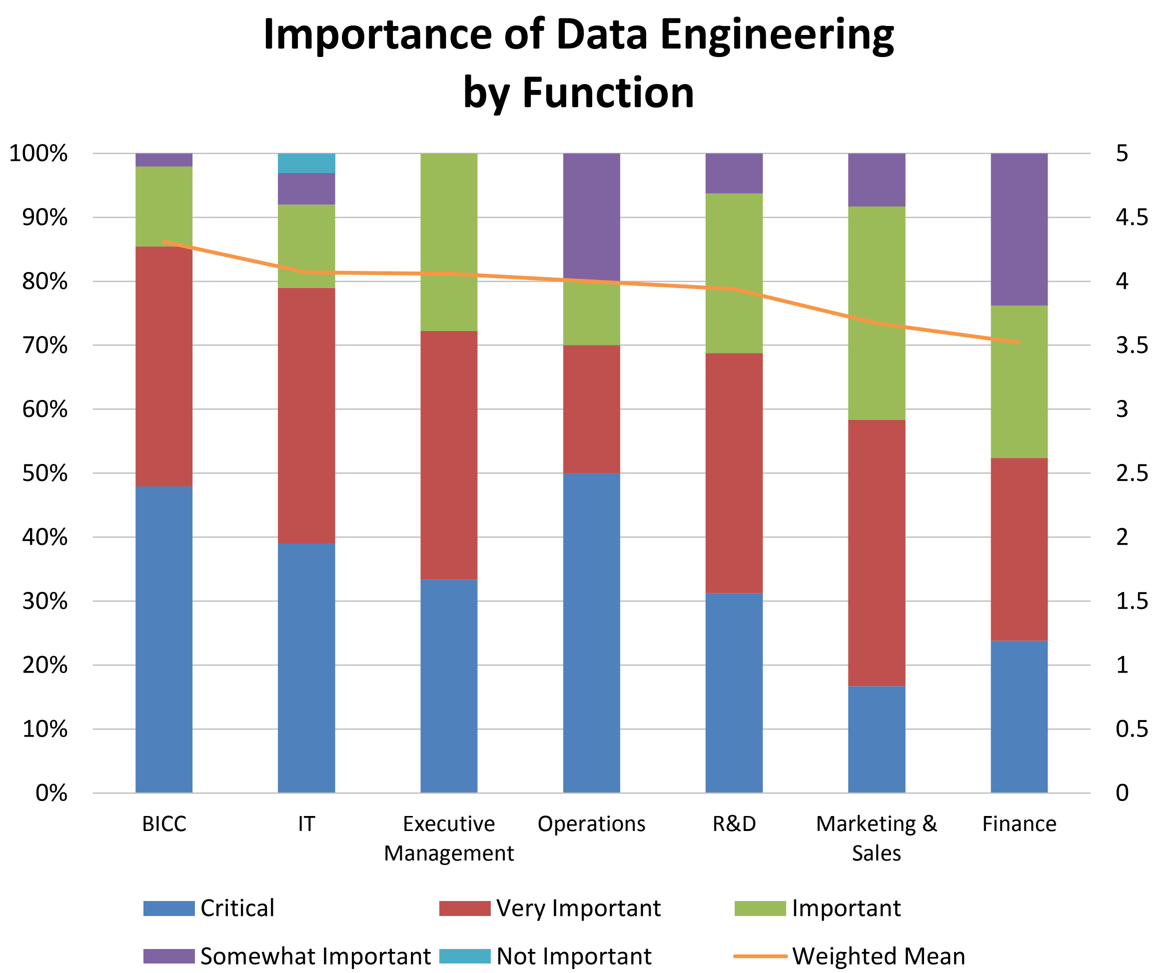


Figure 6 – Importance of data engineering by function

Data engineering is an increasing priority as organization size increases beyond 1,000 employees, likely due to increasing complexity of information infrastructure and scale of digital operations/analysis. *Smaller organizations* (1-100 employees) often use simple query interfaces (e.g., ODBC / JDBC) and spreadsheets to extract their data for analysis—perhaps into an analyst's personal data prep and integration tool. Even so, smaller organizations can have very complex data-integration requirements, as they can have as many data sources for analysis as larger organizations. *Larger organizations* often already made investments into data-engineering platforms and associated best practiecs and approaches.

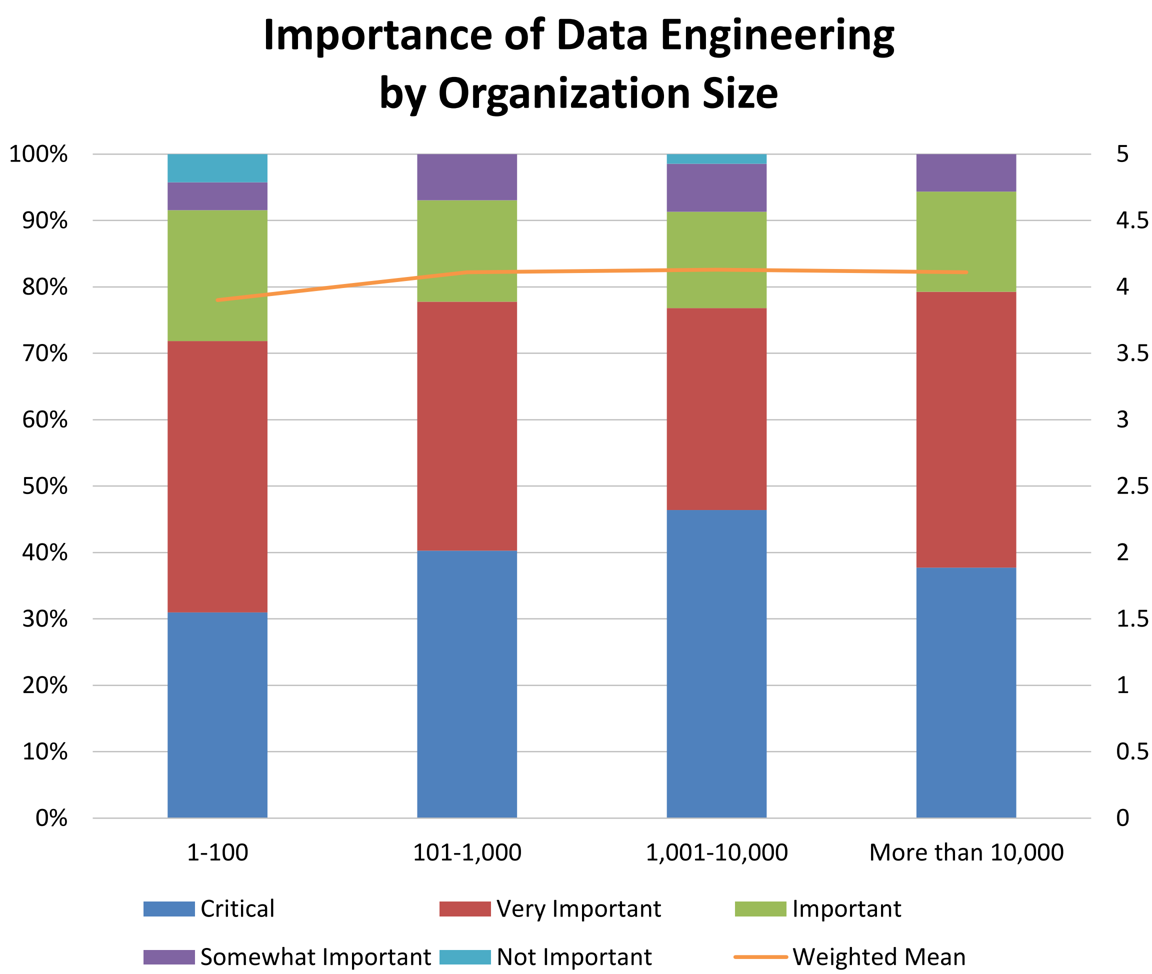


Figure 7 – Importance of data engineering by organization size

Although all industries place a high priority on data engineering, respondents from *technology* and *financial services* industries give it the highest relative priority, followed by *retail and wholesale* and *business services* (fig. 8). *Consumer services* and *education* respondents give data engineering lower importance scores. That said, all industries gave it a mean importance score of over 3.8/5.

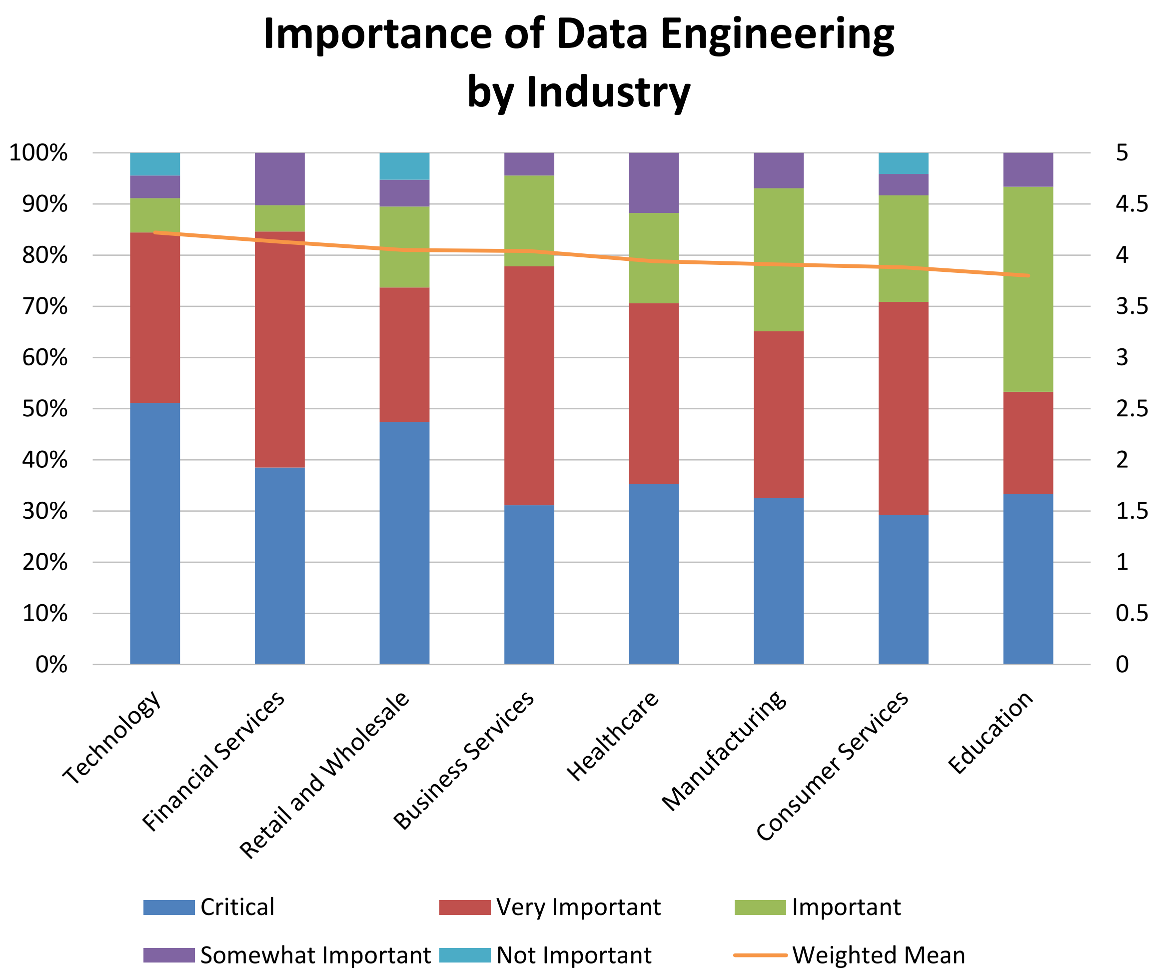


Figure 8 – Importance of data engineering by industry

Taking a data engineering approach to accessing, combining, and preparing data speaks to a degree of maturity and sophistication in leveraging data as a strategic asset (fig. 9). It is not surprising, then, that respondents that rate their BI initiatives as a success place a much higher relative priority on data engineering compared to those organizations that are less successful.

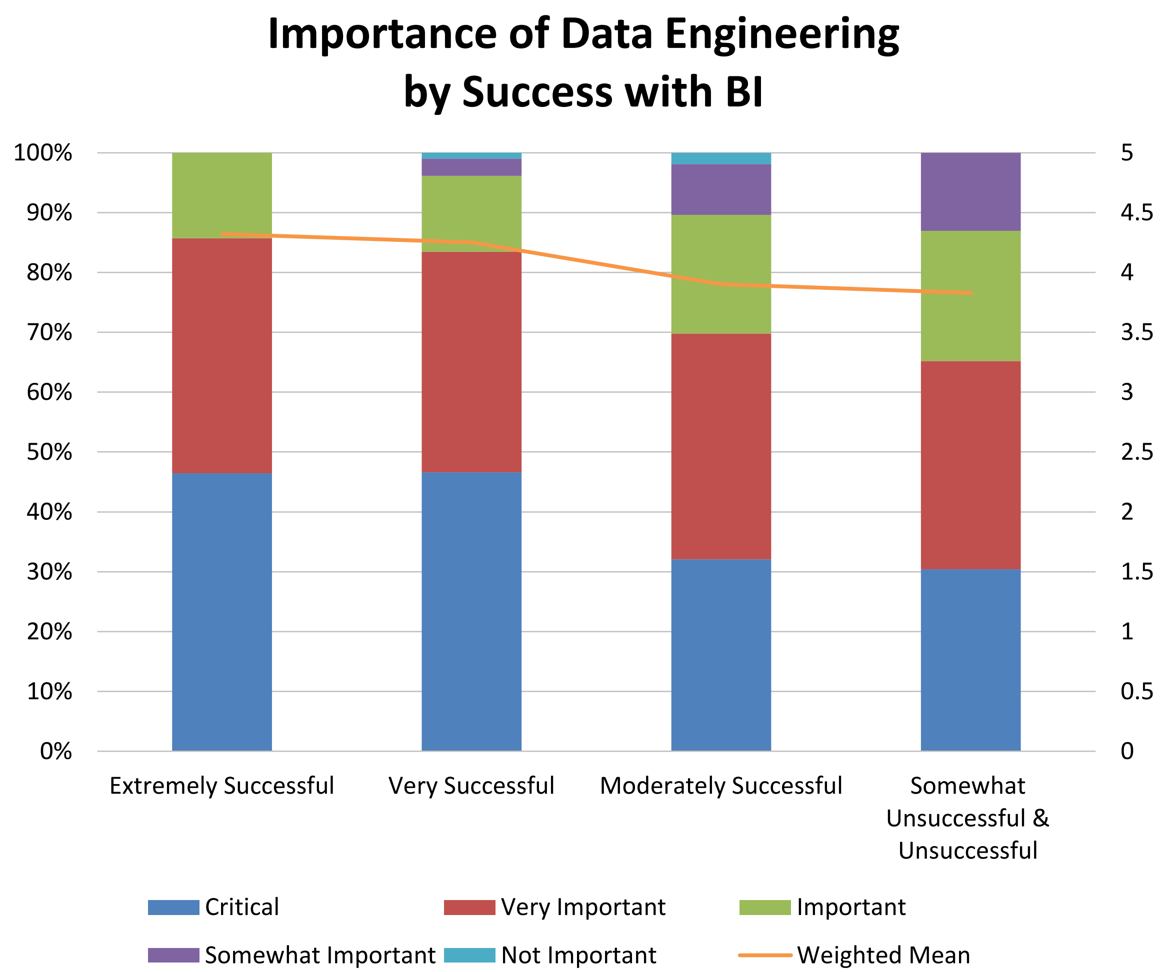


Figure 9 – Importance of data engineering by success with BI

Use and Plans for Data Engineering

We asked respondents about their current use and plans for data engineering. While 65 percent of respondents say their organizations use data engineering capabilities *today*, 20 percent of respondents indicate plans to use data engineering tools within the next *12 months*, indicating a growing demand for data engineering in 2023 (fig. 10). Only 8 percent have no plans to use data engineering capabilities.

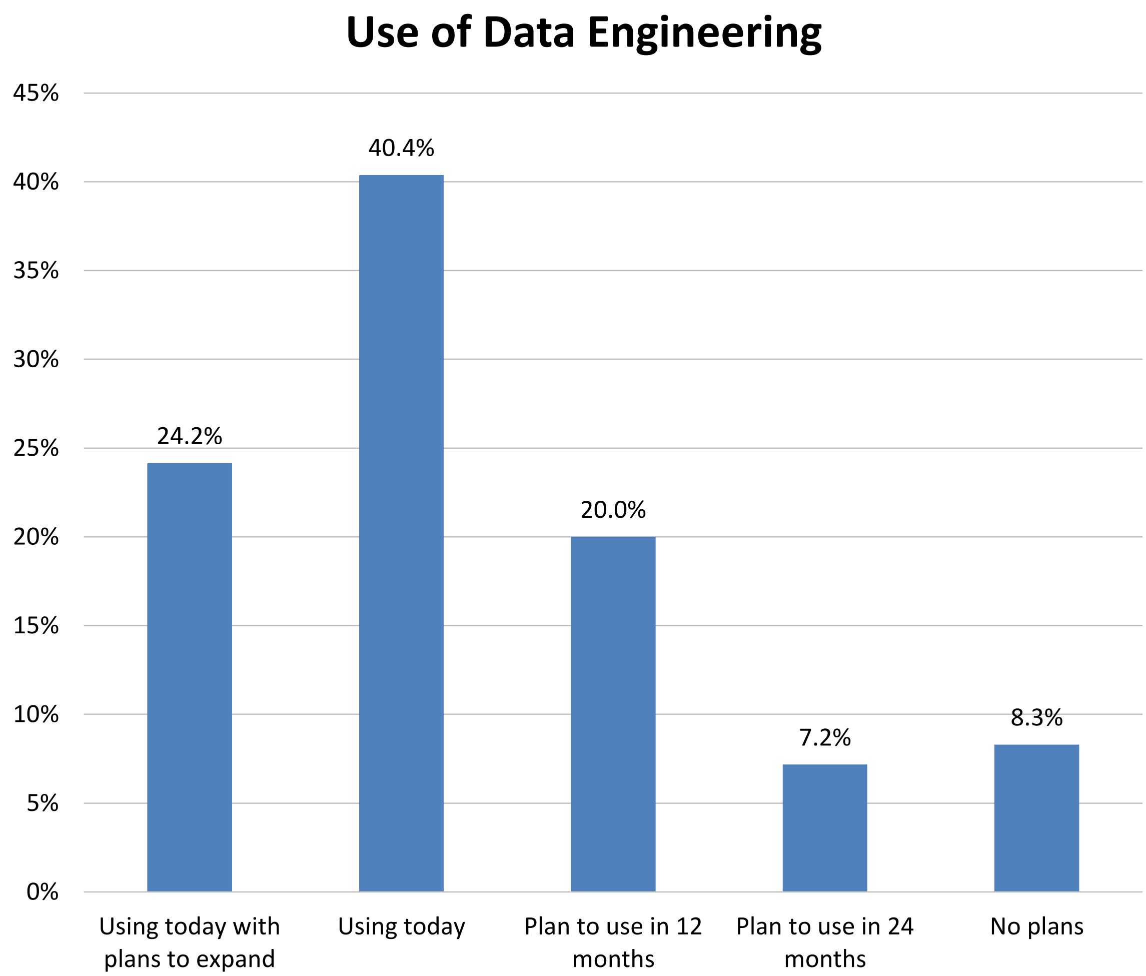


Figure 10 – Use of data engineering

Across geographies, *North American* respondents indicate the highest use and plans to expand their data engineering capabilities (fig. 11). *Latin American* respondents have the lowest percent of use of data engineering but the highest level of plans to use in 12 months. *Asia Pacific* respondents indicate the highest percent (16 percent) of respondents that have plans to use data engineering within 24 months.

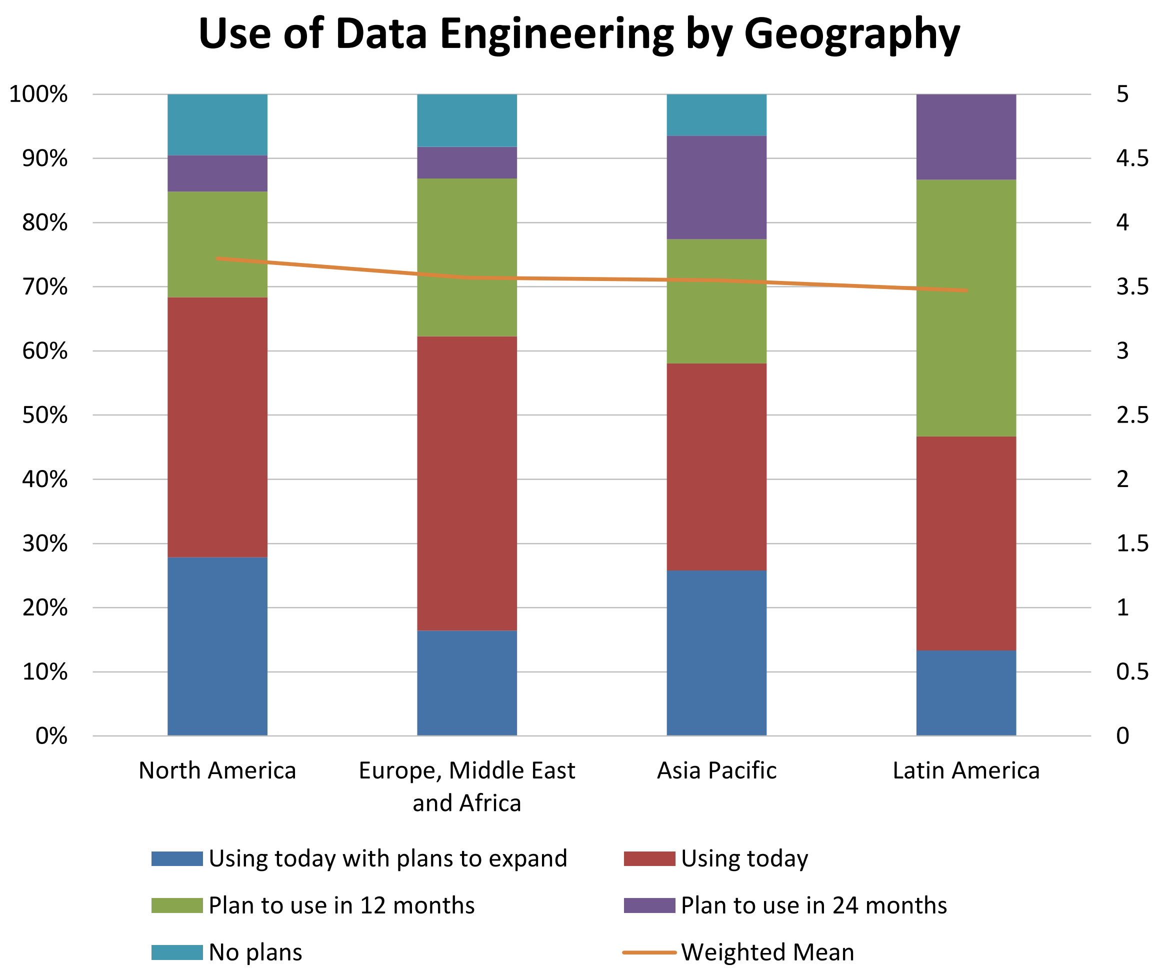


Figure 11 – Use of data engineering by geography

Survey respondents from the *technolgy* and *healthcare* industry segments are among the highest users of data engineering today (fig. 12). *Consumer services* and *financial services* industry respondents are among the highest data enginering users and also the highest percent indicating plans to expand their use. Respondents from the *education* segment indicate the lowest overall use and plans for data engineering.

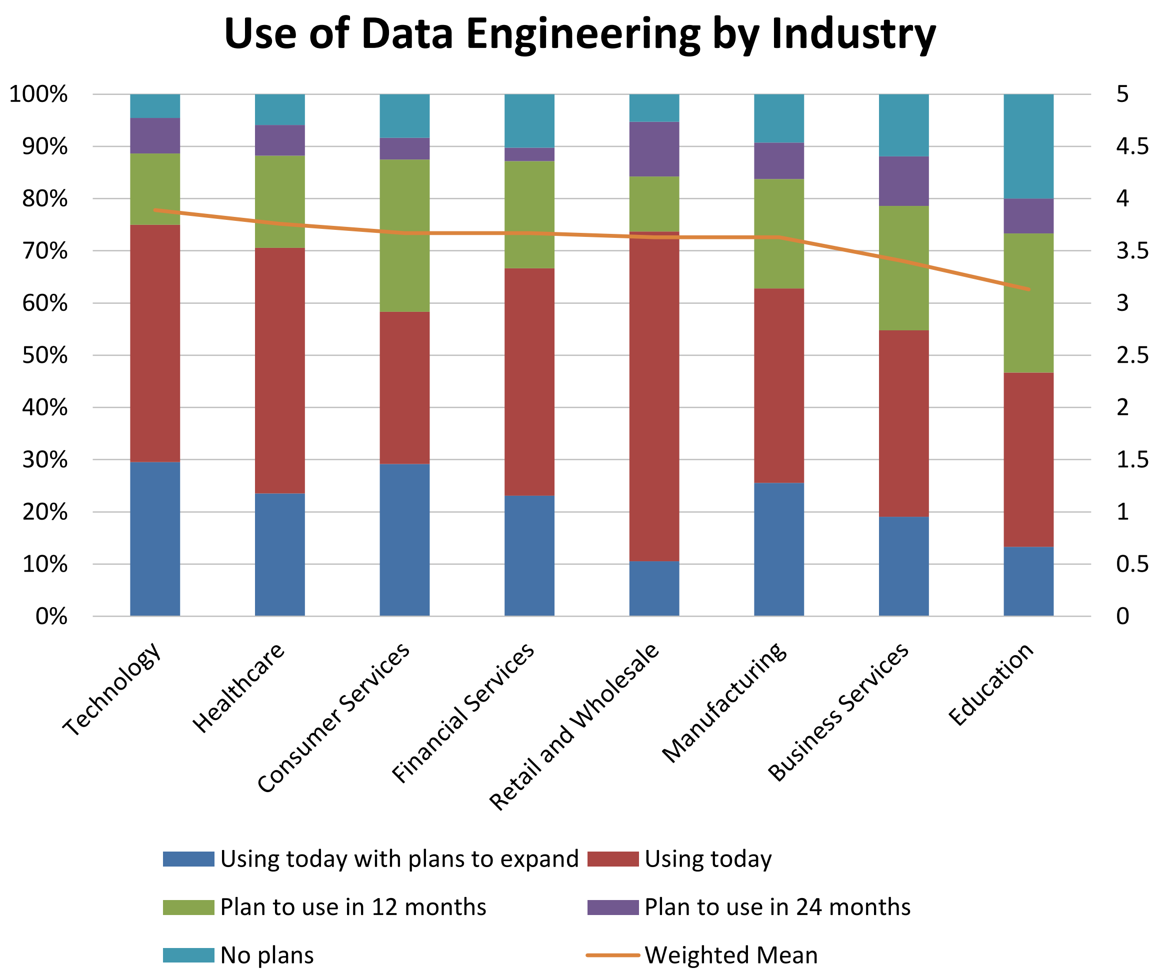


Figure 12 – Use of data engineering by industry

Respondents from the *BICC* functions indicate the highest use of data integration today (fig. 13). According to our annual BICC Market Study report, supporting business units with *analytical model development* and *database design and management* are key competencies for the BICC. As such, it makes sense that data engineering is a critical capability for BICC respondents. As an emerging discipline, *R&D* function respondents report a high level of use today but indicate they have no plans to expand their use.

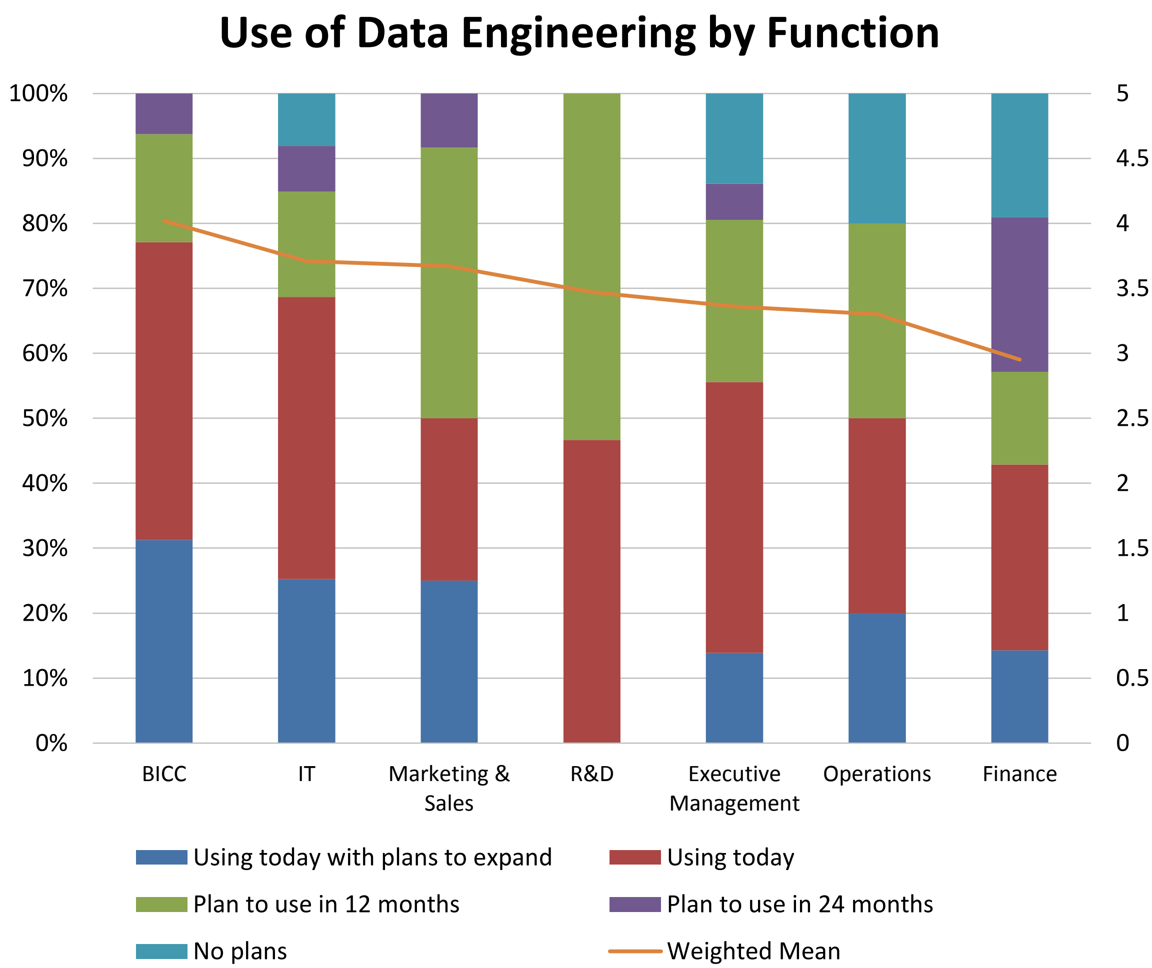


Figure 13 – Use of data engineering by function

The use and plans for data engineering vary by organization size (fig. 14). *Smaller organizations* have the lowest use and plans for data engineering capabilities today, whereas the use and plans go up with the size of the organization. *Larger organizations* often have a degree of complexity with numerous data sources that require investments in data engineering. In contrast, *smaller organizations*, with less complex businesses may use spreadsheets to manage data and may not yet have a need for data engineering capabilities.

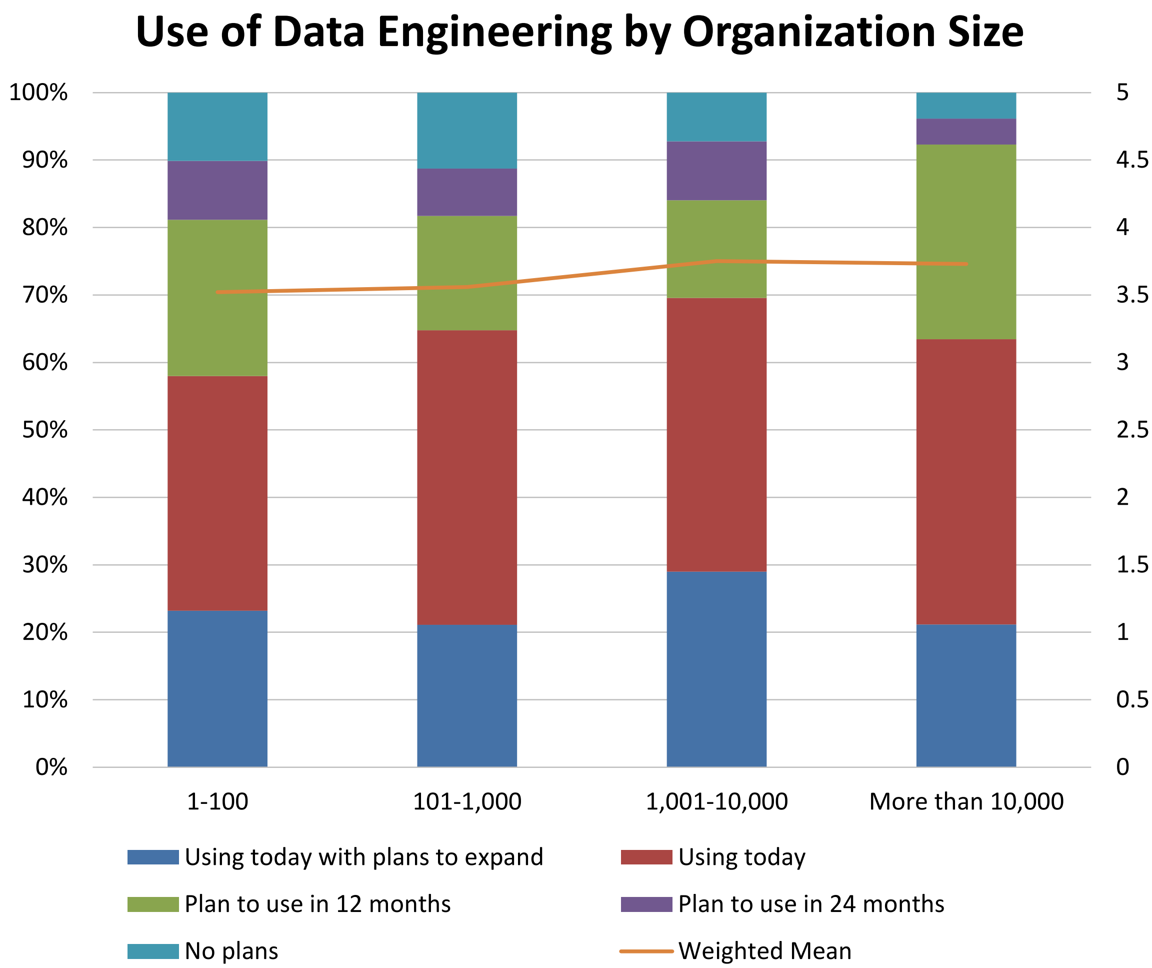


Figure 14 – Use of data engineering by organization size

Use Cases for Data Engineering

Organizations often purchase and use data engineering products for multiple BI use cases (fig 15). We asked our survey respondents to indicate the estimated percent of their BI use cases for which their data engineering products are purchased and used. Twenty-seven percent say they employ 60 percent of their data engineering capabilities for *data integration, cleansing, and transformation workflows for a data warehouse supporting dashboards and reporting*. *Re-platforming or replication of an existing data warehouse* is the lowest use case for data engineering investments. These percentages correspond to the prevalence of the various forms of business intelligence, analytics and data science activities in most organizations.

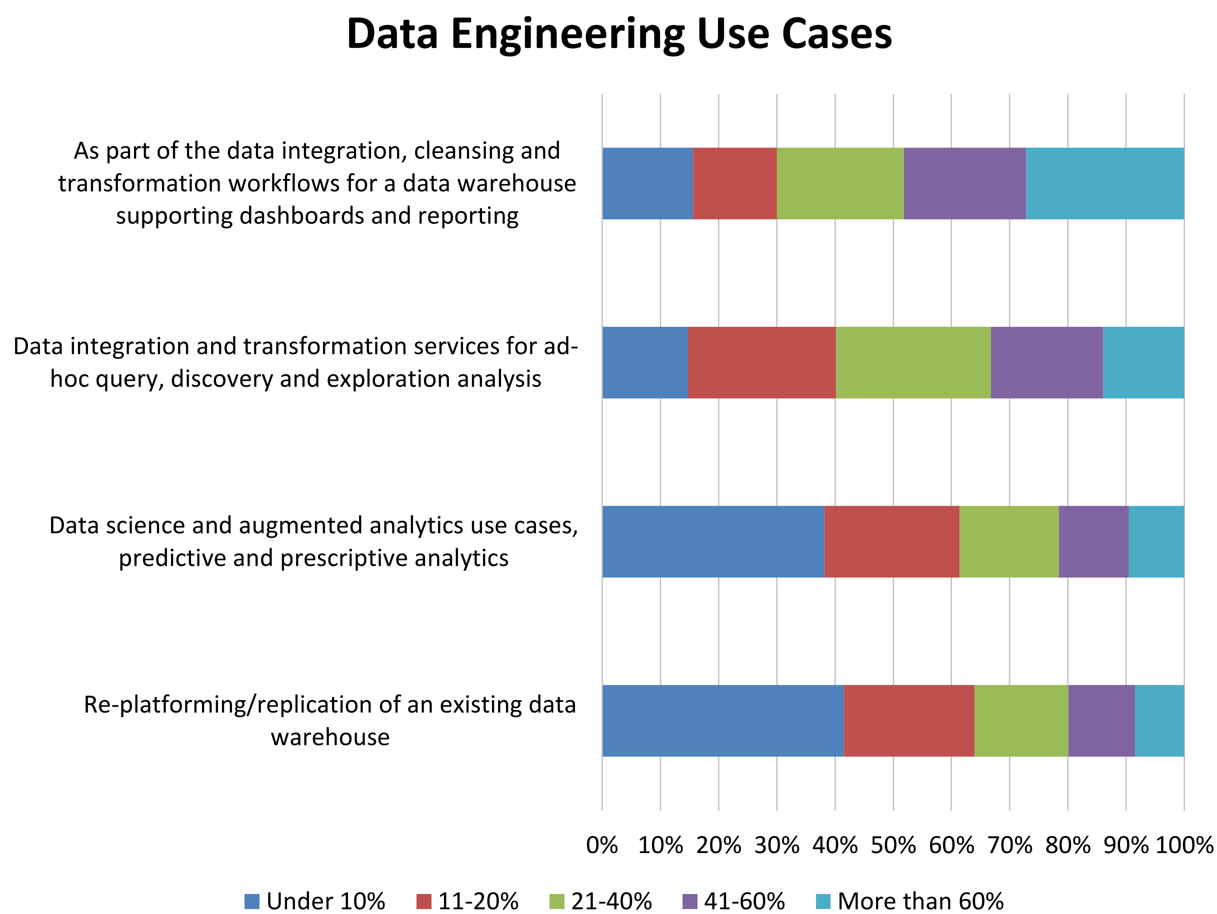


Figure 15 – Data engineering use cases

We filtered the data engineering usage by the highest percent use case responses and geographic regions (fig.16). There are variations of the percentage of use cases by geographies. *North America* and *EMEA* regions have similar percentages for their data engineering investments for supporting dashboards and reporting use cases. *Latin American* respondents indicate the lowest use of data engineering capabilities employed for the data science and augmented analytics use cases, predictive and prescriptive analytics.

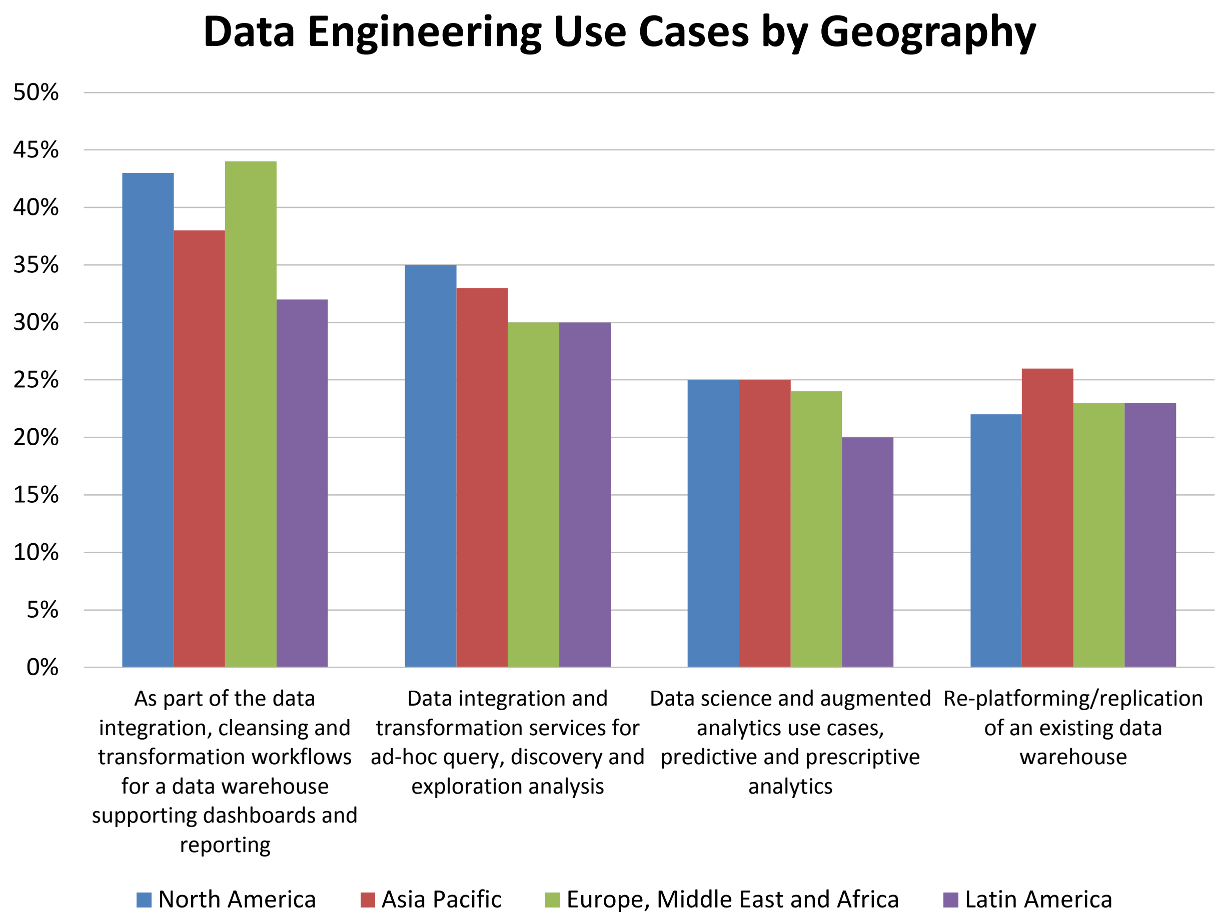


Figure 16 – Data engineering use cases by geography

Data engineering use cases vary by industry (fig 17). The highest use case across all industries is using *data engineering workflows for a data warehouse supporting dashboards and reporting* especially in the *retail and wholesale* and *healthcare* industries. The second-highest use case is for *data integration and transformation services for ad-hoc query, discovery, and exploration analysis*, with respondents again from the *retail and wholesale* and *healthcare* industries indicating this is a very common use case as well. *Data science and augmented analytics use cases* and *replacement/replication of an existing data warehouse* are currently the lowest use cases driving data engineering investments.

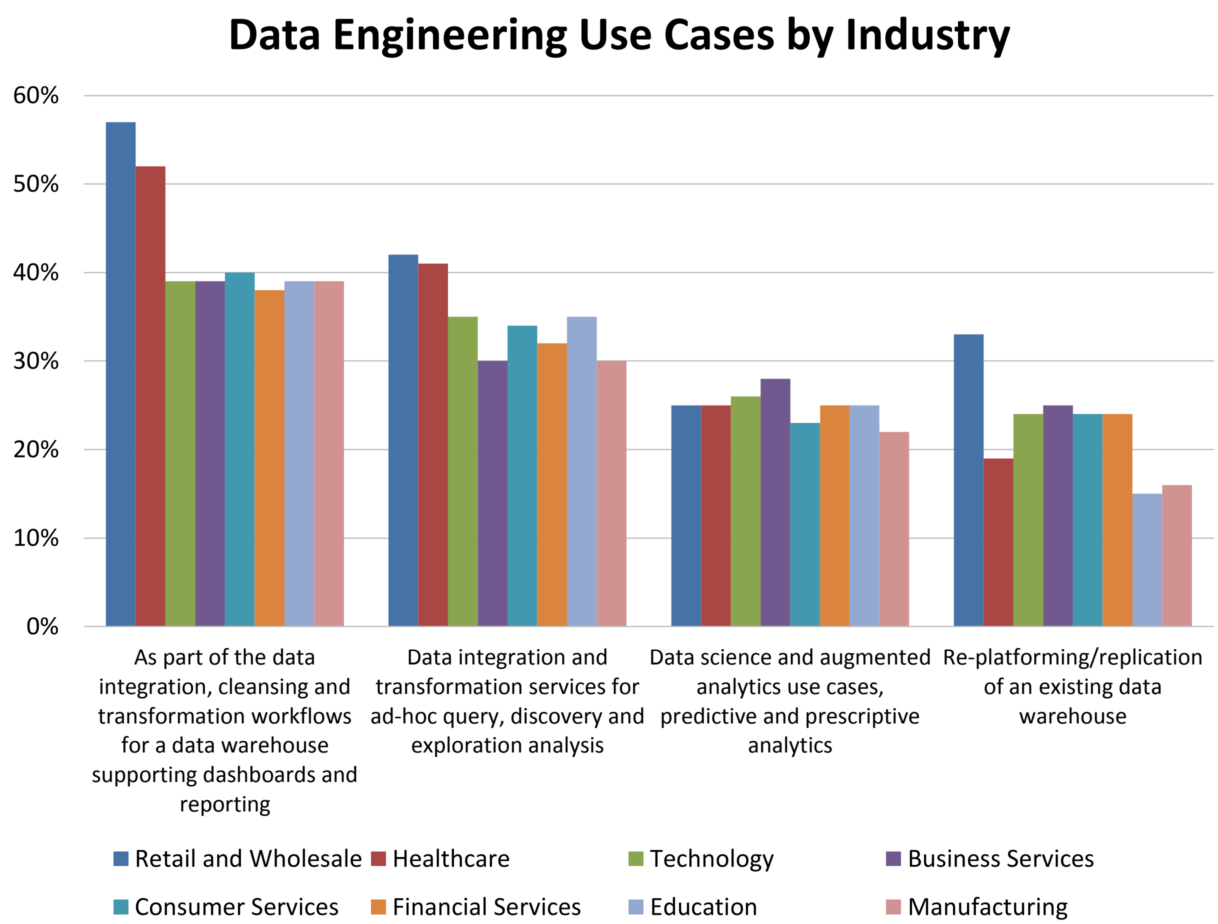


Figure 17 – Data engineering use cases by industry

Data engineering used for data integration, cleansing, and transformation workflows for a datawarehouse for supporting dashboards and reporting is a high use case for respondents from the all functions except *operations*. The *BICC* function rates this use case the highest overall. *R&D* function respondents rate the use of data engineering for *data science and augmented analytics* higher than other functions (fig. 18).

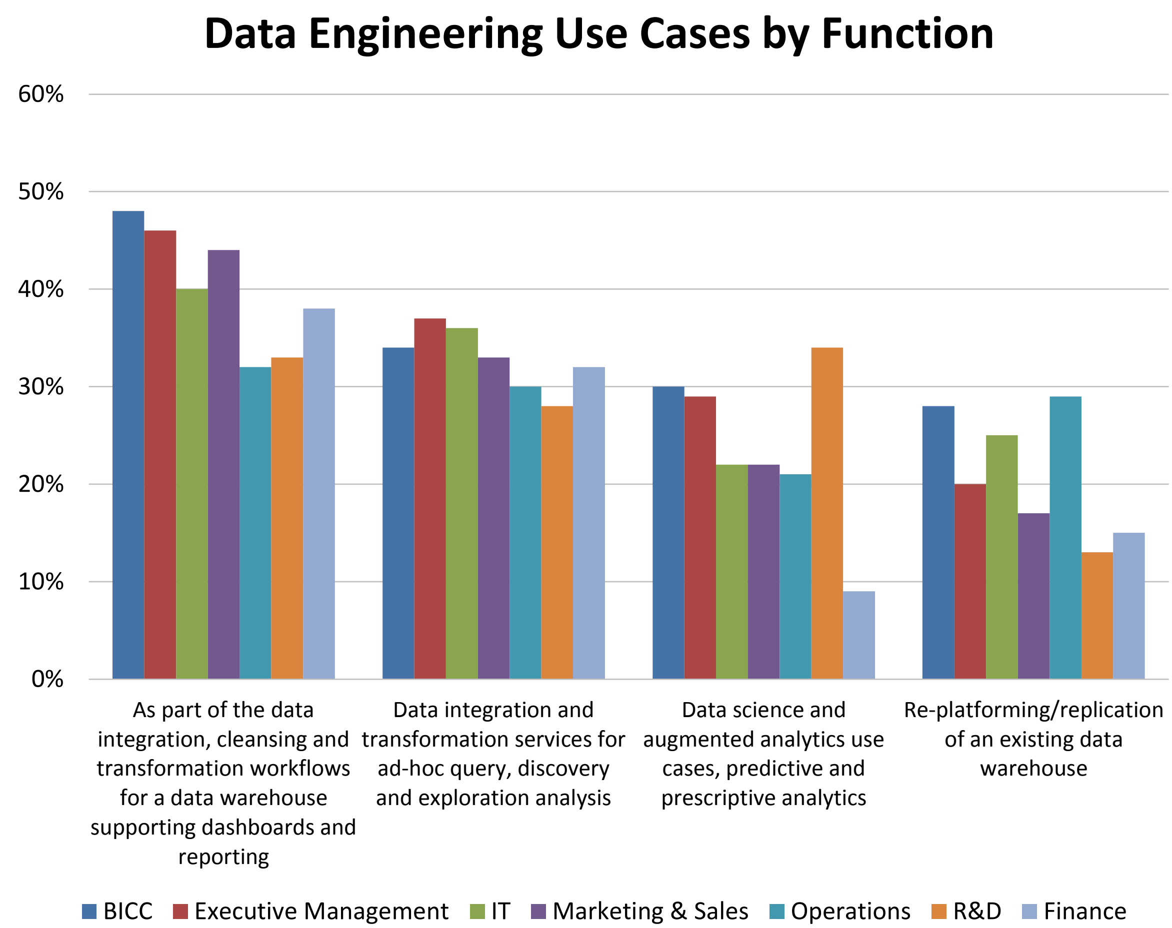


Figure 18 – Data engineering use cases by function

Data Engineering Features

The importance of data engineering features ranked by our survey respondents is presented in fig. 19. The 30 top results are sorted by weighted mean of the importance of the features (critical, very important, etc.). The top features for data engineering workflows include the *ability to aggregate and group data and ETL and ELT workflows* as well as the management of engineering workflow such as *alerting/Job monitoring* and *execution plan, time, and event-based schedulers for jobs*. Least important is support for Kafka and Apache big data services.

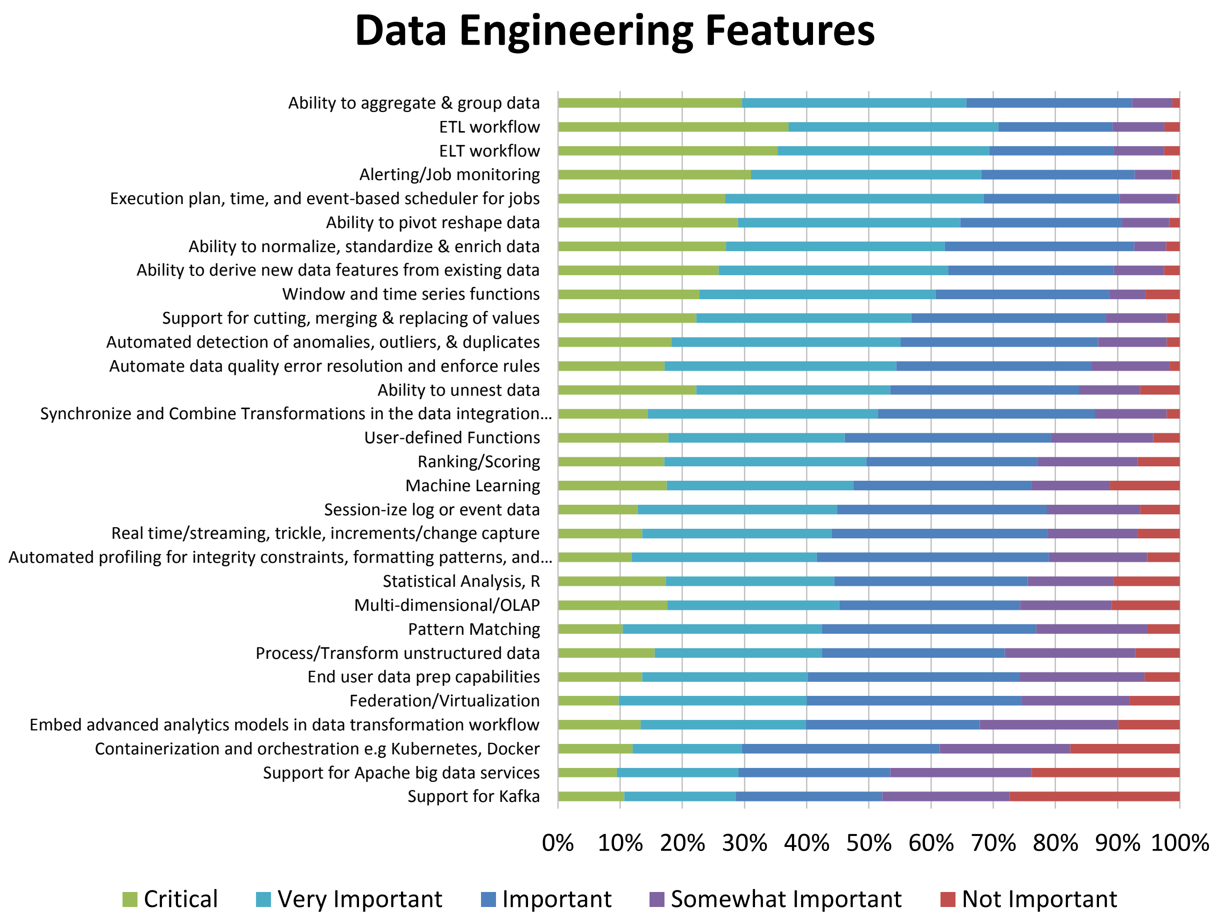


Figure 19 – Data engineering features

Data engineering features vary in importance in minor ways across geographies. We present here the top 15 data engineering features by geographic location of the survey respondents. *North American* and *Asia Pacific* respondents indicate a higher importance on *ELT workflows* compared to ETL workflows (fig. 20). *Asia Pacific* respondents indicate *synchronize and combine transformations* as a higher importance compared to other geographies.

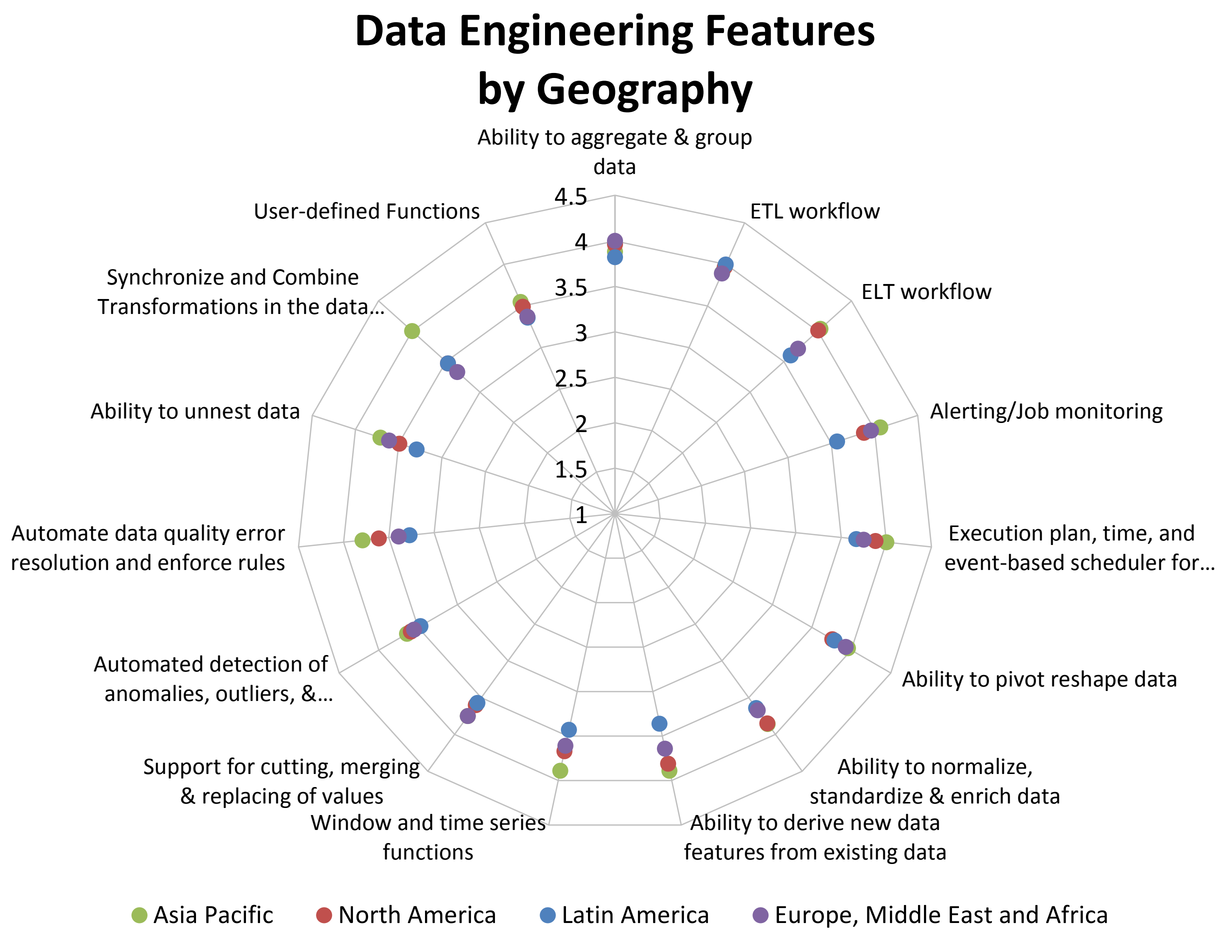


Figure 20 – Data engineering features by geography

The top 15 data engineering features are ranked in importance by industry of the respondents (fig. 21). There are minor variations when we rank the features by industry with *ability to aggregate and group data* at the top and *user-defined functions* of lower relative importance.

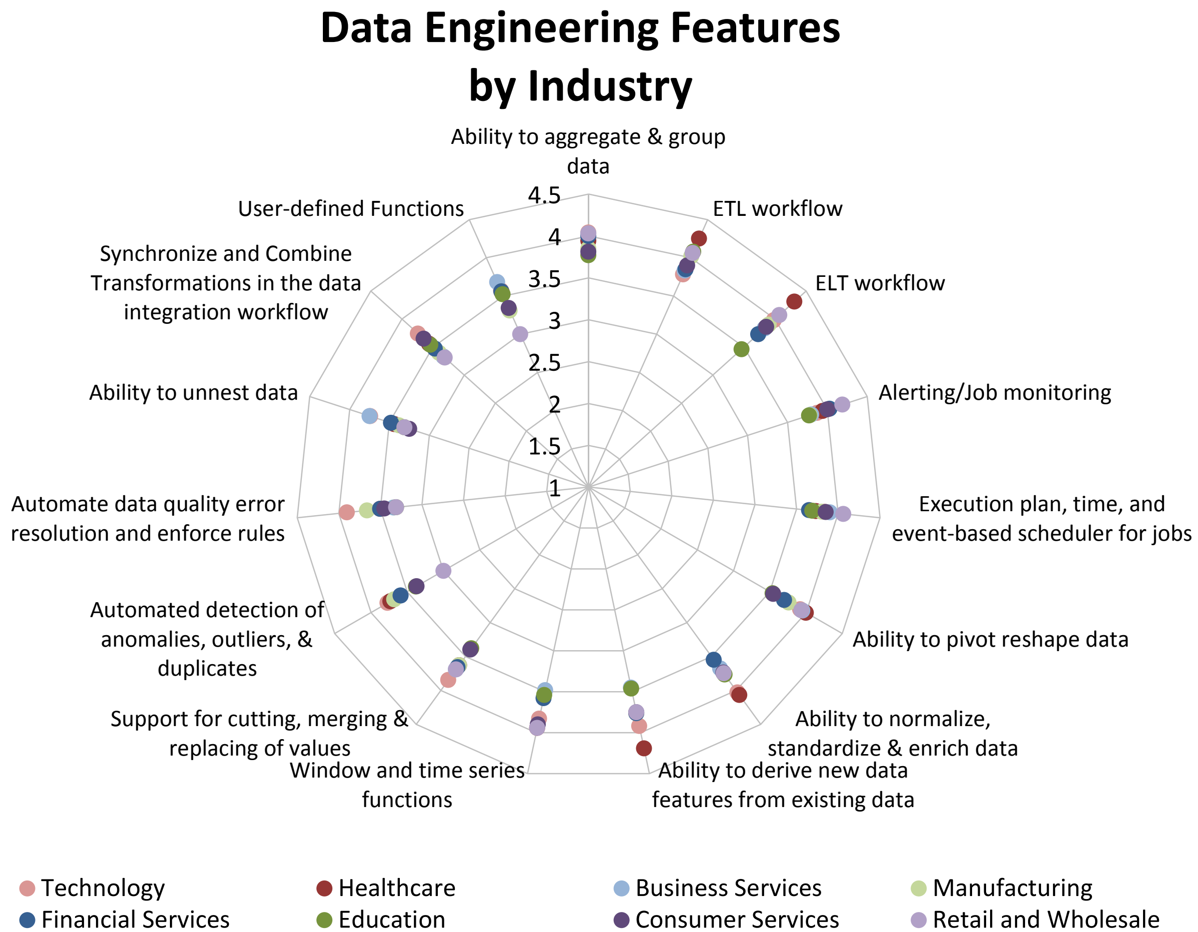


Figure 21 – Data engineering features by industry

The top importance of data engineering features varies slightly across functions. Respondents from *operations* have the highest number of outliers, both high and lower in importance than other functions. *User-defined functions* are the lowest relative importance across all functions surveyed (fig. 22).

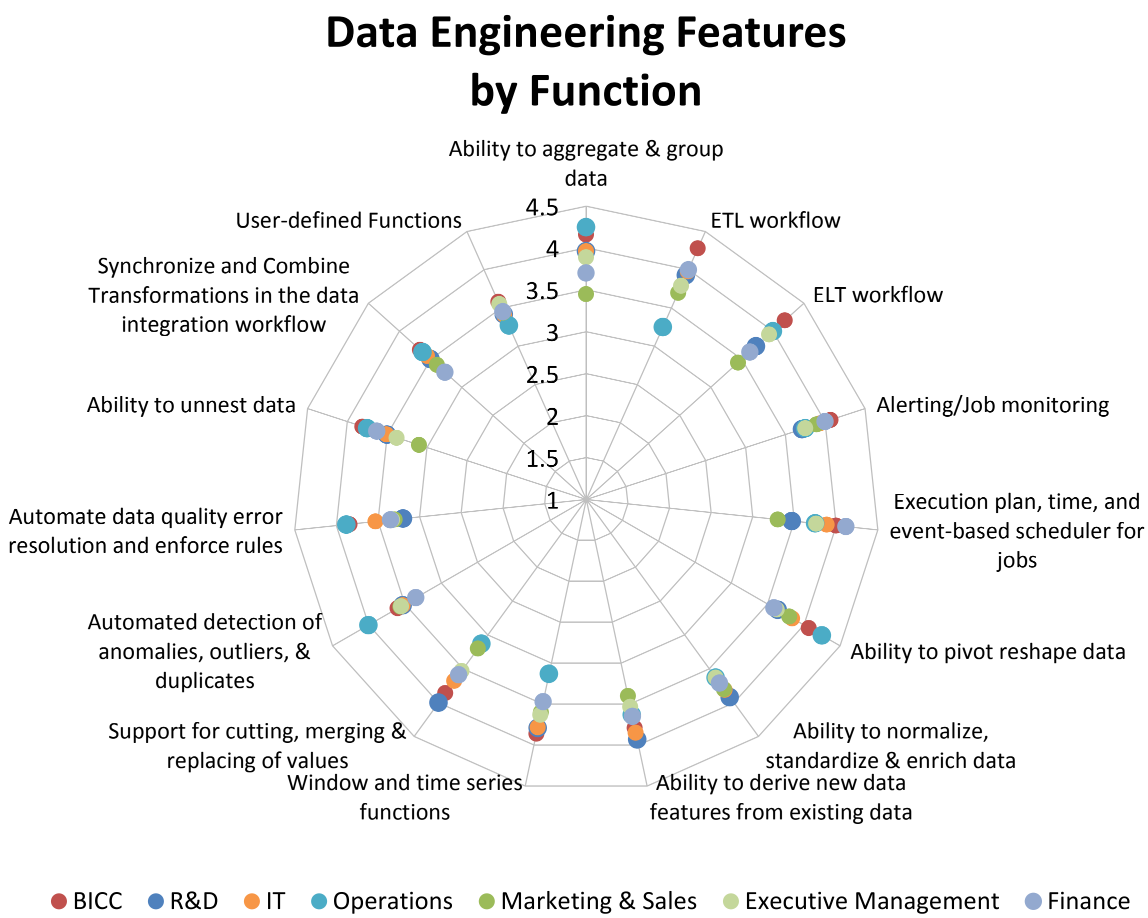


Figure 22 – Data engineering features by function

There are minor variations of importance of data engineering features when we slice the data by organization size. Organizations of *all sizes* indicate *ability to aggregate and group data* as their most important data engineering feature (fig. 23). *Larger organizations* indicate a higher importance for *ETL workflows* versus ELT workflows. *Midsized organizations* (100 – 10,000 employees) place a slightly higher importance on *ELT workflows* versus ETL workflows.

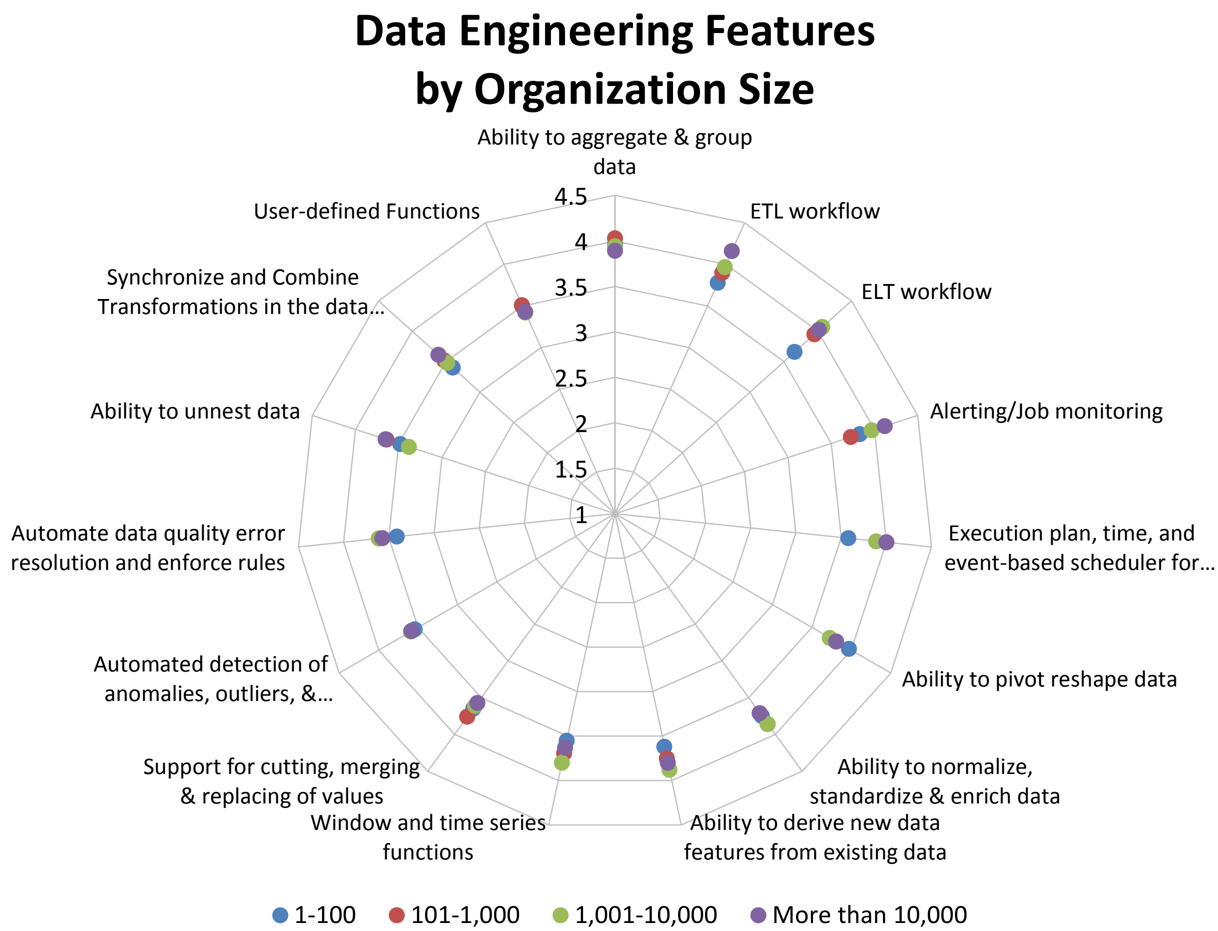


Figure 23 – Data engineering features by organization size

Frequency of Data Engineering

We asked respondents, “How frequently do people in your organization engage in data engineering prior to analysis?” Fifty seven percent of respondents engage in data engineering *constantly or frequently* (fig. 24). Twenty seven percent use data engineering *occasionally*, and less than 17 percent use data engineering *rarely or never*.

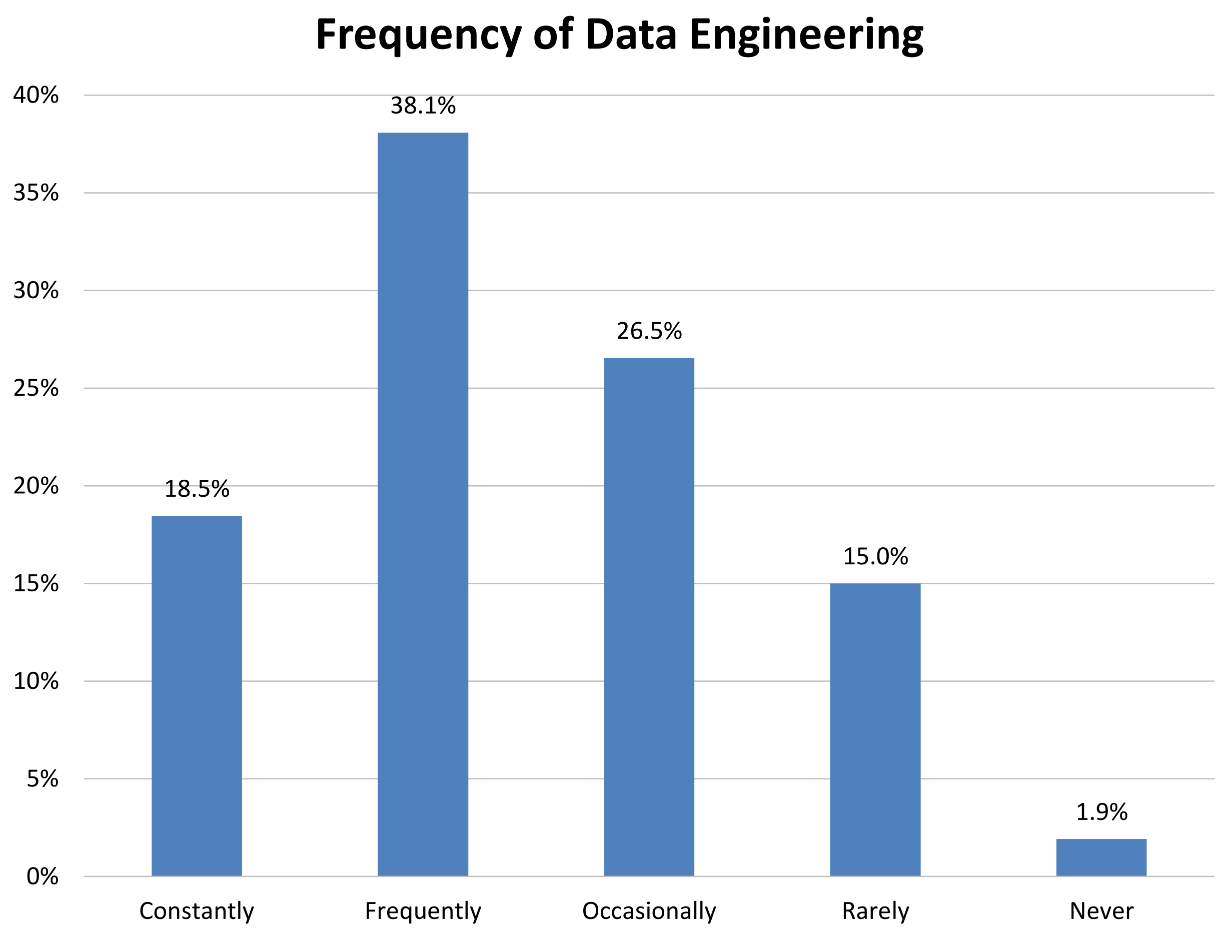


Figure 24 – Frequency of data engineering

All industry respondents tend to use data engineering frequently. *Technology* and *financial services* organizations are more likely to use data engineering *constantly or frequently* compared to other organizations (fig. 25). *Consumer services*, *retail and wholesale*, and *education* respondents are the least likely to use data engineering constantly or frequently.

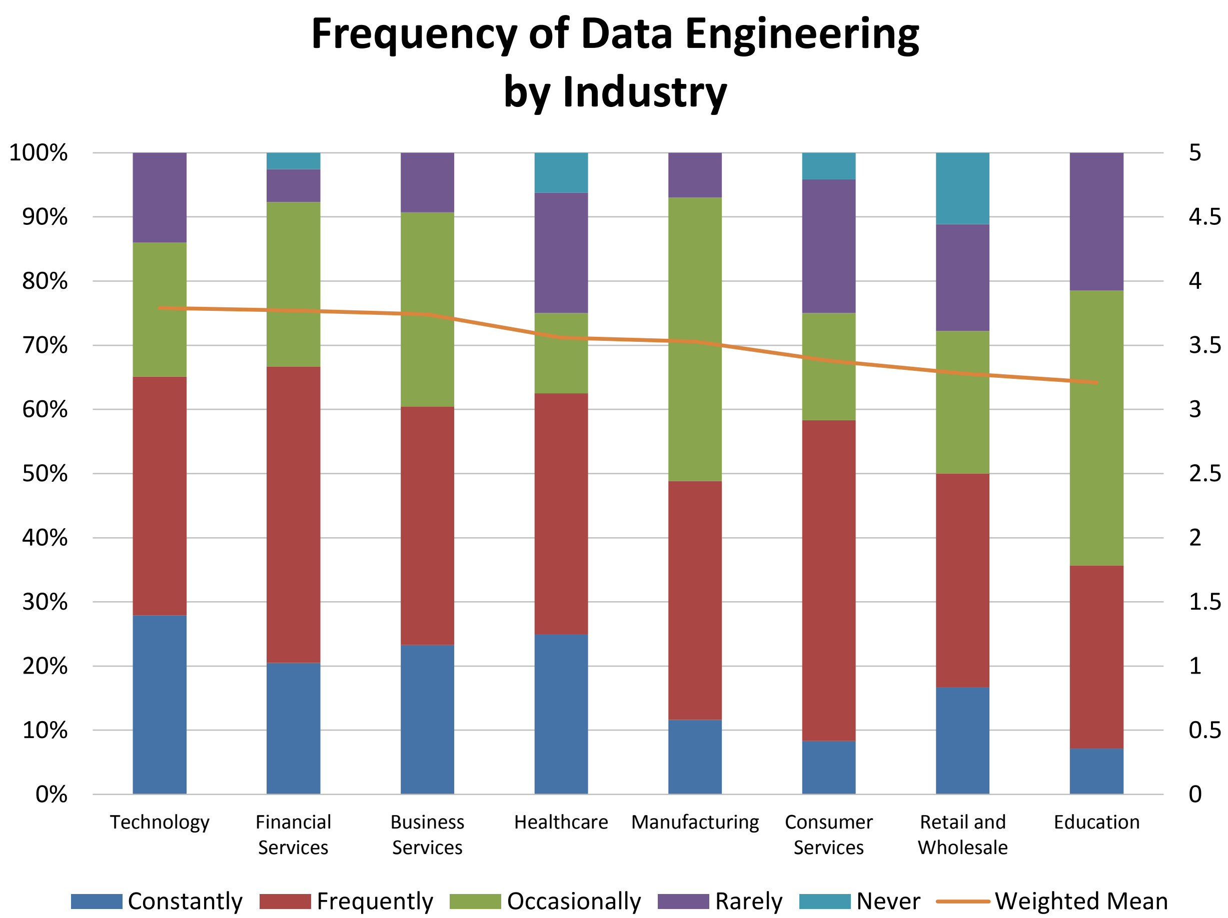


Figure 25 – Frequency of data engineering by industry

When we slice the responses to frequency of data engineering by function, we find some interesting responses. *Operations* and *R&D* functions use data engineering *constantly* (fig. 26). The large response of *rarely* for frequency of data engineering by *finance* is likely associated with the very common use and collaboration based on departmental spreadsheets for which a data engineering workflow may not be required.

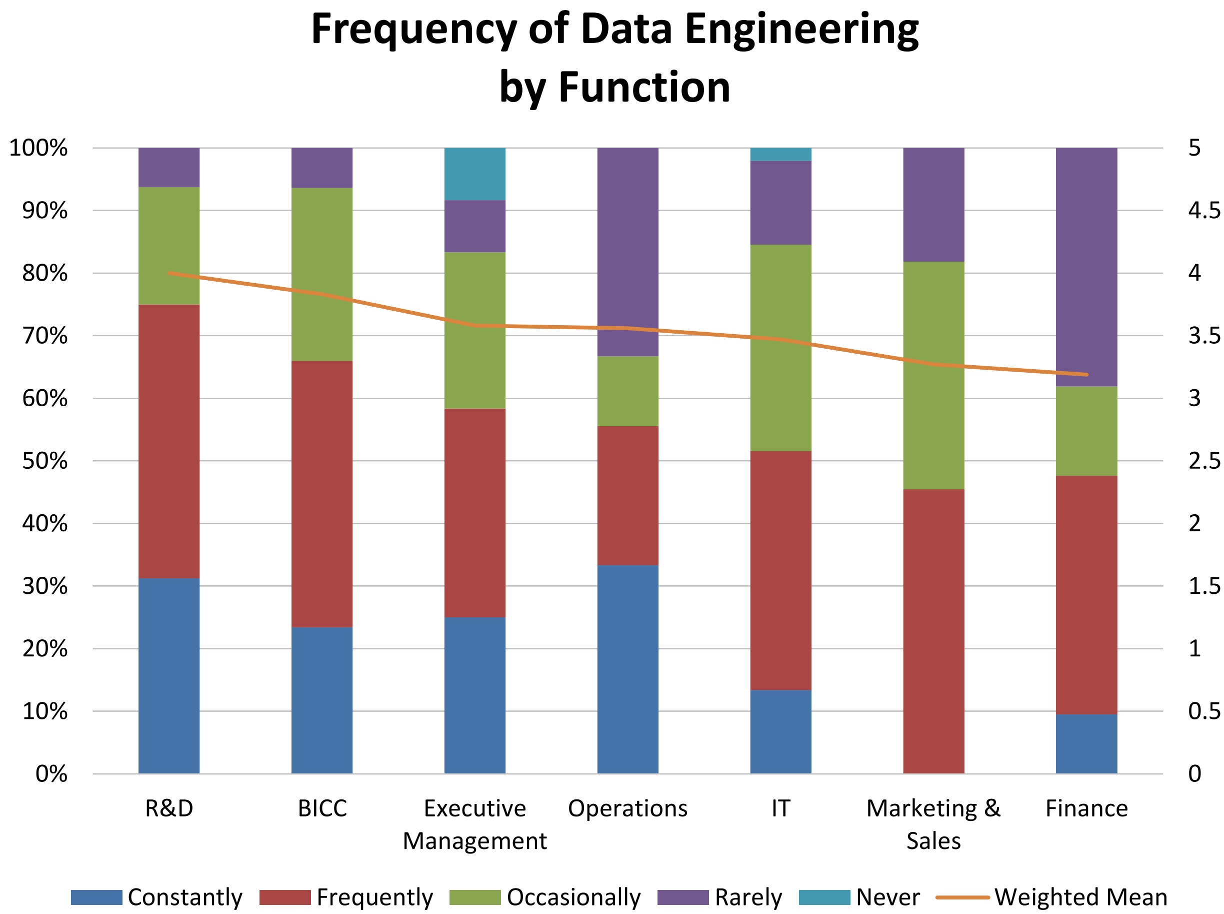


Figure 26 – Frequency of data engineering by function

Larger organizations, with their more complex operational and analytical data infrastructures, employ their data engineering more frequently than smaller organizations (fig. 27). Of interest is that approximately 5 percent of the respondents from organizations with 1*,000-10,000 employees* report that they never use data engineering, which may be due to the extensive use of spreadsheets to run and analyze the business.

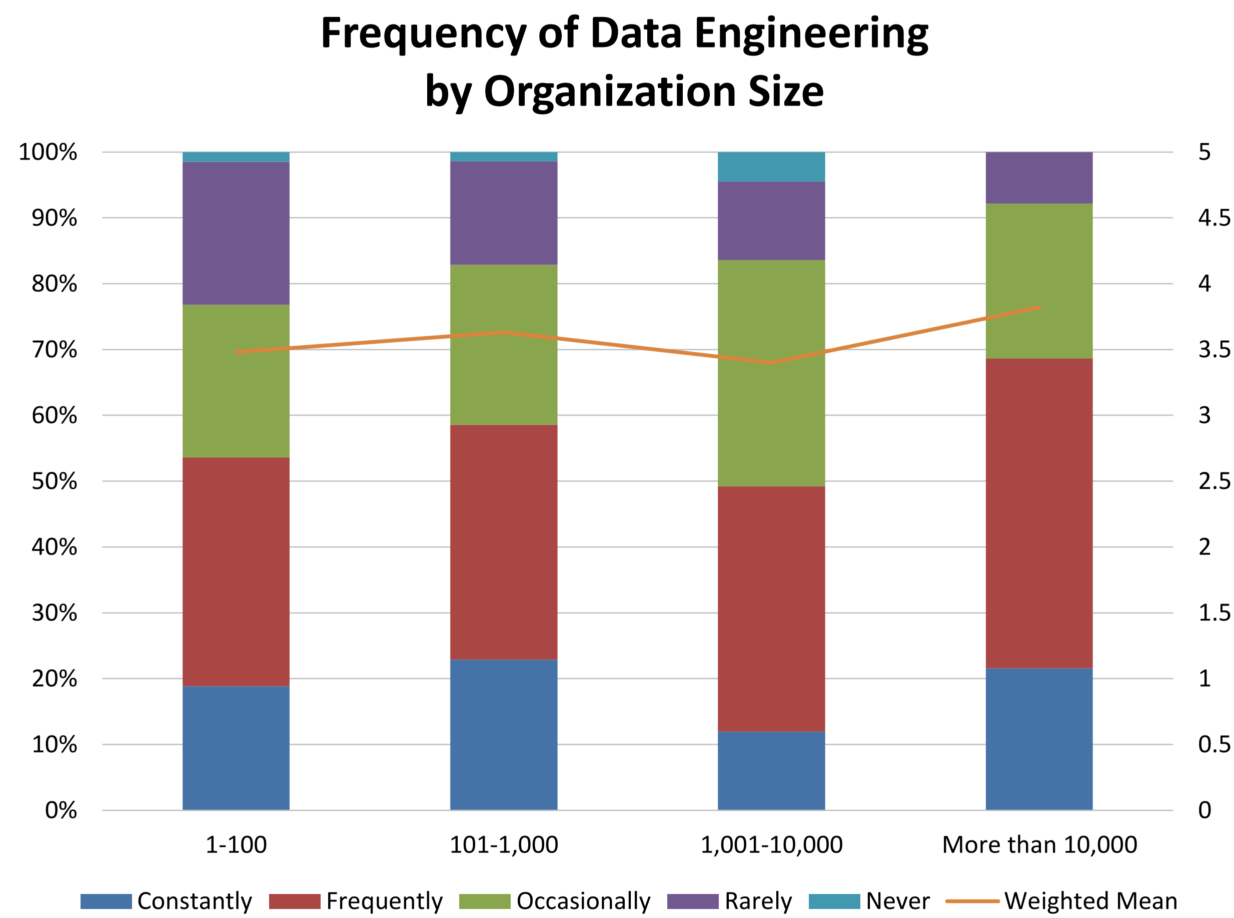


Figure 27 – Frequency of data engineering by organization size

It appears clear from the data (fig. 28) that the more frequent an organization employs data engineering in analyzing data, the more likely that organization will have success with BI. Organizations that more frequently use data engineering can better meet the BI users demand for more timely and relevant data.

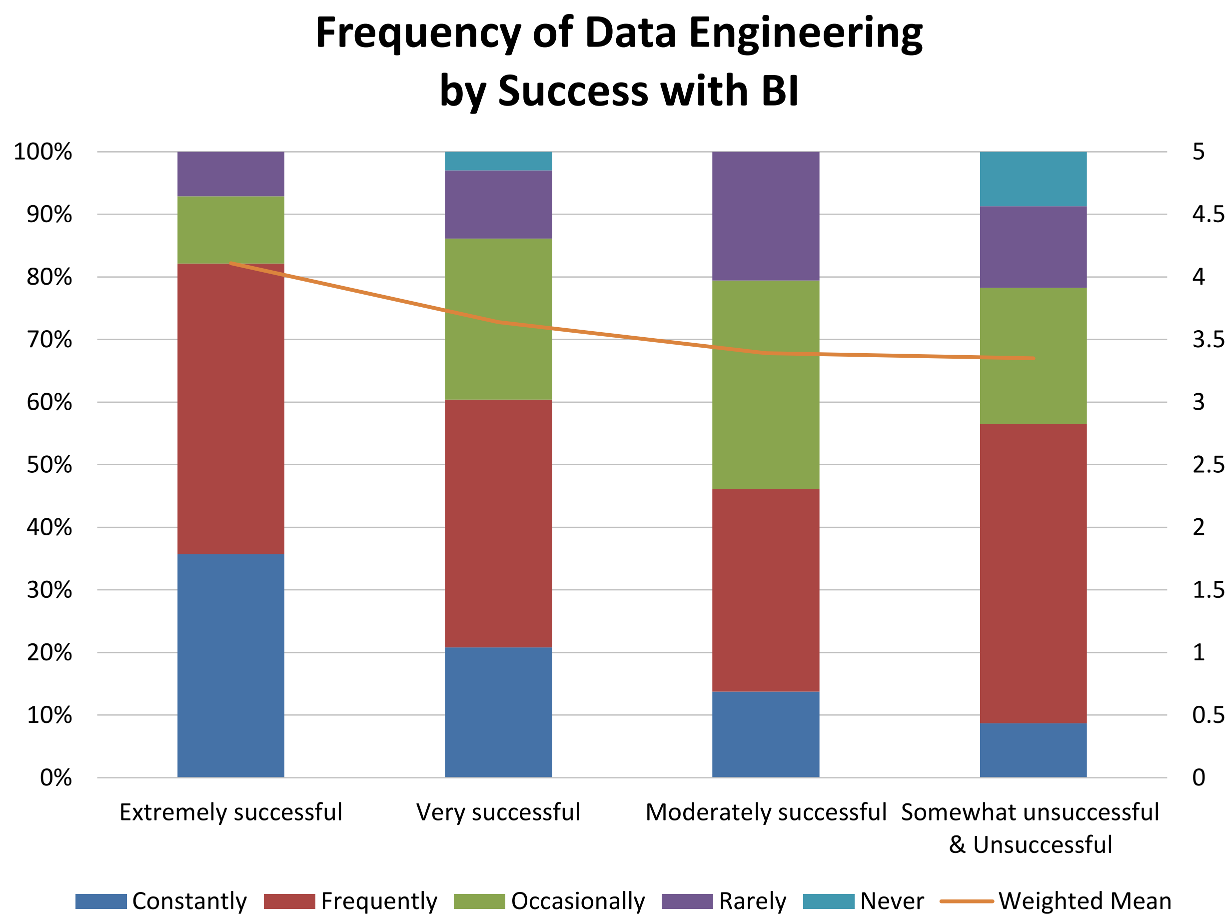


Figure 28 – Frequency of data engineering by success with BI

Data Engineering Usability

Respondents rated the top three usability features to be: *simple to build data workflows*, *scheduler for coordinating data workflows*, and *connectors to simplify data access as critical, very important, or important*. The feature highest rated as critical is *scheduler for coordinating data workflows*. The features most rated as *not important* are *no code data transformations* (also note the low importance of graphical, drag-and-drop capabilities) and *machine learning and recommendations based on usage data gathered across users, groups, or organizations*.

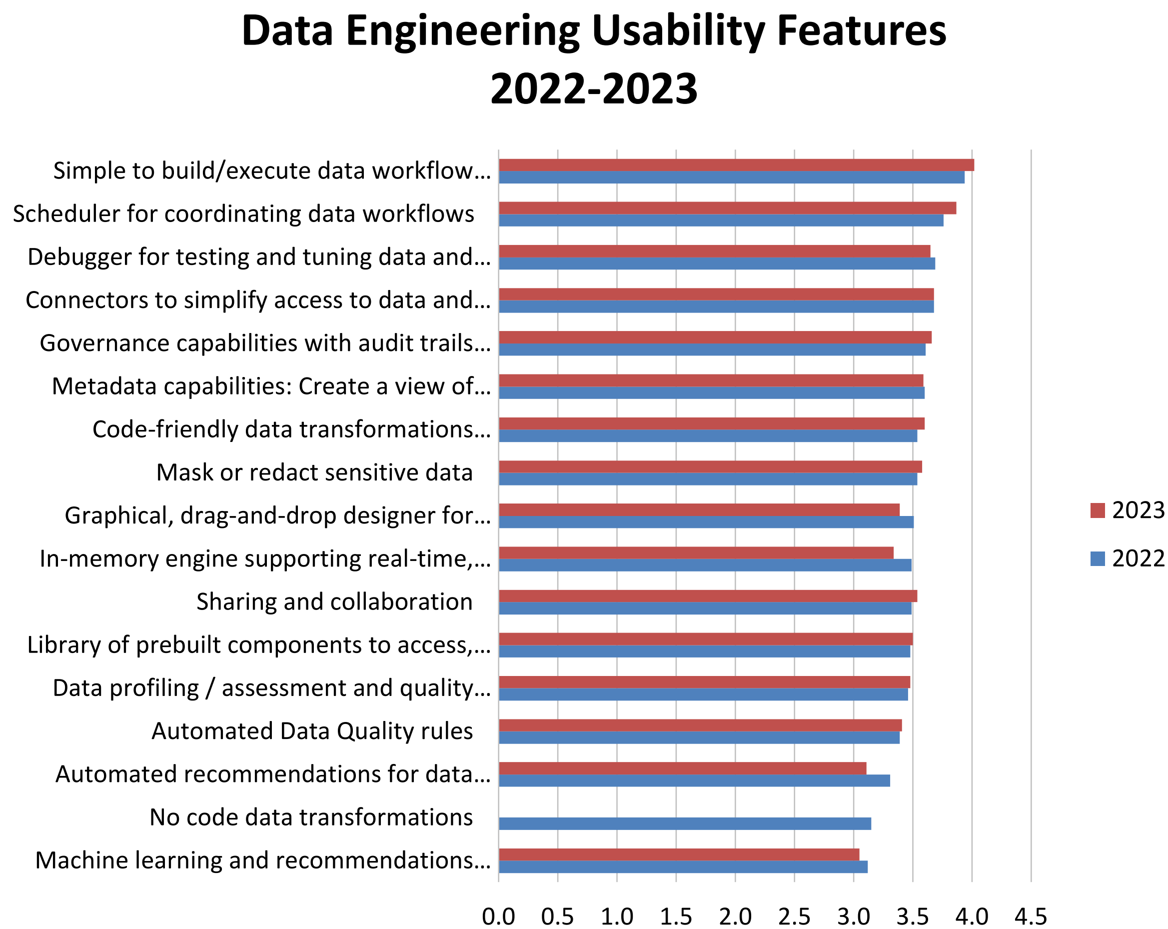


Figure 29 – Data engineering usability features

For all four geographies, *simple to build/execute data workflow scripts* is the highest rated, with average importance scores of above 4 in importance (fig. 30) with *Asia Pacific* respondents rating it the most important. *All geographies* rate *no code data transformation* and *machine learning and recommendations* as the lowest in importance.

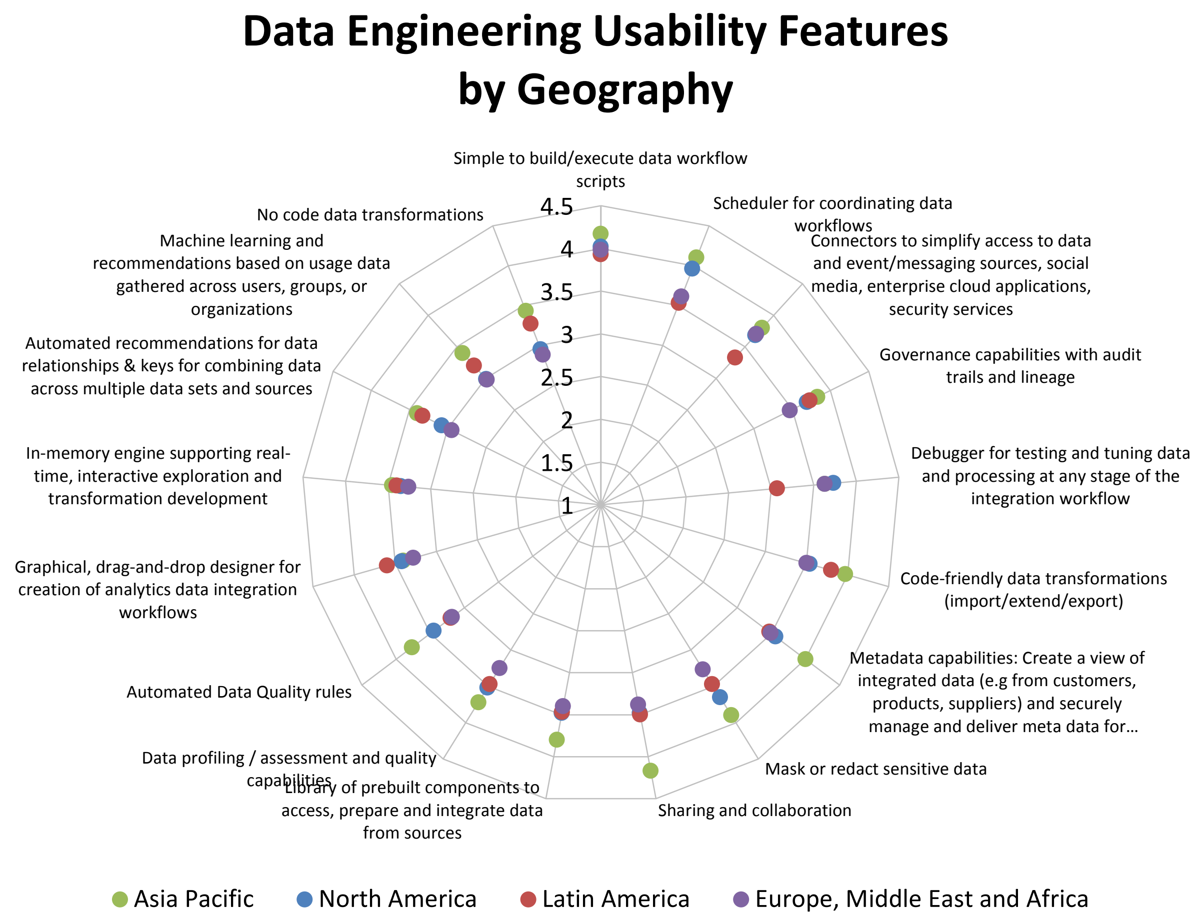


Figure 30 – Data engineering usability features by geography

From an industry perspective, most industries rate *simple to build/execute data workflow scripts* highest in importance for data engineering usability features (fig. 31). Across all industries except *business services*, the lowest importance usability features are those related to *no code data transformations*. *Business services* rates *machine learning and recommendations based on usage data gathered across users, groups, or organizations* as lowest. Conversely, *technology* industry respondents rate the importance of more advanced features such *machine learning and recommendations based on usage data gathered across users, groups, or organizations* and *in-memory features*. *Healthcare* respondents also place a higher-than-industry-average importance on *governance capbilities* and the ability to *mask or redact sensitive data*.

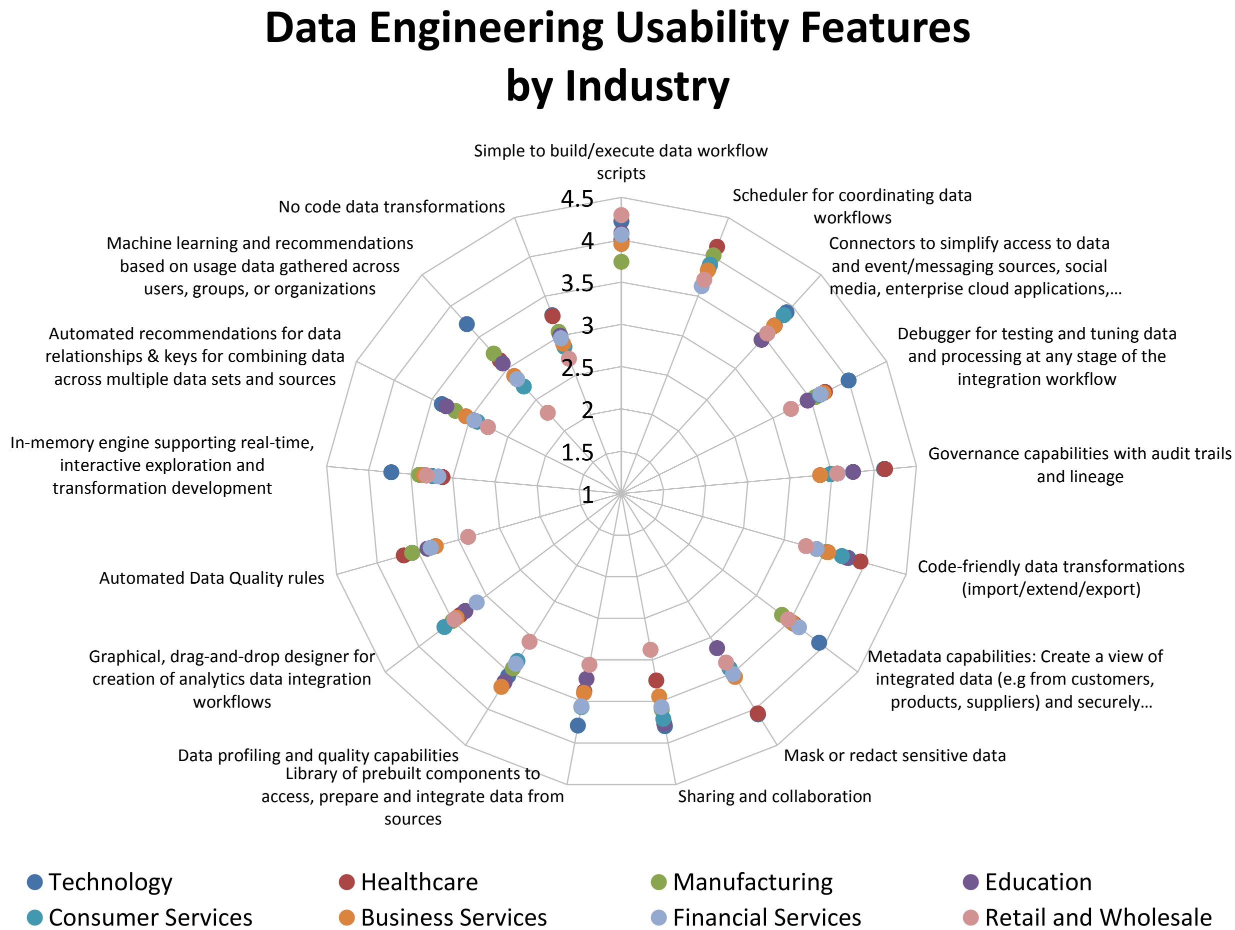


Figure 31 – Data engineering usability features by industry

Respondents in each function rate *simple to build/execute data workflow scripts* as the highest importance for usability features (fig. 32). The lowest importance across all functions, except *finance*, is *no code data transformations*. Outliers include *R&D*'s high importance rating for *machine learning and recommendations*, and *finance* and *operations* respondents' low importance placed on *in-memory* engine capabilities.

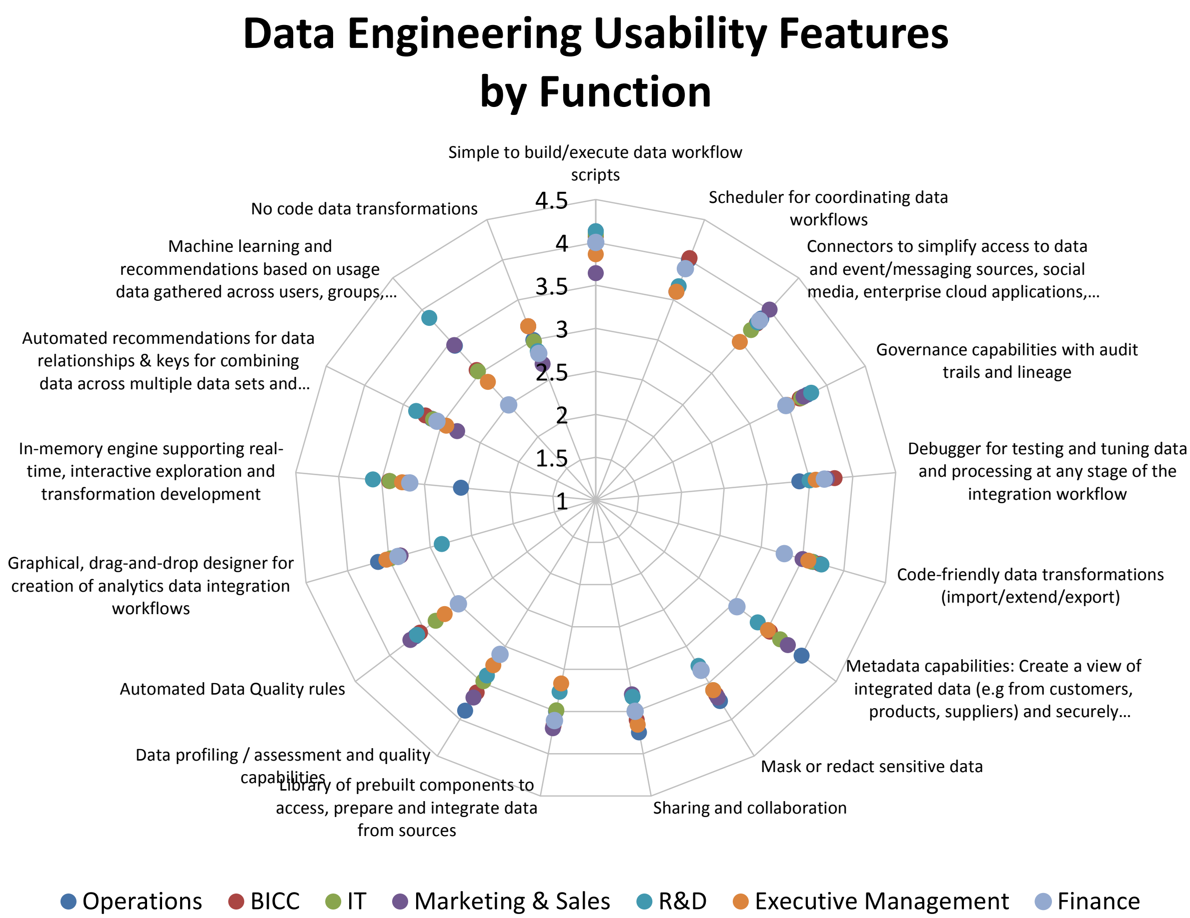


Figure 32 – Data engineering usability features by function

There are small variations on the importance of usability features across organization sizes. Organizations of *all sizes* rate *simple to build/execute data workflow scripts* as their top priority for usability features (fig. 33). *Very large organizations* (more than 10,000 employees) rate *metadata capabilities* as tying for their second-highest rated usability features.

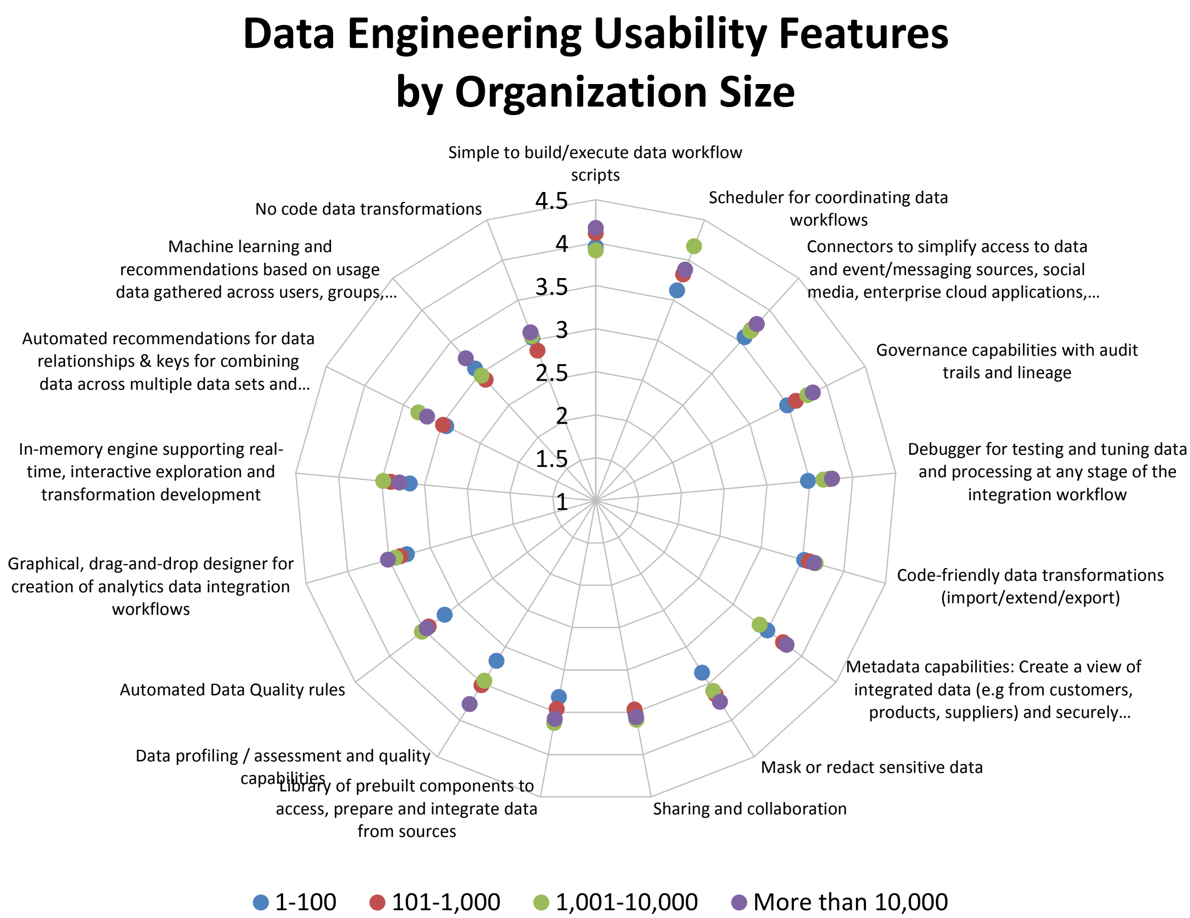


Figure 33 – Data engineering usability features by organization size

Usability features by perceived success with BI indicates is fairly consistent with a few exceptions (fig. 34). In general, organizations that are more *successful* with BI rate most usability features higher than their peers, including more advanced features such as *automated data quality rules* and *machine learning recommendations*. *Less successful* organizations place a greater importance on a *library of prebuilt components to access, prepare and integrate data from sources*.

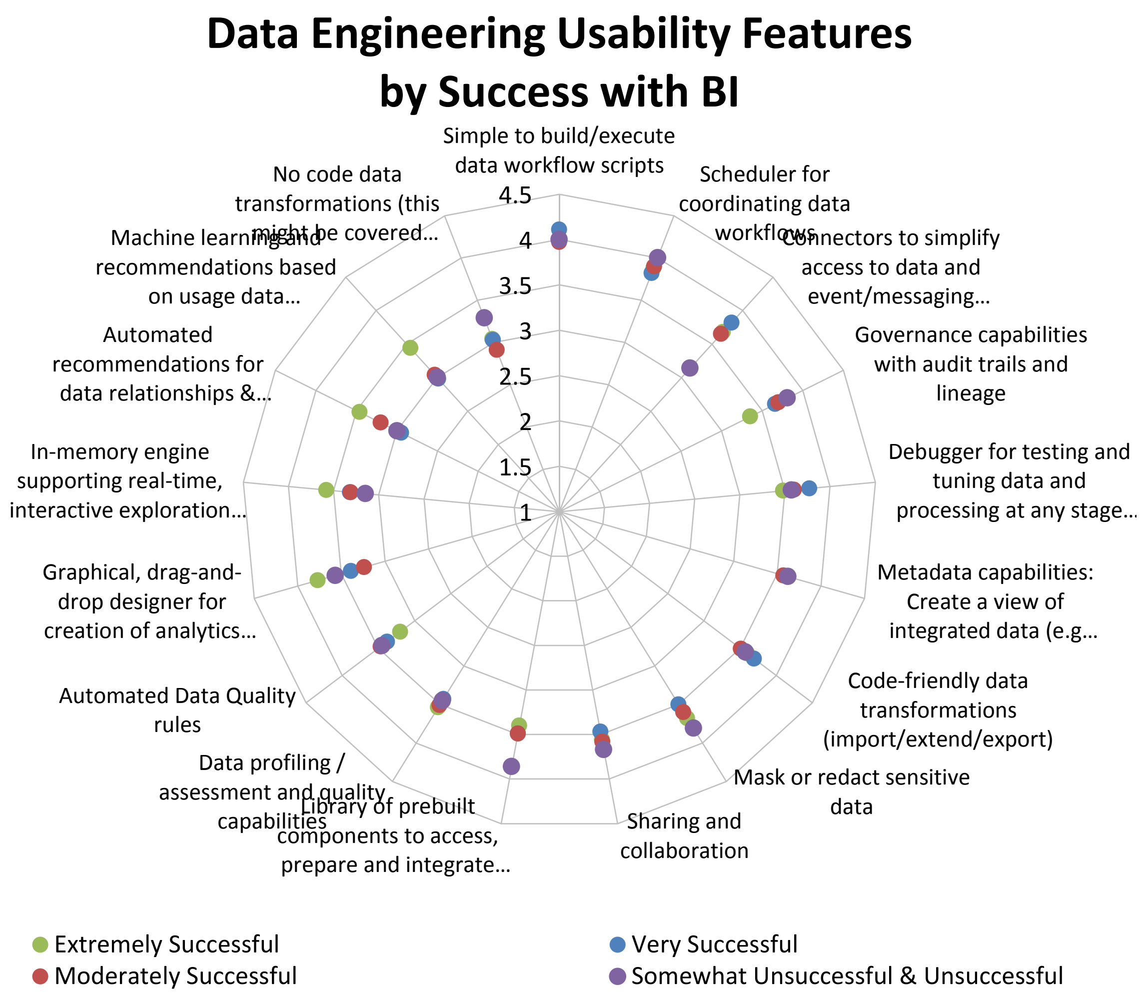


Figure 34 – Data engineering usability features by success with BI

Frequency of Enrichment with Third-Party Data

We asked our survey respondents about the use and frequency of data enrichment with third-party data as part of data engineering workflows (fig 35). Most organizations (35 percent) *occasionally* enrich their data engineering workflows with third-party data. Eleven percent *constantly* enrich with third-party data, and 26 percent *frequently* enrich their data sets with third-party data as part of their data engineering workflows. Only seven percent of organizations never use third-party data for enrichment.

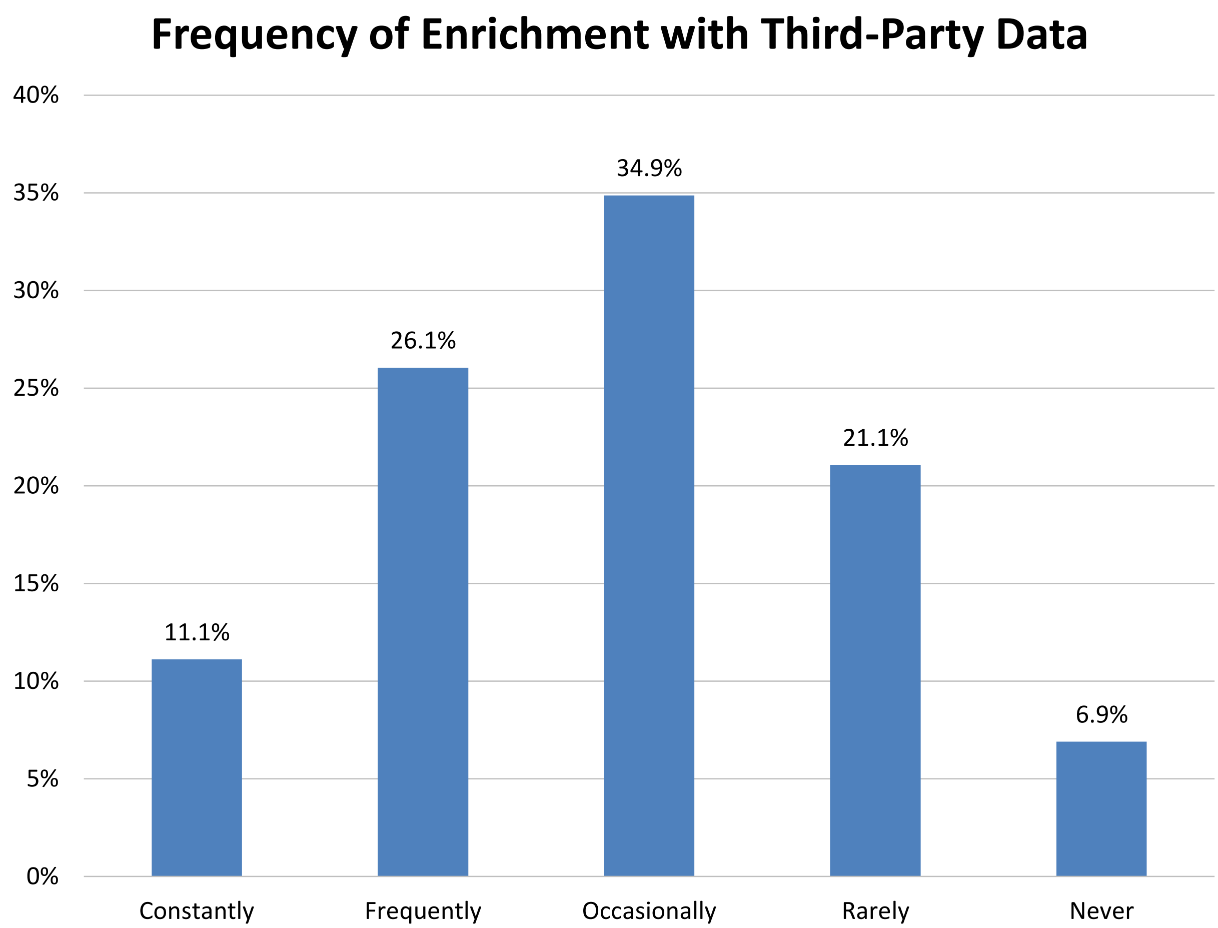


Figure 35 – Frequency of enrichment with third-party data

Asia Pacific respondents enrich their data sets with third-party data more often than other geographies. *Latin American* respondents are the least likely to enrich their data engineering workflows with third-party data, and 27 percent of the respondents indicate that they *never* enrich their data sets using third-party data.

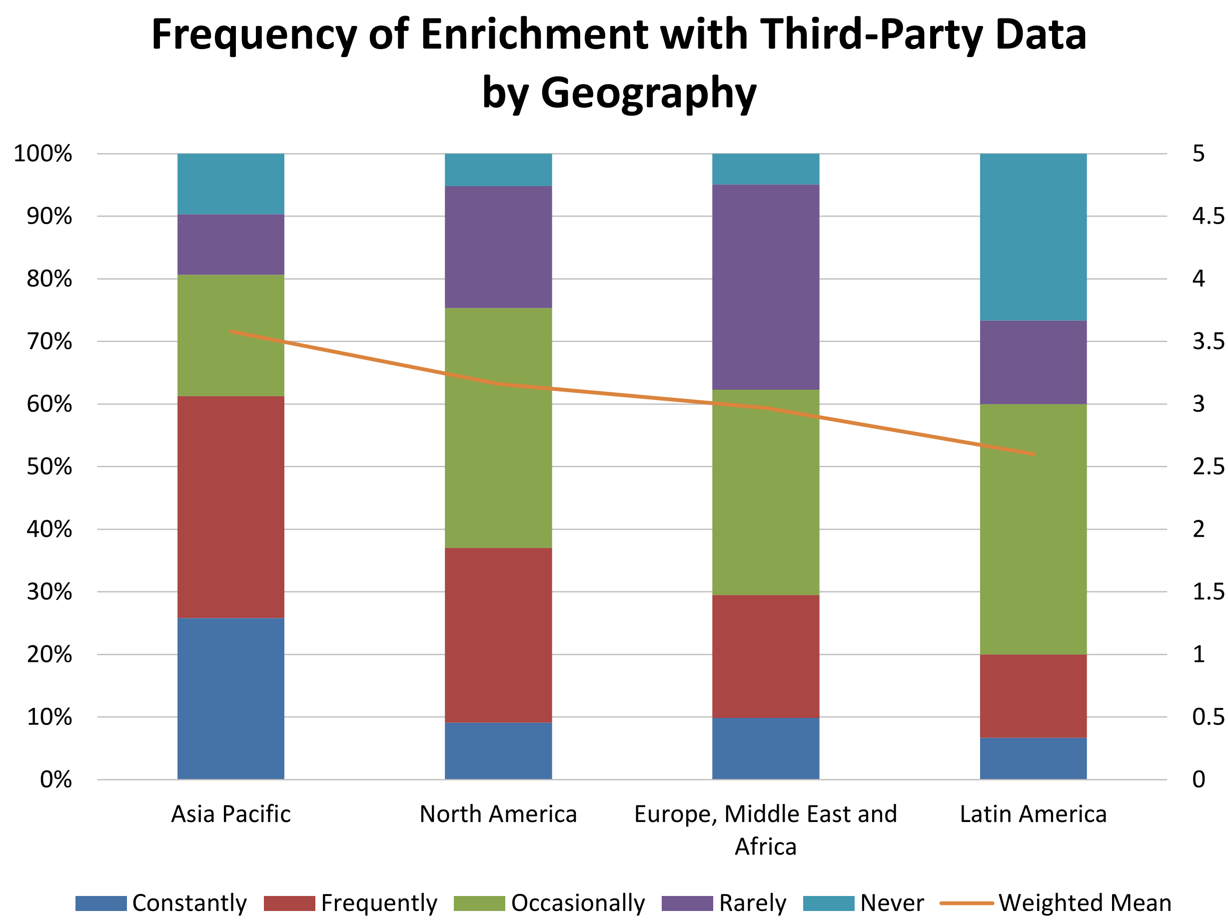


Figure 36 – Frequency of enrichment with third-party data by geography

Healthcare, *technology*, and *manufacturing* respondents use third-party data to enrich their data engineering workflows the most, compared to other industries (fig. 37). *Education* industry respondents are the most likely to *never* use third-party data enrichment.

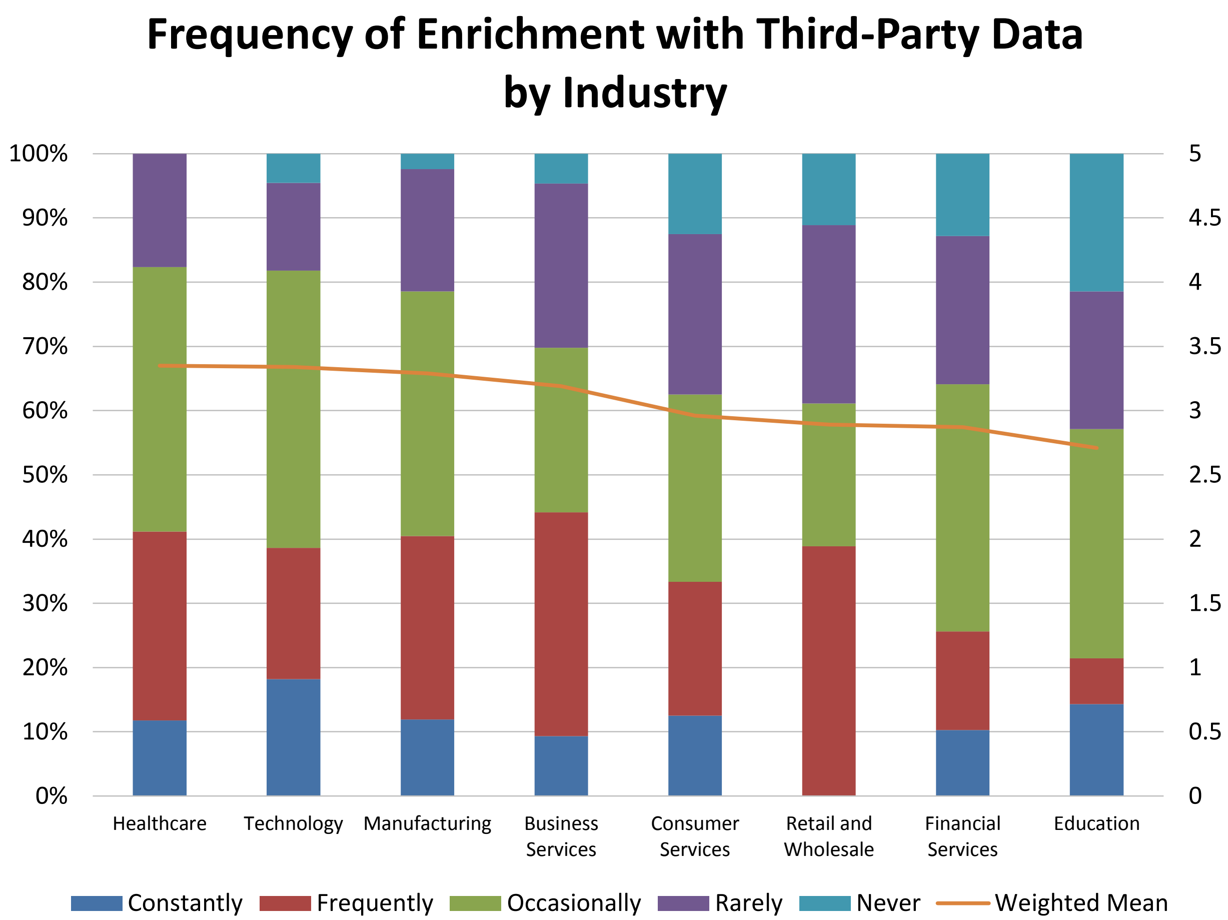


Figure 37– Frequency of enrichment with third-party data by industry

Operations functions have the highest use of enrichment of their data sets with third-party data via their data engineering workflows (fig. 38). (Also note the importance of data engineering to operations respondents in fig 6.) *Finance*, which mostly relies on internal revenue and expense data sets places the lowest frequency and importance on data enrichment from third-party sources.

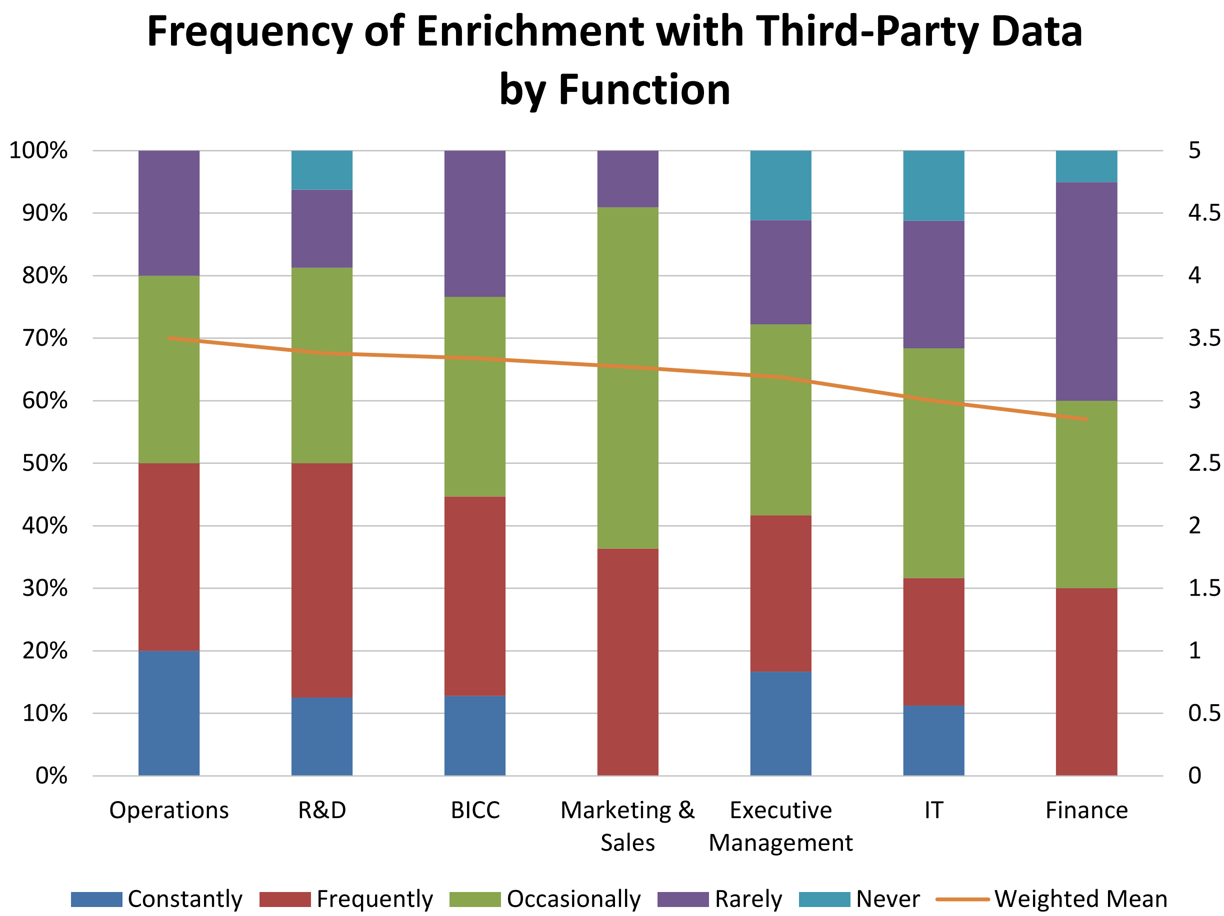


Figure 38 – Frequency of enrichment with third-party data by function

Enrichment of data engineering pipeline workflows with third-party data varies by organization size (fig. 39). The distribution of responses appears bimodal. The *smallest organizations* (1-100 employees) and the *largest organizations* (more than 10,000 employees) indicate the highest frequency of enrichment of their data sets with third-party data.

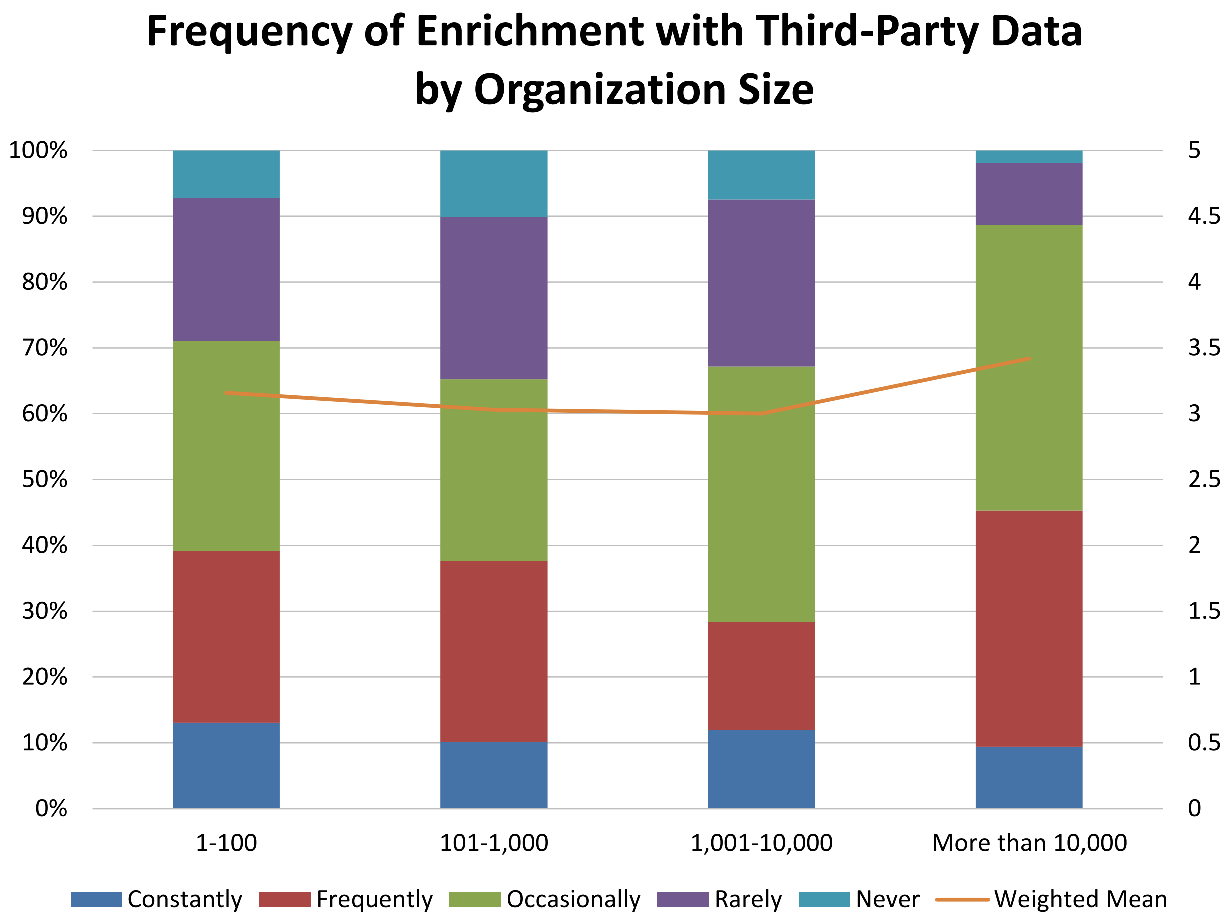


Figure 39 – Frequency of enrichment with third-party data by organization size

There appears to be a reasonable correlation between BI success and the use of third-party data set enrichment with internal data sets (fig. 40). Successful BI initiatives have the highest use of third-party data.

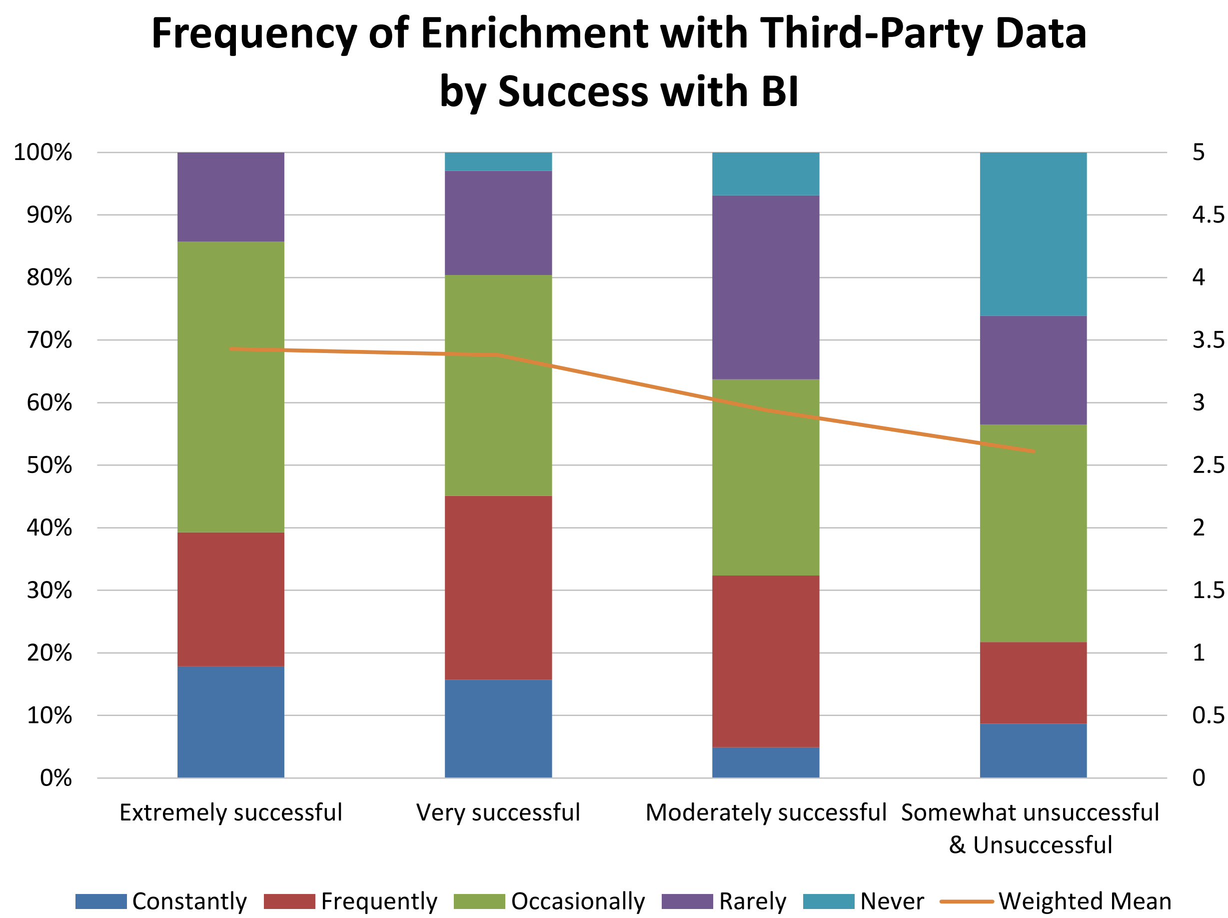


Figure 40 – Frequency of enrichment with third-party data by success with BI

Data Sources and Targets for Data Engineering

Overall, respondents indicate that *relational databases*, *file systems,* and *applications* are the top data sources for data engineering workflows. Of note is *object stores* (e.g. Amazon S3, Google Cloud Storage, Microsoft Azure blob storage) are the predominant target for data engineering workflows but also highly used as a source. This indicates their role as an offline storage and staging repository for data engineering workflows.

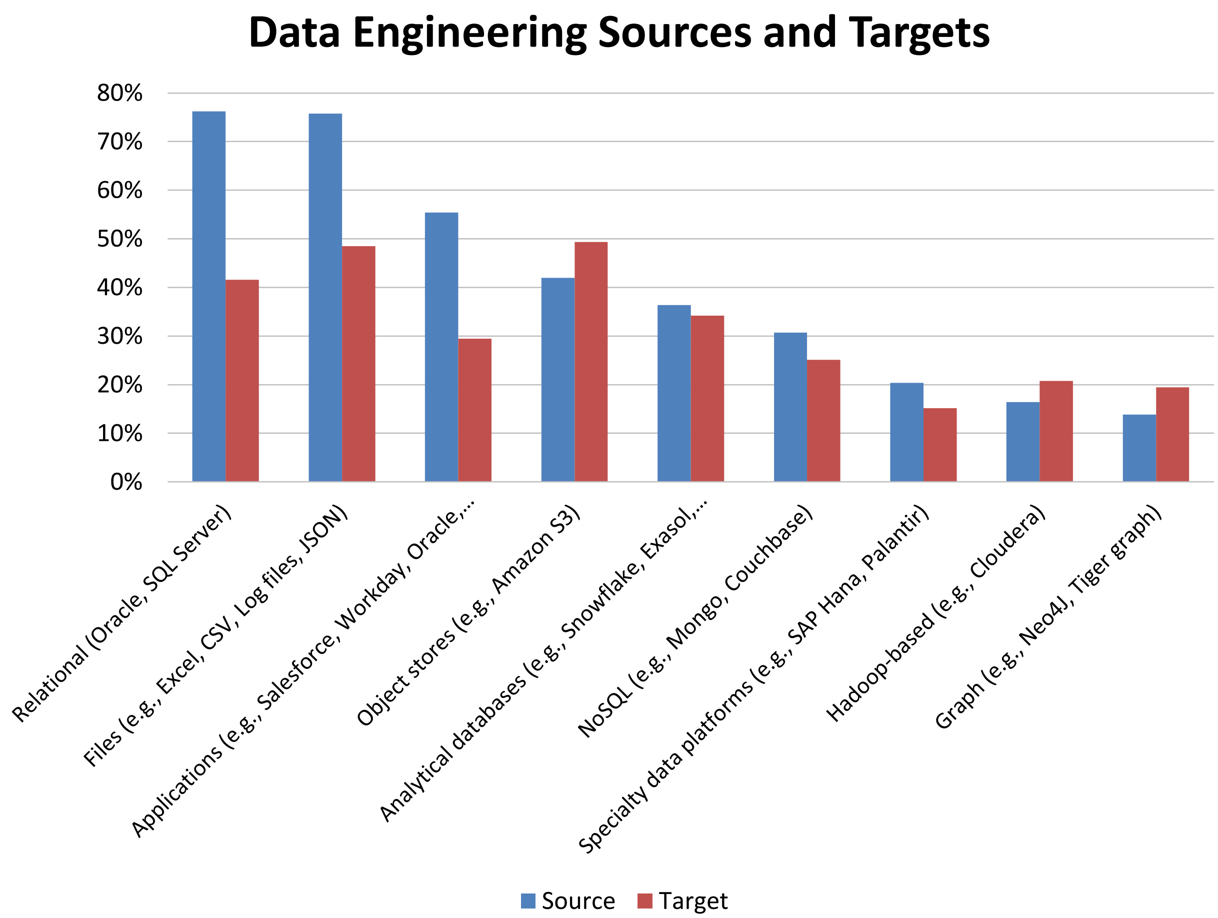


Figure 41 – Data engineering sources and targets

The use of *files* (e.g., Excel, CSV, Log files, JSON), *object Stores,* and *analytical databases* as sources of data engineering (fig. 42) is on the increase. The rising use of object Stores and analytical databases in data engineering workflows reflects the rising use of cloud versus on-premises-based analytical data sources.

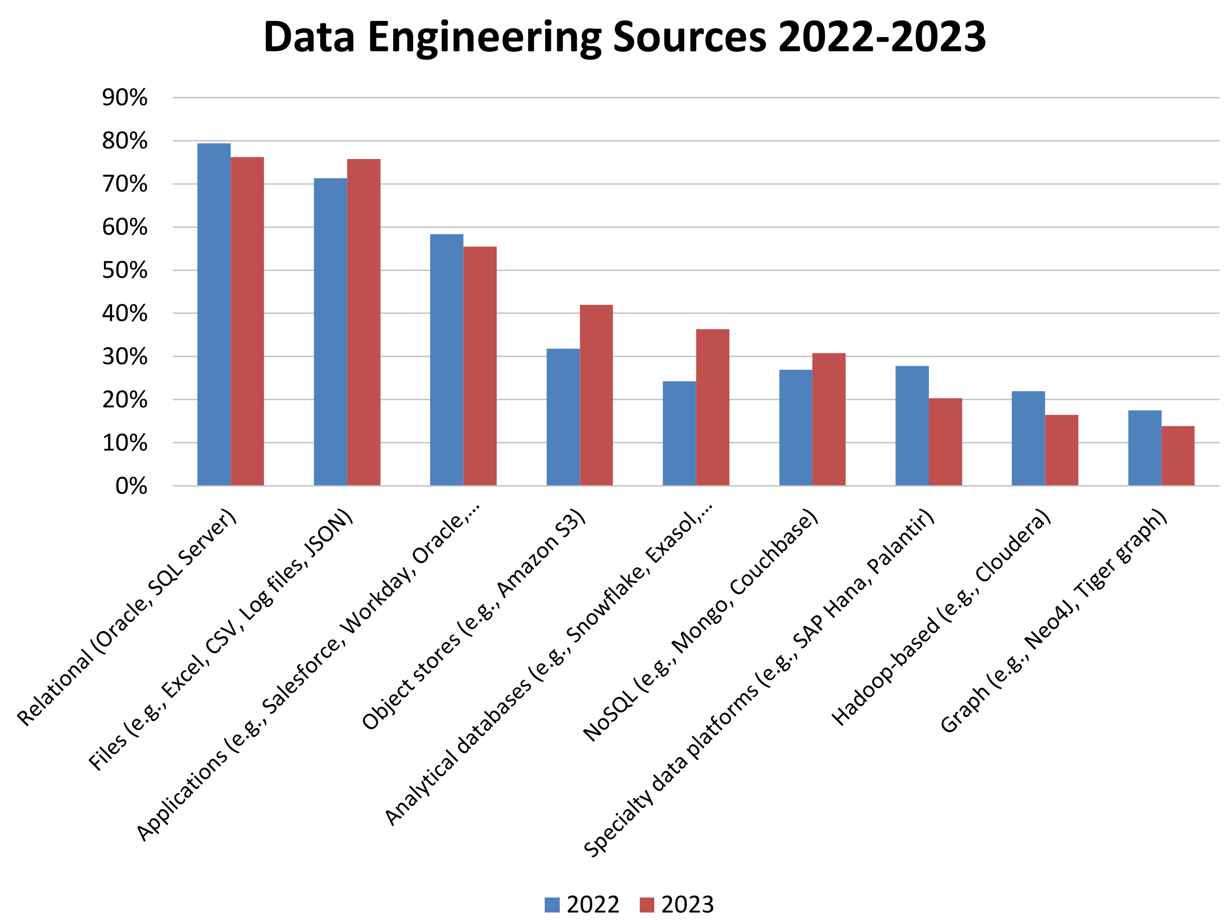


Figure 42 – Data engineering sources 2022-2023

The trend for the targets of data engineering workflows is on the increase for most targets, except for *relational* and *graph* databases (fig. 43). *Analytical* databases show the largest increase as targets year over year.

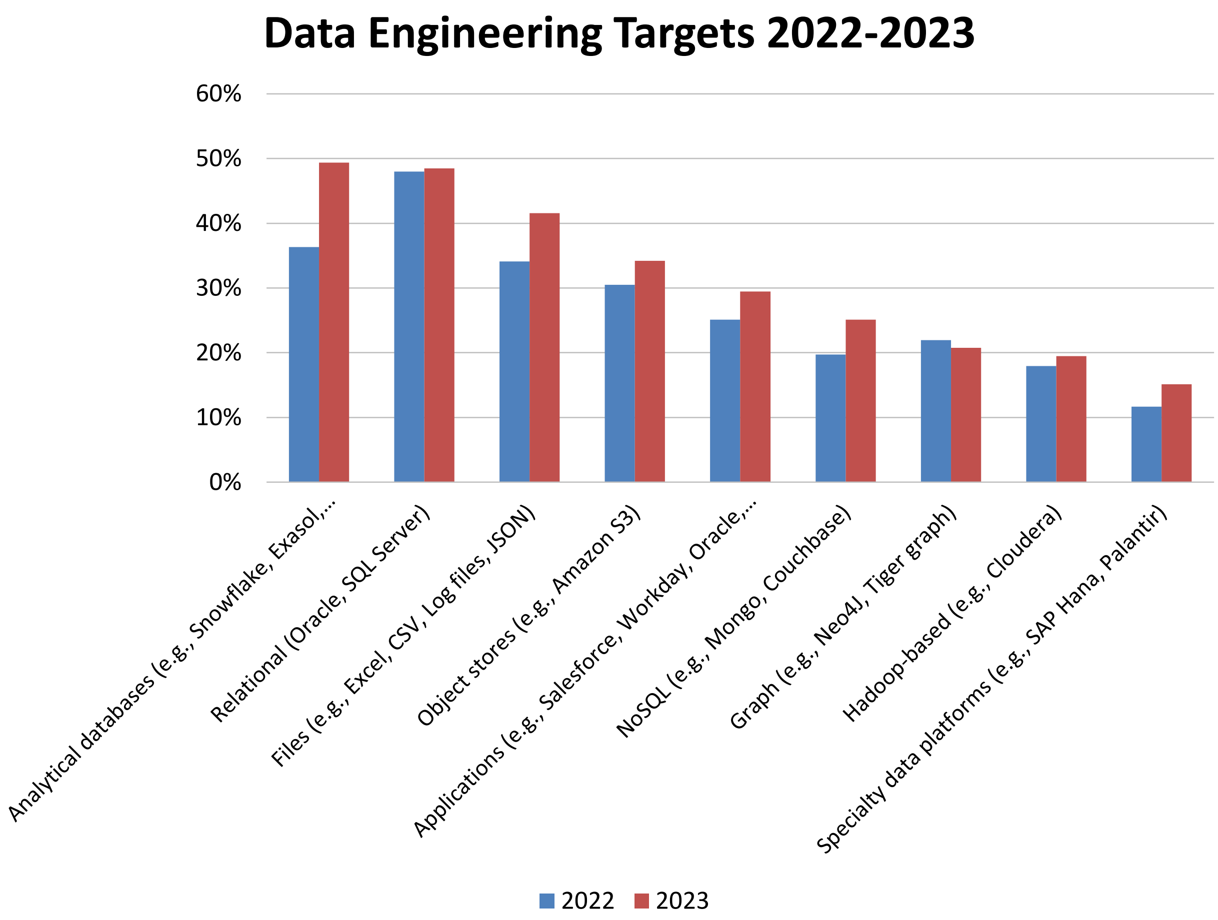


Figure 43 – Data engineering targets 2022-2023

Because of its legacy, *relational* databases remain the top source across *all geographies*. The next most common sources for data engineering workflows are *files* (especially in NA) and *applications* (fig. 44). Note the lower-than- market-average response from *EMEA* respondents for applications as sources. *Hadoop=based* and *graph* data sources are the least common data sources across all geographies.

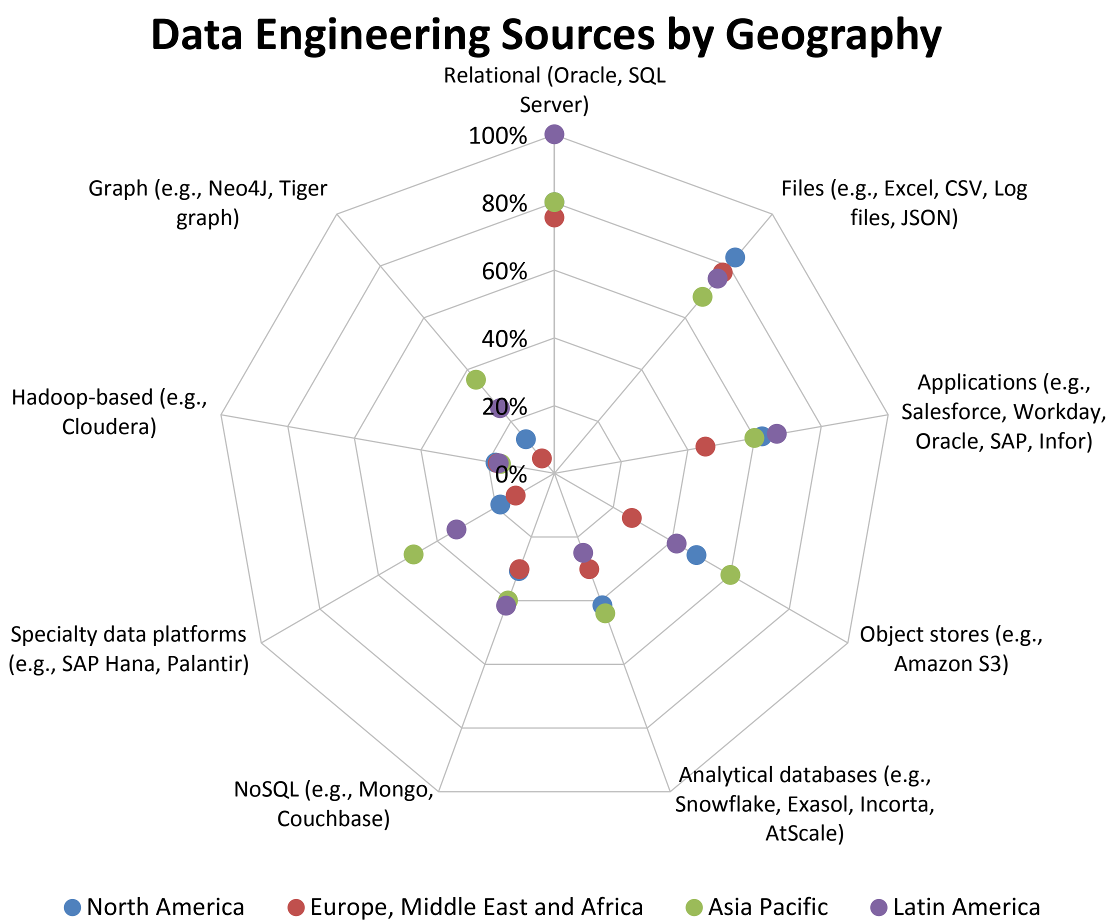


Figure 44 – Data engineering sources by geography

Last year, *relational* databases were the most-used target for data engineering (fig. 45)[but this year, the *analytical* databases are more likely to be the target. Of all geographies we study, *Asia Pacific* places a high priority on several more modern targets including *analytical databases*, *object stores*, and *graph*. In contrast, *Latin America* places a low priority for most of these. Both *North America and EMEA* report mixed priorities for both traditional and modern targets.

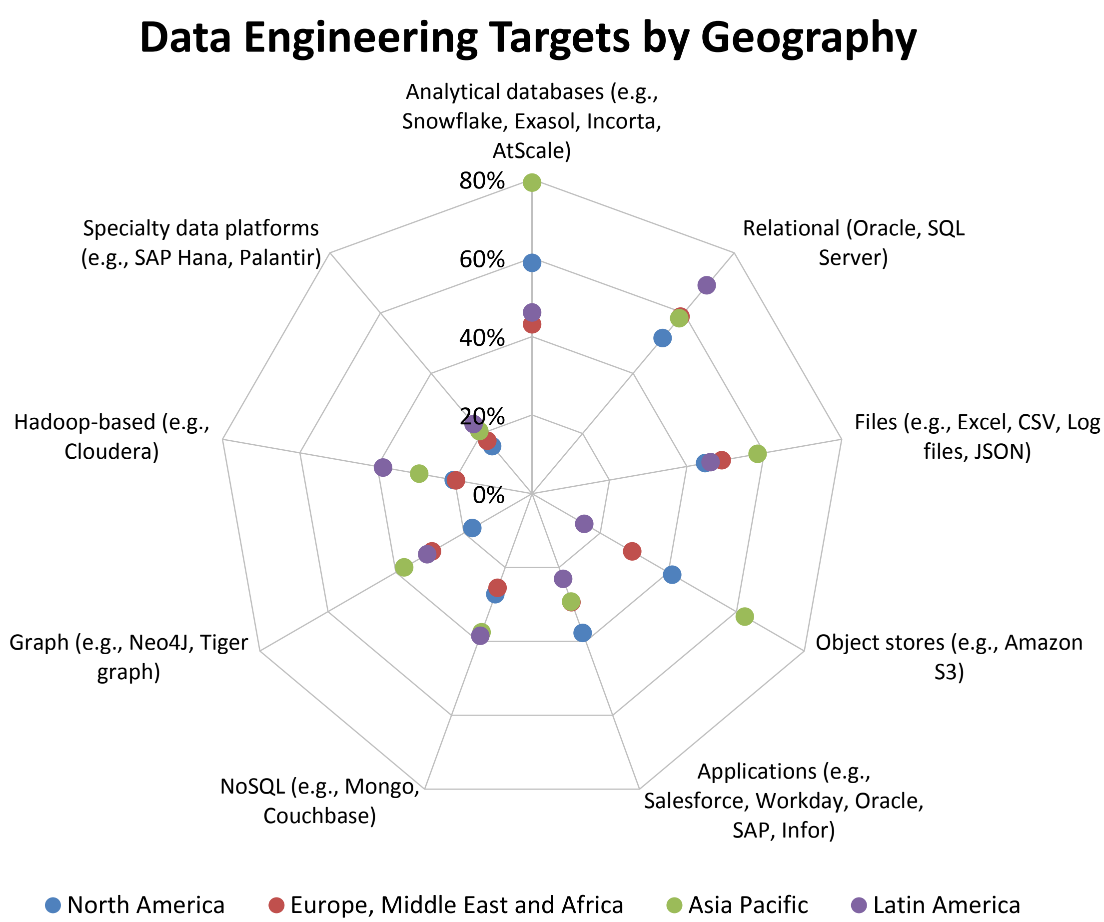


Figure 45 – Data engineering targets by geography

All industries represented in the study indicate that *relational* and *files* are the top two sources, except for *consumer services*, where files, followed by *applications* are their top two priorities (fig. 46). Also noteworthy is *technology* placing a relatively high priority upon object stores and *analytical databases*, and *manufacturing* reporting a greater preference for specialty data platforms. *Consumer services and financial services* place higher priority on *object stores* at 60 percent and 58 percent, respectively.

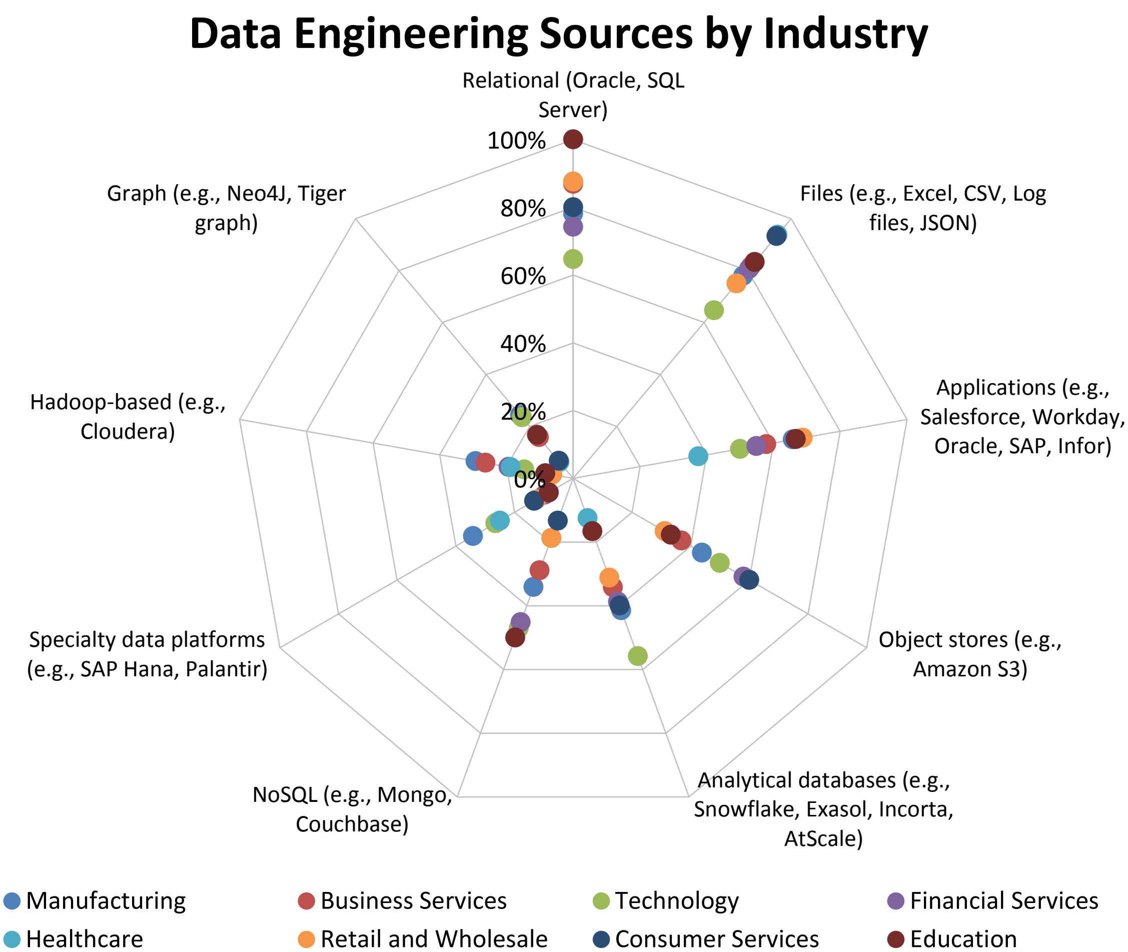


Figure 46 – Data engineering sources by industry

Overall, when we slice the use of sources and targets for data engineering by the industry of the respondents, we see that the most used target overall for data engineering is *analytical databases* (fig. 47). *Relational databases* are the second most important targets for data engineering. No industries rate specialty platforms as highly used targets.

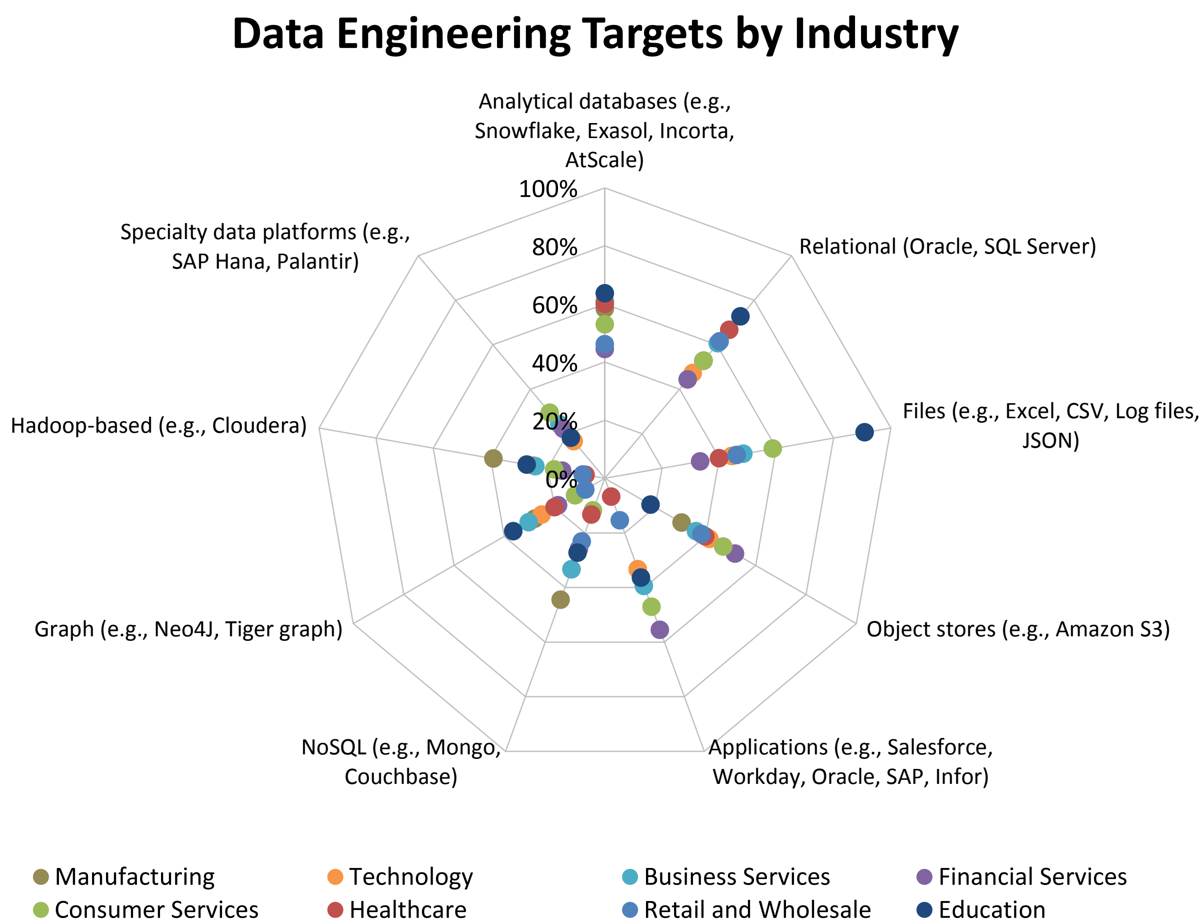


Figure 47 – Data engineering targets by industry

With the exception of *operations*, all other functions represented in the study indicate that *relational databases* is the most used source for data engineering. Operations indicates it most-used data engineering sources and importance are *files* and *applications* (fig. 48) and rates other high diversity of data sources including *graph* and *Hadoop-based* data sources. This diversity of sources mirrors the higher-than-average importance placed on data engineering by *operations*.

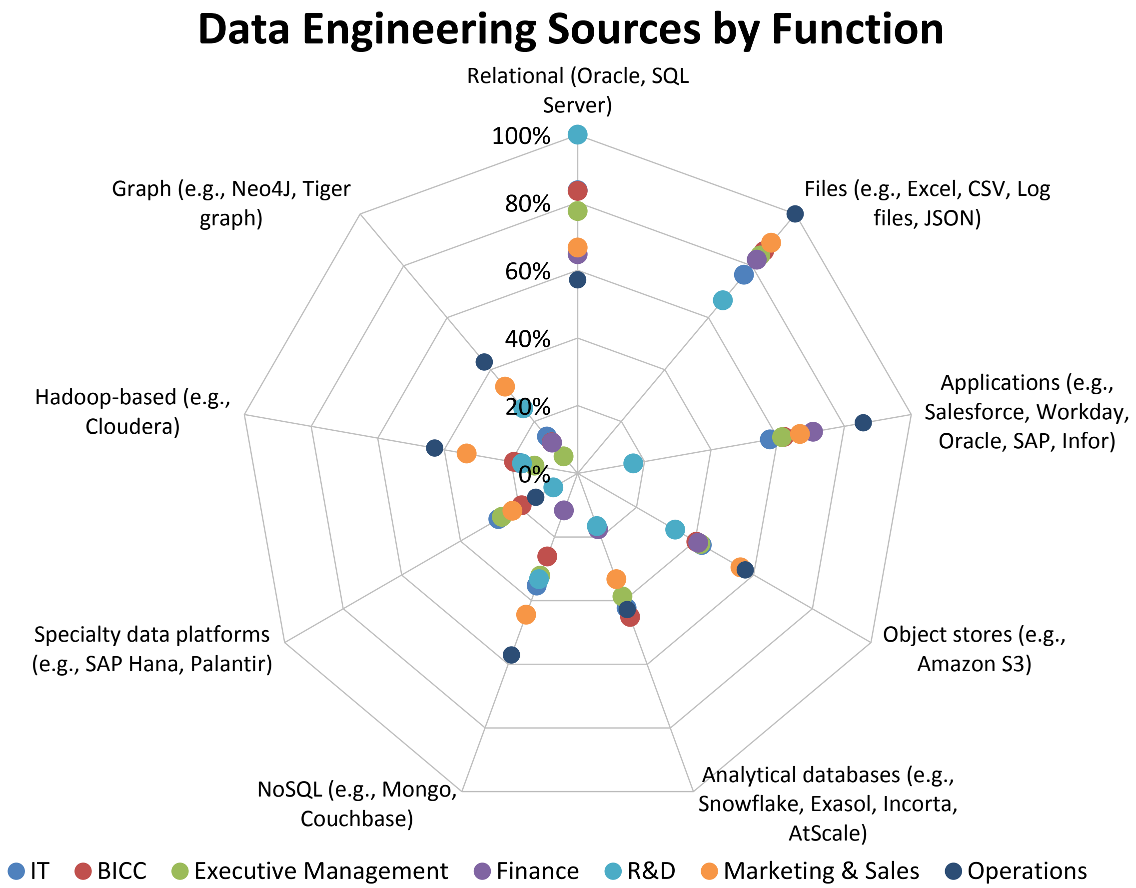


Figure 48 – Data engineering sources by function

All functions represented in the survey indicate that the most used targets for data engineering are *analytical databases*, followed by *relational databases*. Specialty data platforms and Hadoop-based data sources are the lowest targets by function (fig. 49).

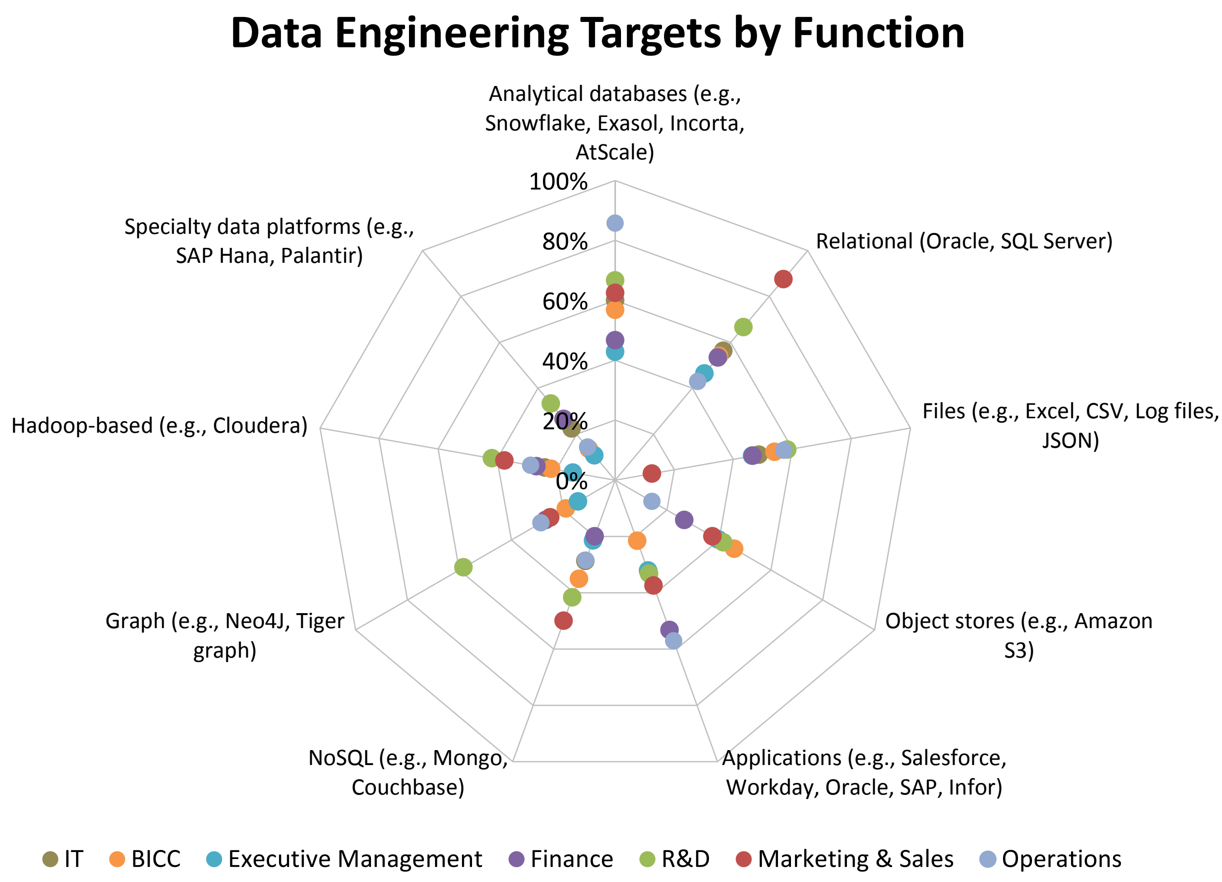


Figure 49 – Data engineering targets by function

It is no surprise that the *relational database*, with its longstanding dominance in the industry, is the top priority source across *all organization sizes* (fig. 50). However, with increasing organization size comes greater source diversity, including *Hadoop and specialty data platforms*.

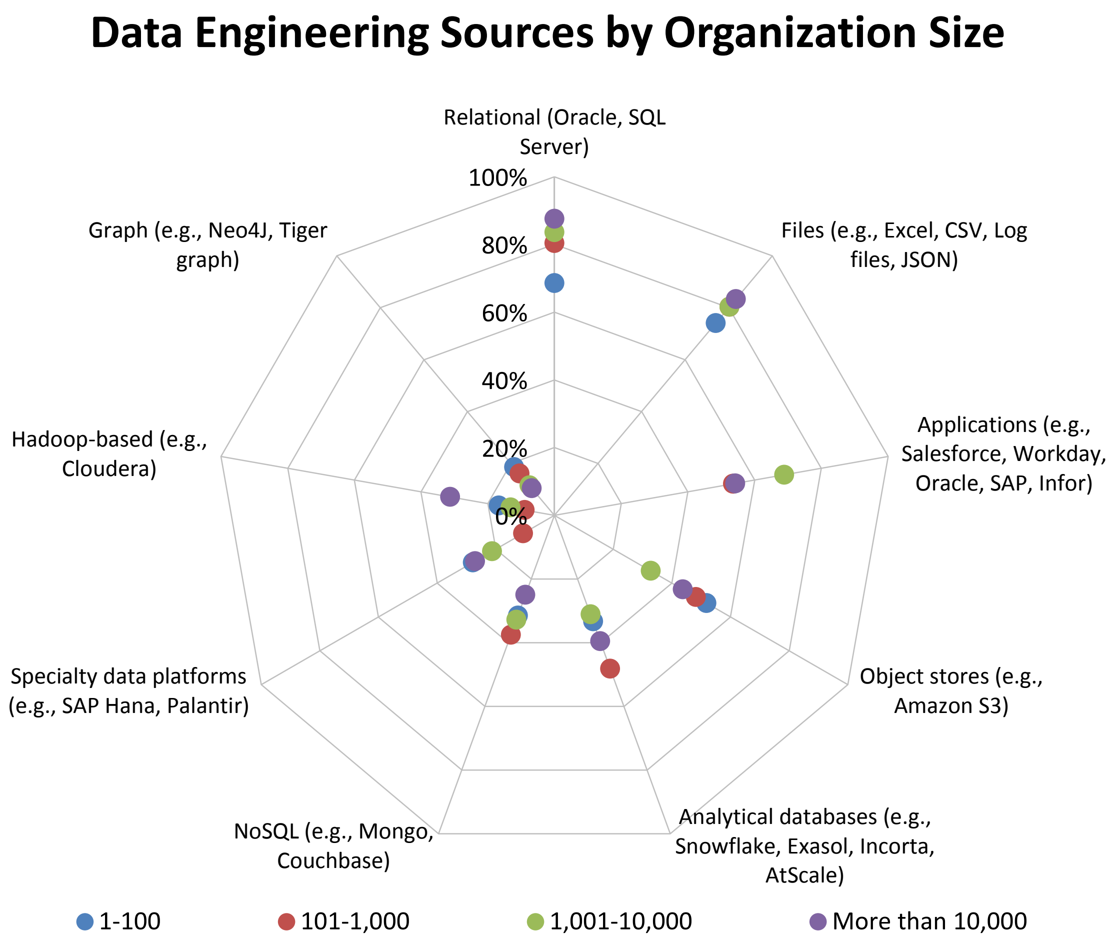


Figure 50 – Data engineering sources by organization size

Consistent with the data in other parts of this report, *organizations of all sizes* indicate that the most important/used target for data engineering is *analytical databases* (fig. 51). All sizes of organizations indicate that *specialty data platforms* are the least important targets.

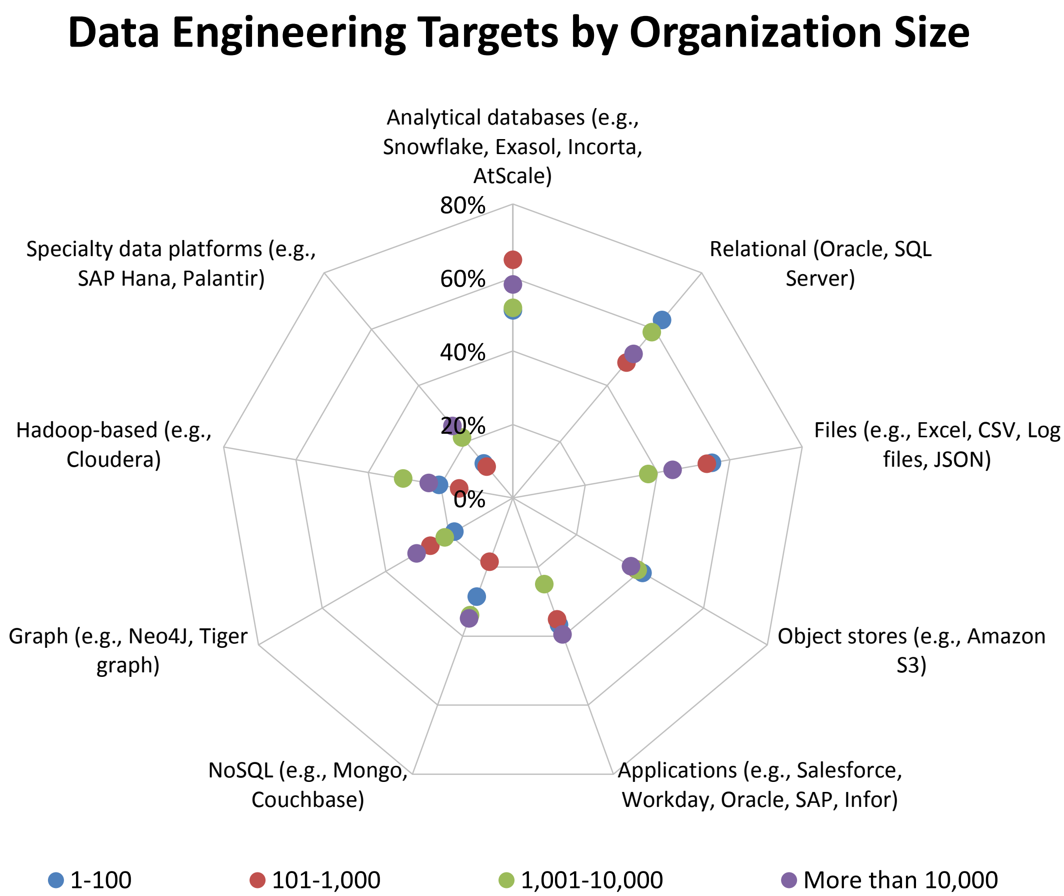


Figure 51 – Data engineering targets by organization size

Effectiveness of Current Approach to Data Engineering

There appears to be room for improving current approaches to data engineering. Twenty six percent of our survey respondents rate their current approach to data engineering as highly effective (fig. 52). Fifty seven percent of our survey respondents rate their current approach to data engineering as somewhat effective. Seventeen percent of the respondents rate their current approach as somewhat or totally ineffective.

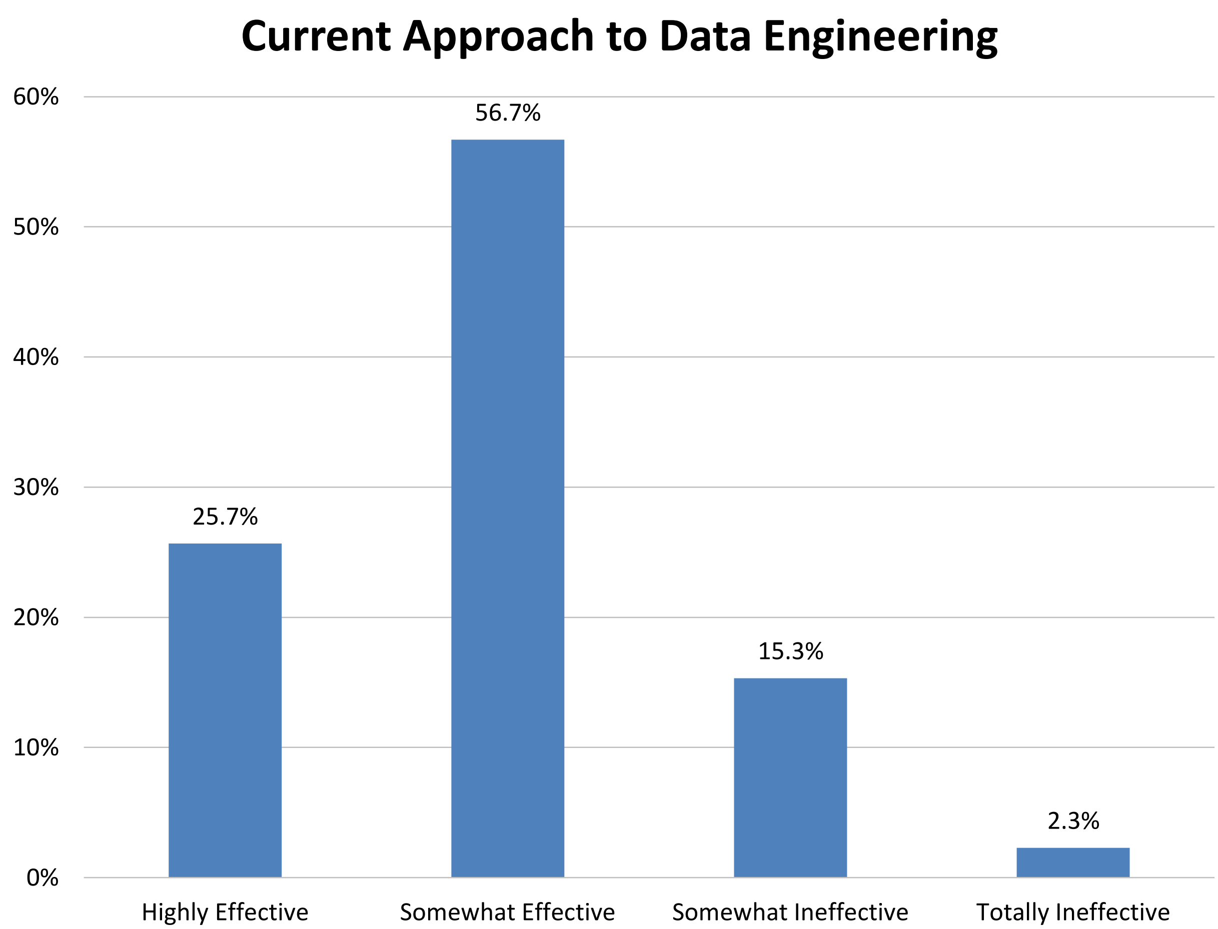


Figure 52 – Current approach to data engineering

Organizations are becoming more effective with their approaches to data engineering. Looking at the levels of effectiveness for approaches organizations use for data engineering, we see a trend of increasing effectiveness overall (fig. 53). The percent of organizations that are *highly effective* is on an upward trend, and the percent of organizations that are somewhat ineffective or totally ineffective is on a declining trend.

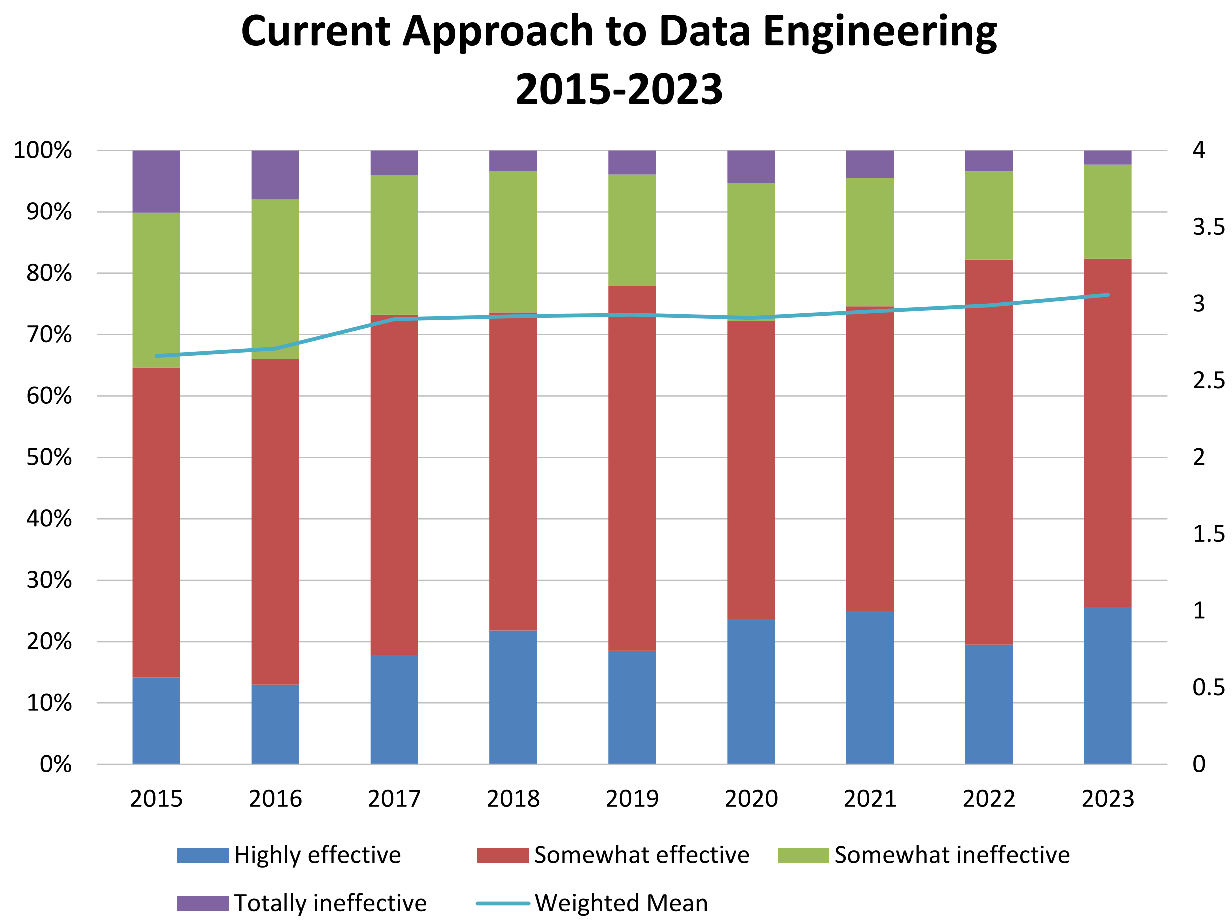


Figure 53 – Current approach to data engineering 2015-2023

Respondents from *Asia Pacific and North America* are the most satisfied with their current approach to data engineering. Respondents from *EMEA and Latin America* are less likely to rate their data engineering approaches as effective (fig. 54).

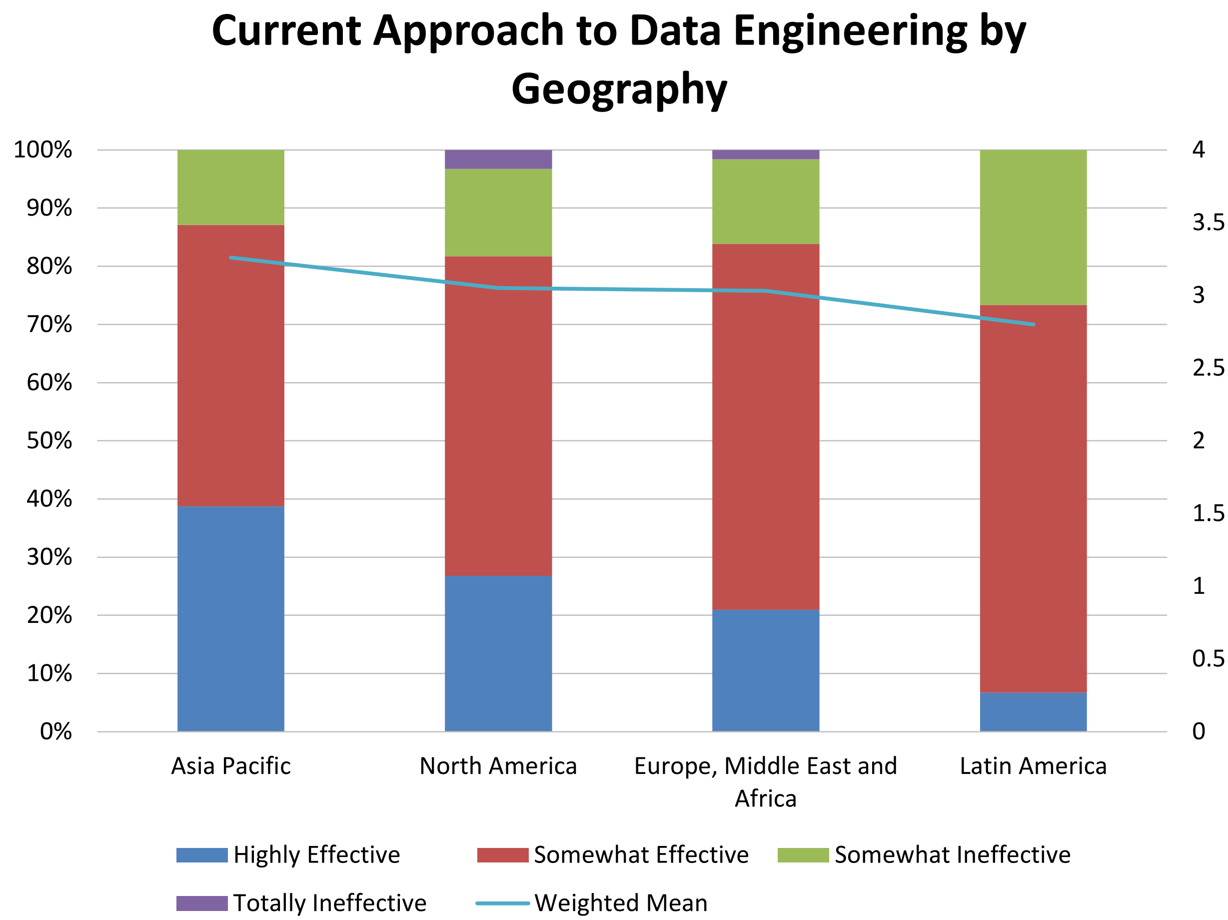


Figure 54 – Current approach to data engineering by geography

Midsized organizations (less than 1,000 employees) say they are *less effective* with their data engineering approaches compared to other organizations (fig. 55). This is not surprising, given smaller organizations most likely have smaller staffs and lower skill levels available to support data engineering efforts. The *smallest organizations* most likely use scripts, spreadsheets, and self-service data-prep tools for data engineering.

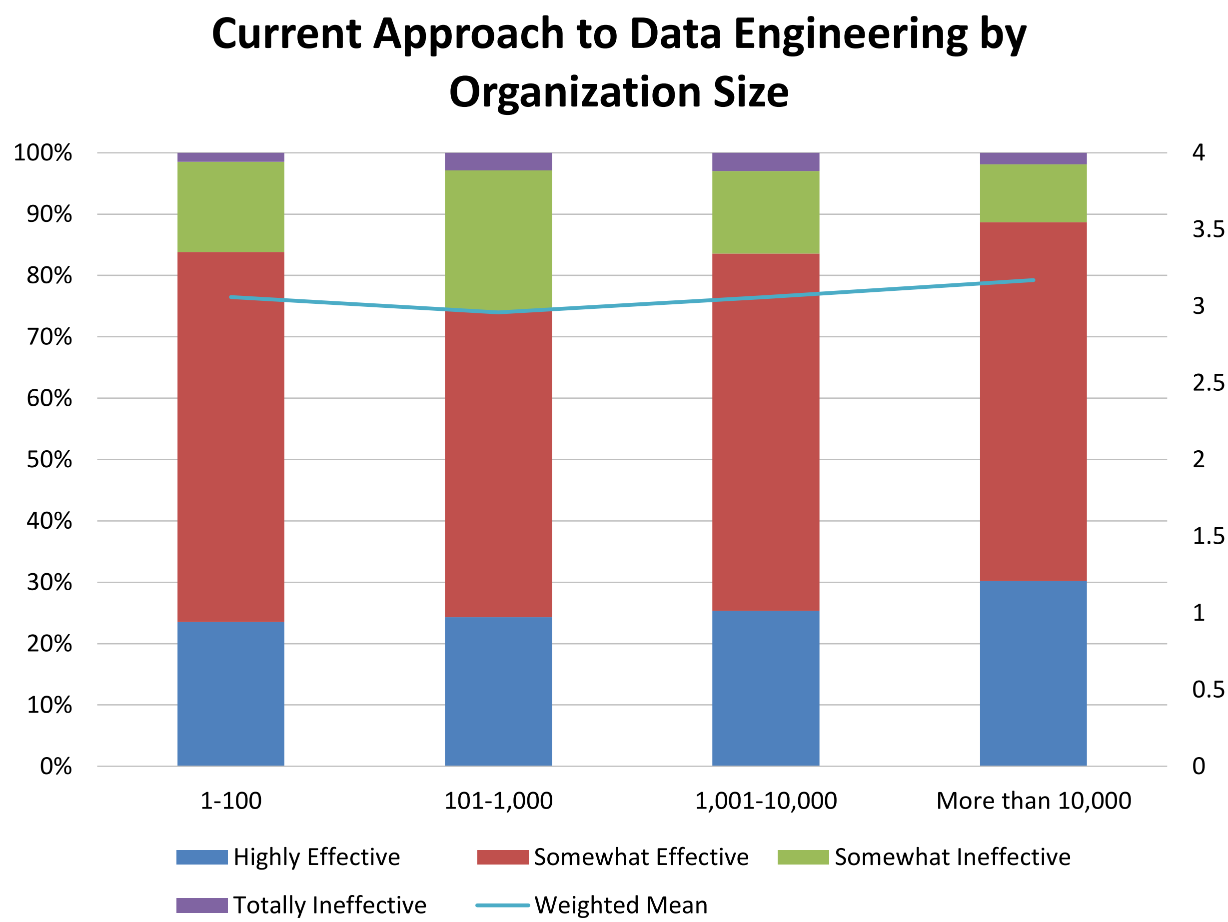


Figure 55 – Current approach to data engineering by organization size

As shown in the chart below, the level of effectiveness with data engineering approaches can materially affect the overall success of a BI initiative (fig. 56). Organizations cannot overlook investments in data engineering approaches and skills when evaluating how to make BI initiatives more successful.

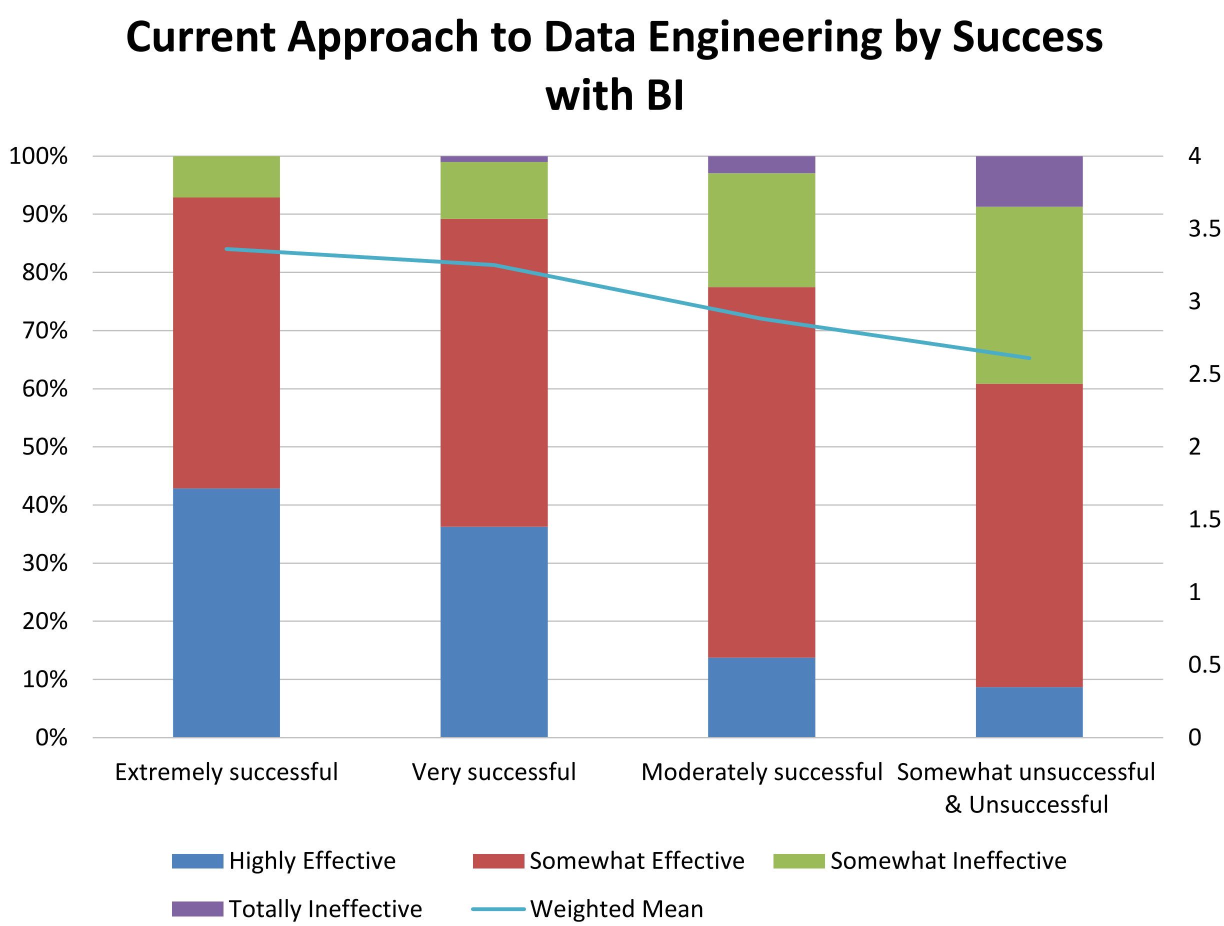


Figure 56 – Current approach to data engineering by success with BI

Deployment of Data Engineering Capabilities

The most common deployment priorities for data engineering capabilities are via a *private or public cloud* (fig. 57). This follows the larger trend we see in the priority for deployment of Analytic Data Infrastructures to the cloud. *On-Premises* deployments are the lowest priority for data engineering deployments.

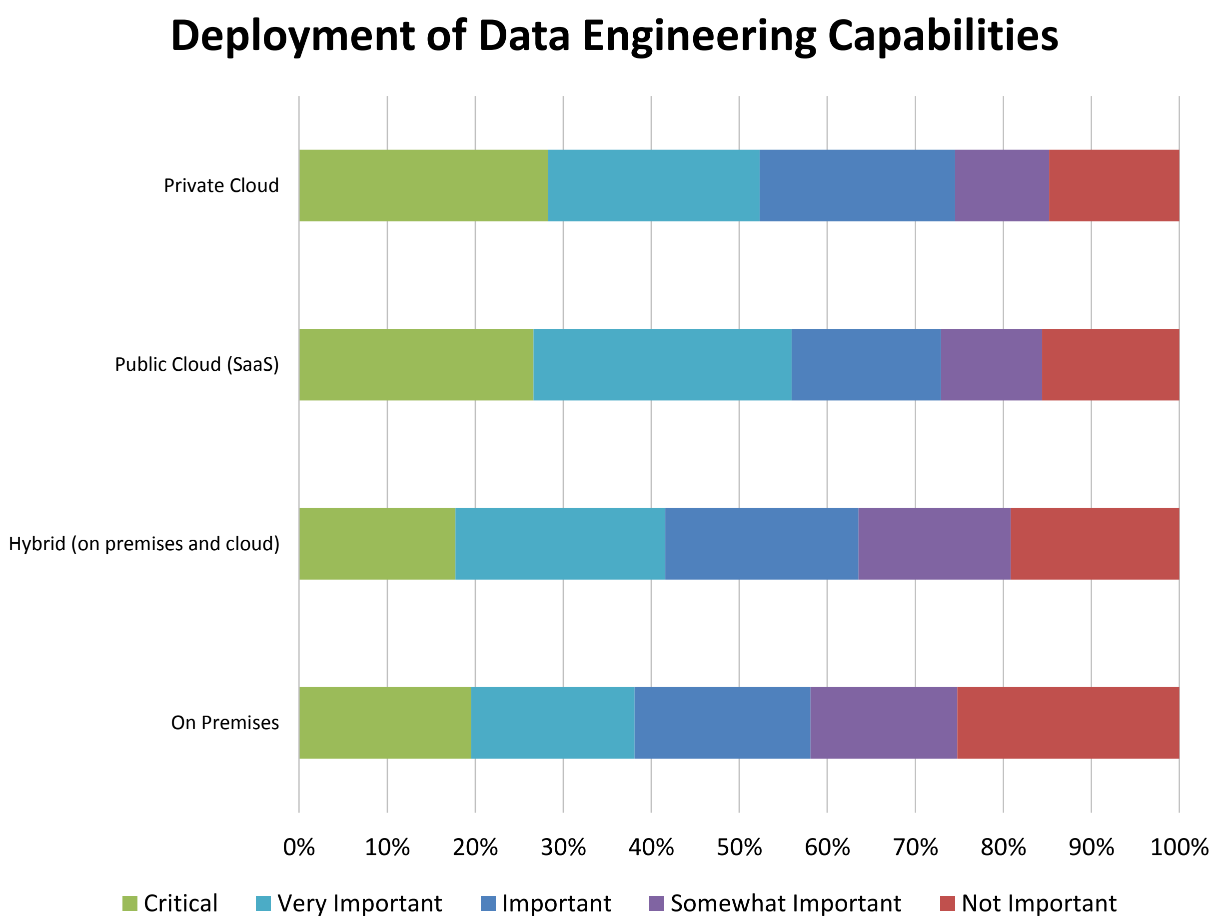


Figure 57– Deployment of data engineering capabilities

The shift in deployment options over the past few years to the cloud for data engineering capabilities is quite clear in fig 58. This shift follows the shift of applications and analytic infrastructure to the cloud over the past few years, albeit with somewhat of a lag compared to applications that were deployed to the cloud sooner. The lake of alignment between the timing of the deployment of capabilities such as applications, analytic data infrastructures, and data engineering to the cloud is a symptom of a larger problem of fragmentation of data across on-premises and cloud deployments.

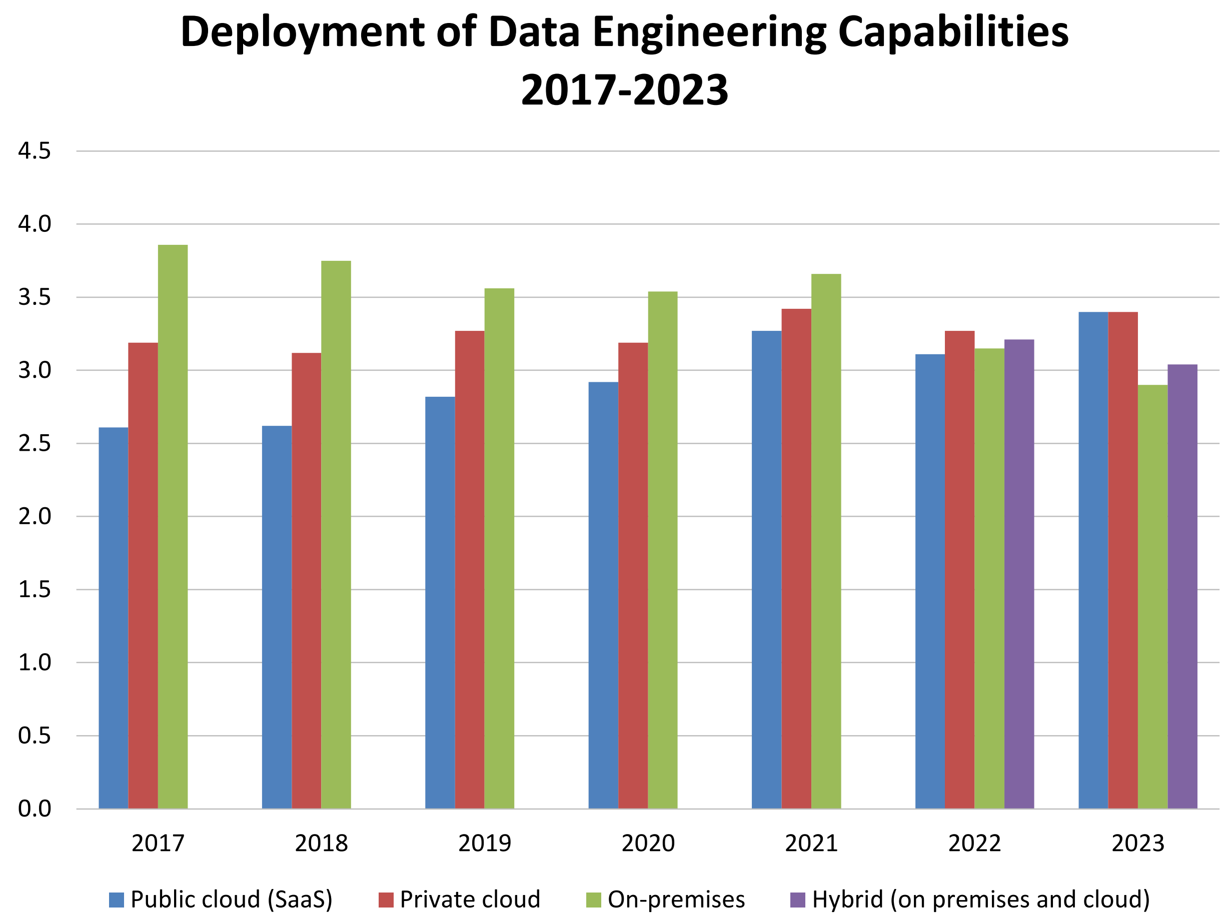


Figure 58 – Deployment of data engineering capabilities 2017-2023

Respondents from *Latin America place* on-premises deployment as their highest priority for deploying data engineering capabilities (fig. 59). *North American* respondents place *on-premises* deployment as their lowest priority. There is variation in the priority of *private cloud or public cloud* across regions as well.

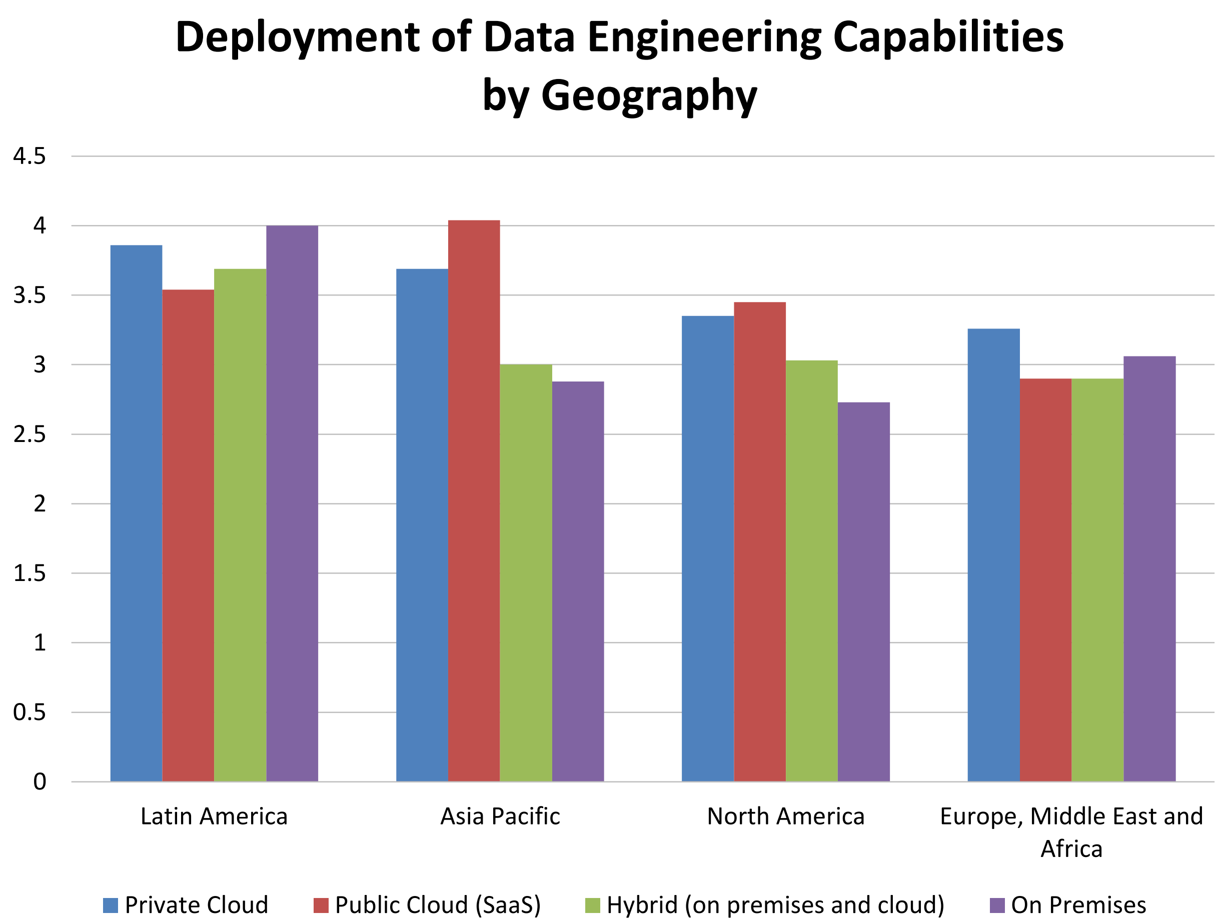


Figure 59 – Deployment of data engineering capabilities by geography

Deployment priorities for data engineering capabilities vary by industry. *Technology* respondents indicate the highest preference for *public cloud* deployments (fig. 60). *Healthcare* respondents indicate the highest priority for *on-premises* deployment. *Financial services* industry respondents indicate the lowest priority for *on-premises* deployments. *Education* respondents place the lowest priority on *public cloud* deployments.

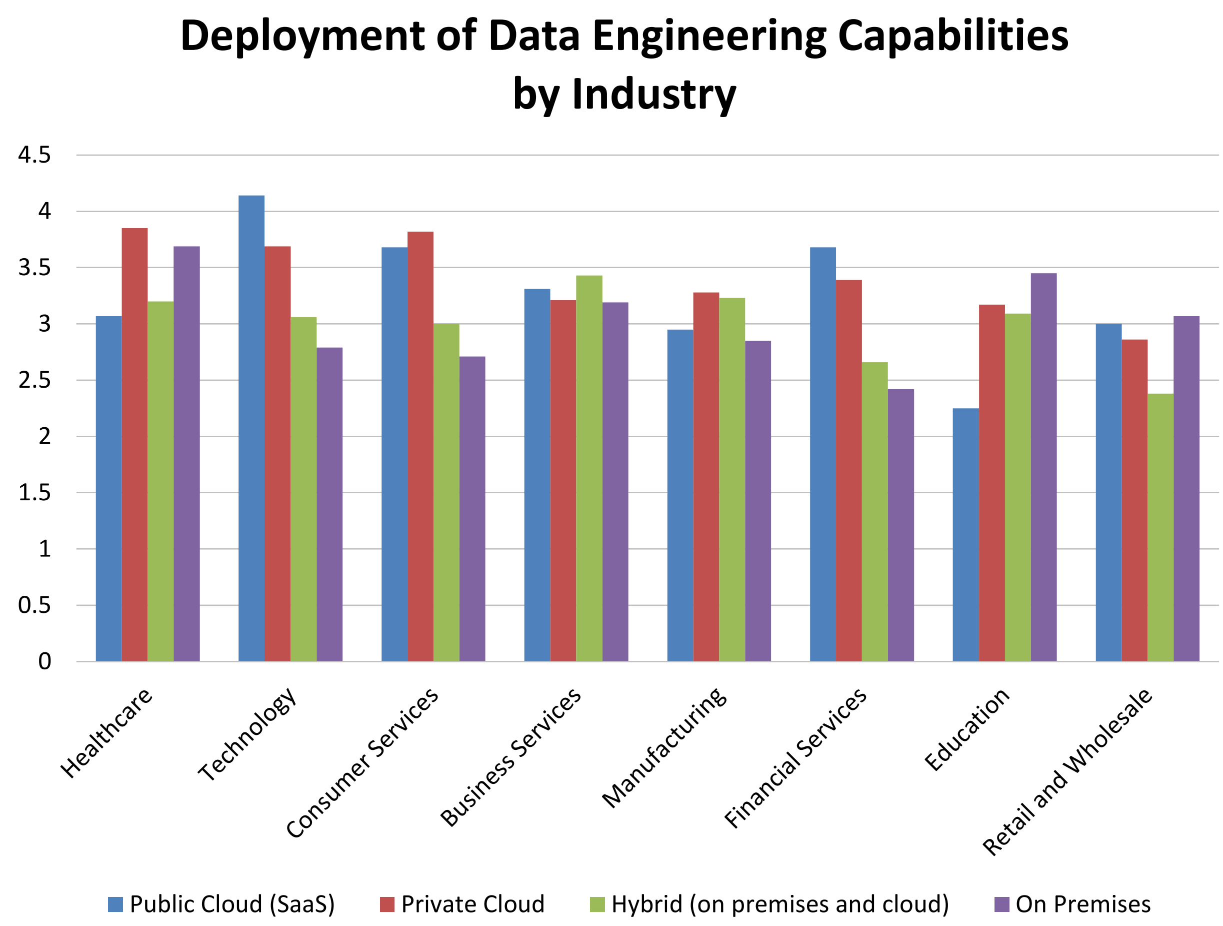


Figure 60 – Deployment of data engineering capabilities by industry

There is variation of data engineering capabilities across organizations for deployment of data engineering capabilities (fig. 61). *Executive management* and *finance* place their lowest priorities for *on-premises* deployments. The *BICC* and *IT* functions place a higher priority on *public cloud* deployments. *Marketing and Sales* prefer *hybrid* deployments. *R&D* functions appear to show no significant preference or priority on deployments.

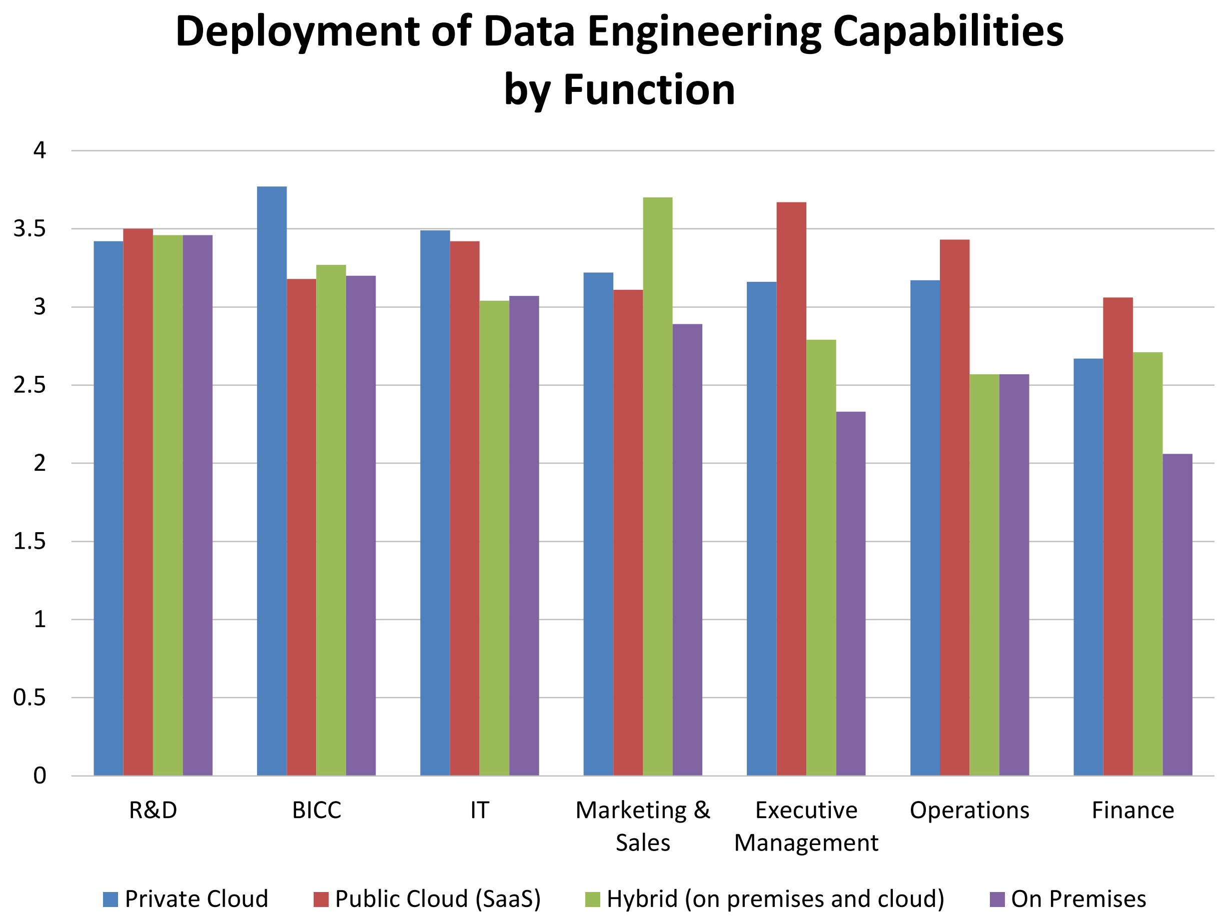


Figure 61 – Deployment of data engineering capabilities by function

Industry and Vendor Analysis

Industry and Vendor Analysis

Data Engineering Industry Importance

We asked industry respondents to indicate the importance of data engineering. Respondents include vendors that do not have a data engineering solution. The greatest majority (65 percent) told us that data engineering is *critically important*, followed by *very important* at 29 percent, with none saying it isn't important. This is a case where the industry is ahead of users, with 39 percent of users ascribing *critical* importance to data engineering (fig. 5).

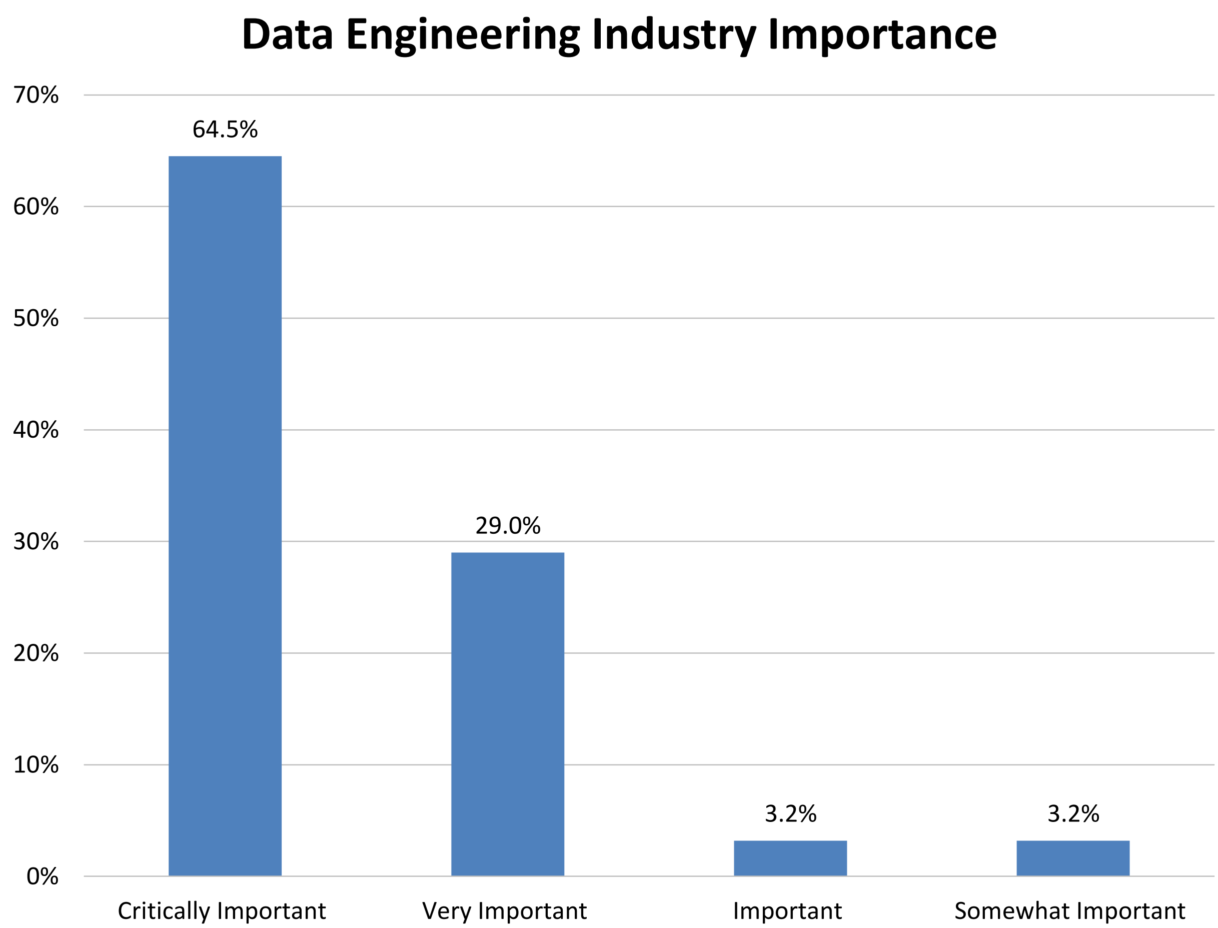


Figure 62 – Data engineering industry importance

Industry Support for Data Engineering Deployment Options

As with users, vendors made the shift to cloud-based deployment, with SaaS/cloud support at 91 percent (fig. 63). Vendors and users differ in priorities beyond that, with users preferring private cloud, followed by hybrid, and then on-premises deployment (see fig. 58).

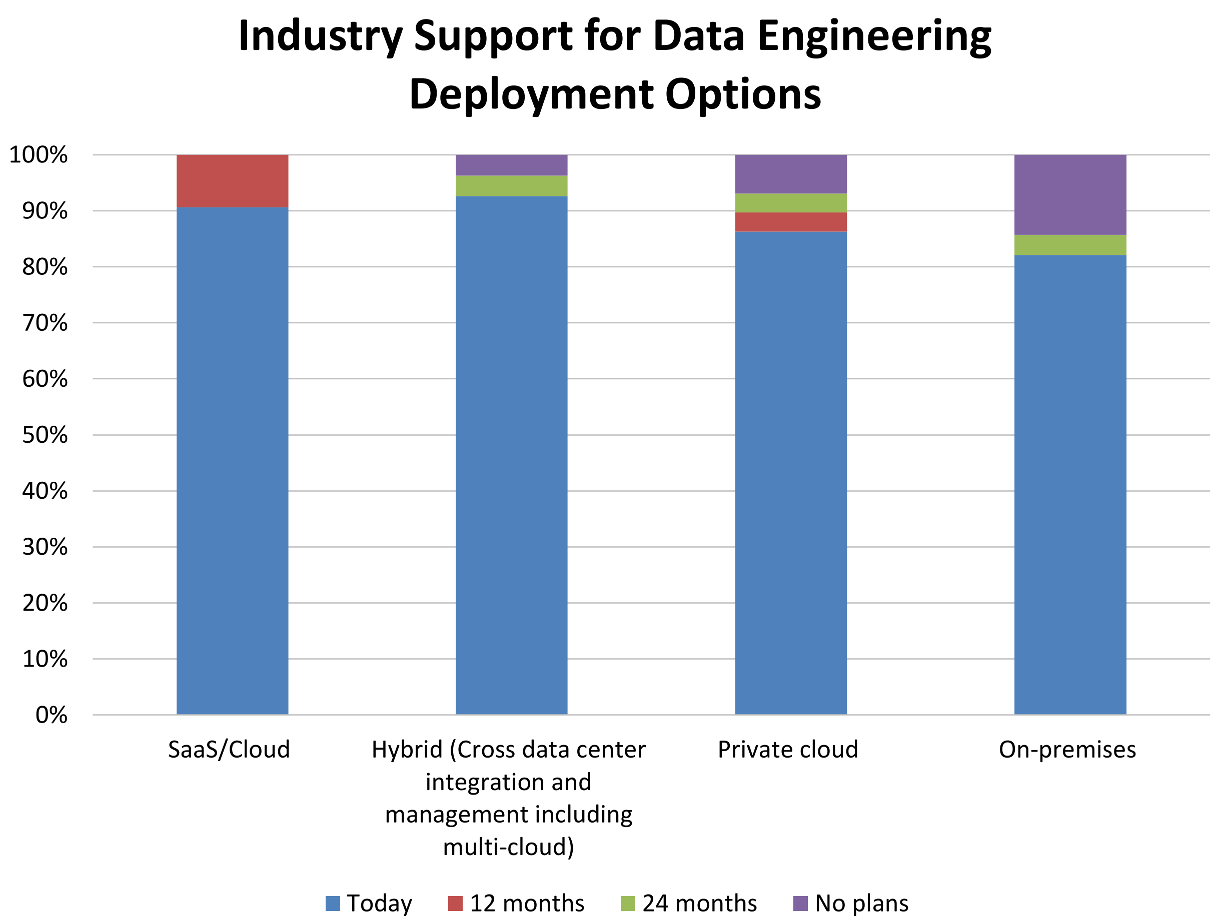


Figure 63 – Industry support for data engineering deployment options

Data Engineering Pricing Models

Unlike traditional pricing models for BI tools, users can expect subscriptions/licensing of data engineering solutions to be based upon a combination of platform utilization and data volumes/capacity (fig. 64). Unlike named user licensing, these sorts of pricing models favor smaller, less active deployments, which can become costly as usage and demand grows over time.

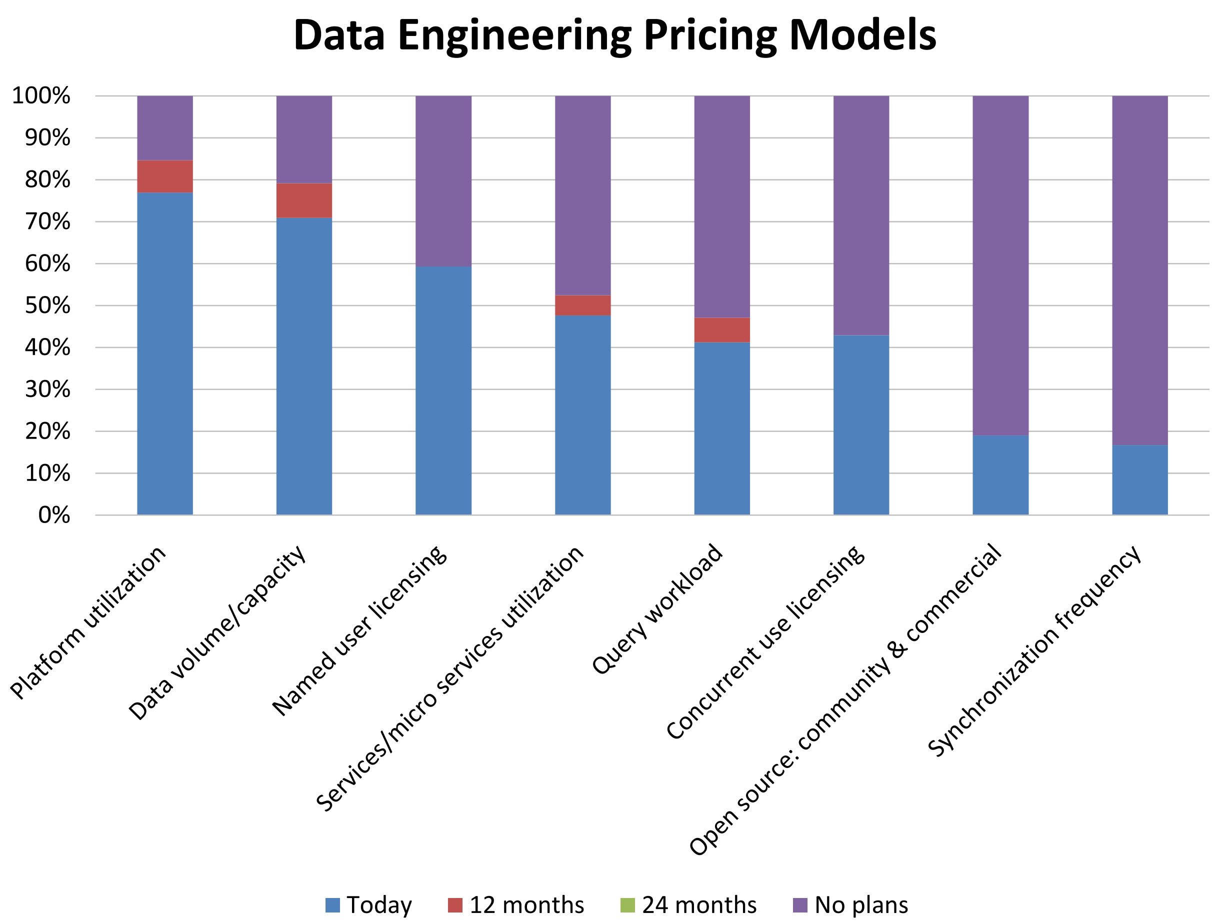


Figure 64 – Data engineering pricing models

Industry Support for Usability Features

In large part, users can expect a majority of vendors to meet their usability requirements for data engineering solutions. That said, user and vendor priorities do not always match (fig. 65). For example, metadata capabilities is in the top half of priorities for end users (see fig. 29), but is at the bottom of the industry list, with 85 percent supporting it.

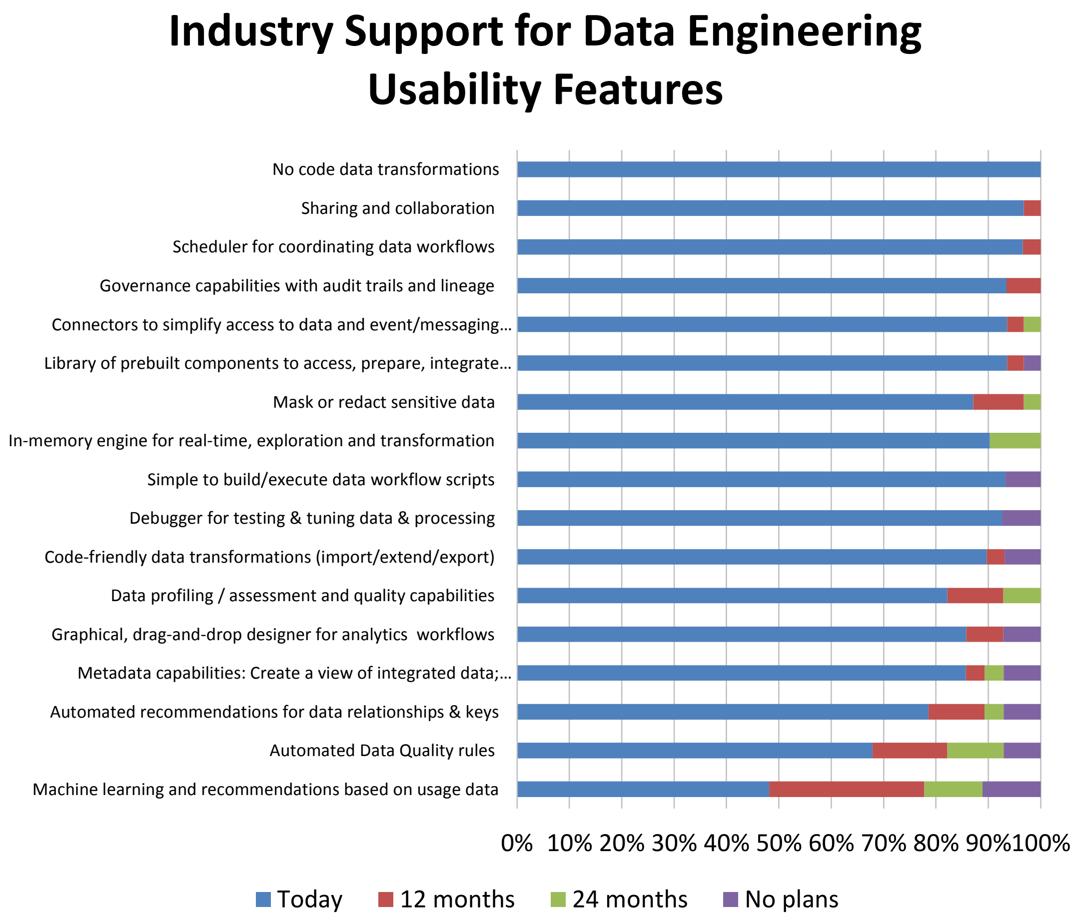


Figure 65 – Industry support for data engineering usability features

Industry Support for Transformation Features

As with usability, users can expect a majority of vendors to meet their transformation requirements for data engineering solutions (fig. 66). However, there is a degree of misalignment between users and suppliers. For example, *ETL workflow* is a top 2 priority for users (see fig. 19) but is only mid-tier for vendors, with 11 percent not supporting it at all.

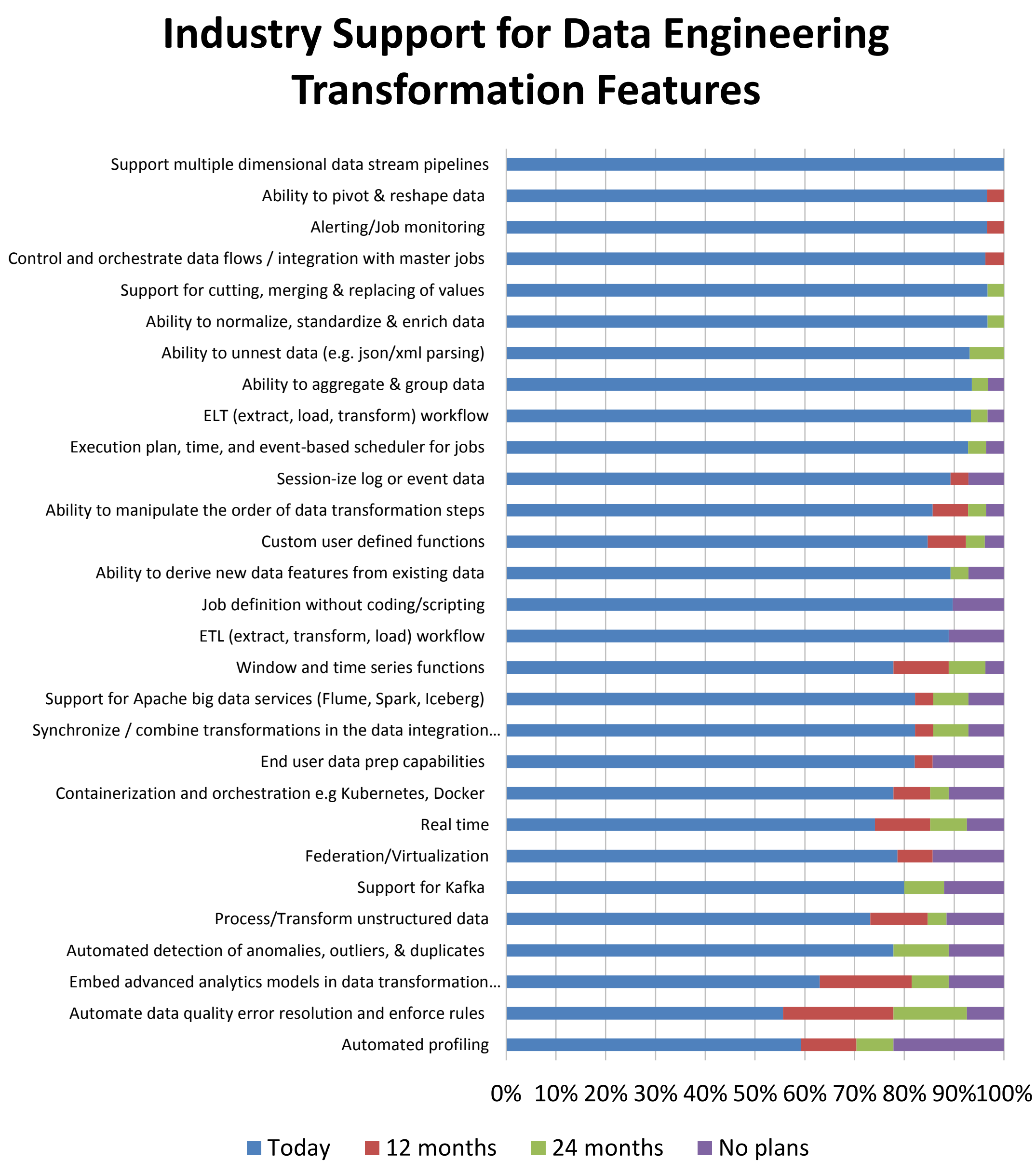


Figure 66 – Industry support for data engineering transformation features

Data Engineering Vendor Ratings

We include 25 vendors in our 2023 data engineering ratings (fig. 67). For each vendor, we consider the following features: development and usability, process and transformation of data, and data sources and targets. In this report, we include only vendors that score 50 percent or greater.

Top vendors include Alteryx (1st), Informatica (1st), Palantir (2nd), RapidMiner (3rd), Domo (4th), and Pyramid Analytics (5th).

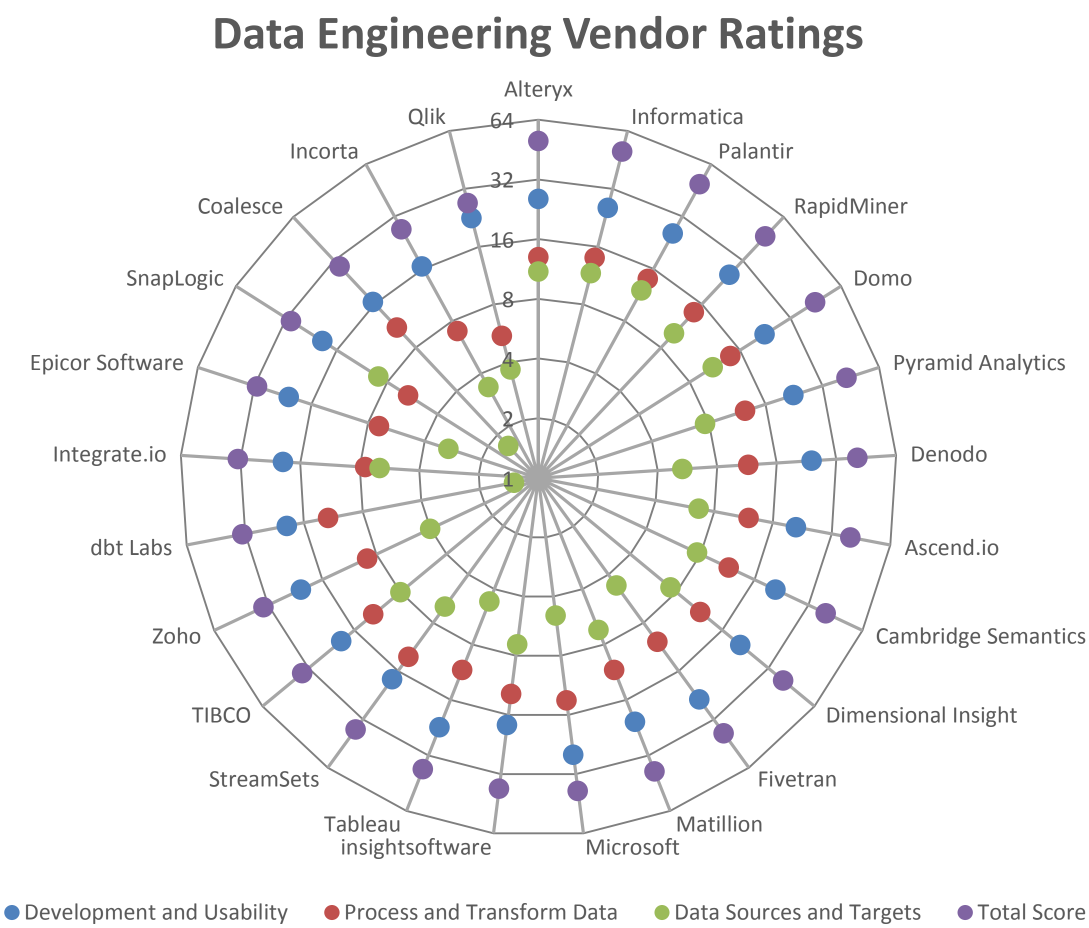


Figure 67 – Data engineering vendor ratings

Other Dresner Advisory Services Research Reports

- Wisdom of Crowds® "Flagship" Business Intelligence Market Study
- Analytical Data Infrastructure
- Analytical Platforms
- BI Competency Center
- Big Data Analytics
- Cloud Computing and Business Intelligence
- Data Catalog
- Data Science and Machine Learning
- Embedded Business Intelligence
- Enterprise Performance Management
- Financial Consolidation, Close Management and Reporting
- Sales Performance Management
- Self-Service BI
- Small and Mid-Sized Enterprise Business Intelligence
- Small and Mid-Sized Enterprise Performance Management

Appendix: 2023 Data Engineering Survey Instrument

Please provide your contact information below:

First Name*: ____

Last Name*: ____

Company: ____

Email Address*: ____

Major Geography*

( ) Asia Pacific

( ) Europe, Middle East and Africa

( ) Latin America

( ) North America

Please specify your city and country

City: ____

Country: ____

Please provide your contact information below:

Address 1: ____

Address 2: ____

City: ____

State: ____

Zip: ____

Country: ____

What is your current title?

____

What function are you a part of?

( ) Business Intelligence Competency Center
( ) Executive Management
( ) Finance
( ) Human Resources
( ) Information Technology (IT)
( ) Marketing
( ) Operations (e.g., Manufacturing, Supply Chain, Services)
( ) Research and Development (R&D)
( ) Sales
( ) Strategic Planning Function
( ) Other - Write In: ____

Please select an industry

( ) Advertising
( ) Aerospace
( ) Agriculture
( ) Apparel & Accessories
( ) Automotive
( ) Aviation
( ) Biotechnology
( ) Broadcasting
( ) Business Services
( ) Chemical
( ) Construction
( ) Consulting
( ) Consumer Products
( ) Defense
( ) Distribution & Logistics

( ) Education (Higher Ed)

( ) Education (K-12)

( ) Energy

( ) Entertainment and Leisure

( ) Executive Search

( ) Federal Government

( ) Financial Services

( ) Food, Beverage and Tobacco

( ) Healthcare

( ) Hospitality

( ) Insurance

( ) Legal

( ) Manufacturing

( ) Mining

( ) Motion Picture and Video

( ) Not for Profit

( ) Pharmaceuticals

( ) Publishing

( ) Real Estate

( ) Retail and Wholesale

( ) Sports

( ) State and Local Government

( ) Technology

( ) Telecommunications

( ) Transportation

( ) Utilities

( ) Other - Write In: ____

How many employees does your company employ worldwide?

( ) 1-100

( ) 101-1,000

( ) 1,001-2,000

( ) 2,001-5,000

( ) 5,001-10,000

( ) More than 10,000

How important is data engineering to your organization?*

( ) Critical

( ) Very Important

( ) Important

( ) Somewhat Important

( ) Not Important

What are your plans for employing data engineering?

( ) Using today

( ) Using today with plans to expand

( ) Plan to use in 12 months

( ) Plan to use in 24 months

( ) No plans

44) Vendors and Products

Which vendor/product are you using for data engineering?:

____

How satisfied are you with your vendor and product for data engineering?

( ) Extremely satisfied

( ) Mostly satisfied

( ) Somewhat satisfied

( ) Somewhat unsatisfied

( ) Unsatisfied

How effective is the current approach to data engineering today?

( ) Highly Effective

( ) Somewhat Effective

( ) Somewhat Ineffective

( ) Totally Ineffective

How frequently do people in your organization engage in data engineering prior to analysis?

( ) Constantly

( ) Frequently

( ) Occasionally

( ) Rarely

( ) Never

How often do users enrich internal data with third-party data (e.g. Google, Facebook, Dun & Bradstreet, US Census)?

( ) Constantly

( ) Frequently

( ) Occasionally

( ) Rarely

( ) Never

How to you define, generate, and extract features for machine learning models?

	Primary	Secondary	Tertiary	N/A

Custom SQL, Python, Java Script	( )	( )	( )	( )
End user data prep tool	( )	( )	( )	( )
ETL tool	( )	( )	( )	( )
Data prep capabilities of Analytic Data Infrastructure platform	( )	( )	( )	( )

Please indicate the estimated percent of your BI use cases for which your data engineering products are purchased and being used;

	Under 10%	**11-20%**	**21-40%**	**41-60%**	**More than 60%**
As part of the data integration, cleansing and transformation workflows for a data warehouse supporting dashboards and reporting	( )	( )	( )	( )	( )
Data integration and transformation services for ad-hoc query, discovery and exploration analysis	( )	( )	( )	( )	( )

Data science and augmented analytics use cases, predictive and prescriptive analytics	( )	( )	( )	( )	( )
Re-platforming/replication of an existing data warehouse	( )	( )	( )	( )	( )

Should data engineering be a standalone capability or part of another tool?

( ) Standalone

( ) Part of business intelligence tools

( ) Other - Write In: ____

Please indicate the importance of the following development and usability features for data engineering software:

	Critical	**Very Important**	**Important**	**Somewhat Important**	**Not Important**
Simple to build/execute data workflow scripts	( )	( )	( )	( )	( )
Graphical, drag-and-drop designer for creation of analytics data integration workflows	( )	( )	( )	( )	( )
In-memory engine supporting real-time, interactive	( )	( )	( )	( )	( )

exploration and transformation development					
Automated recommendations for data relationships & keys for combining data across multiple data sets and sources	( )	( )	( )	( )	( )
Machine learning and recommendations based on usage data gathered across users, groups, or organizations	( )	( )	( )	( )	( )
Library of prebuilt components to access, prepare and integrate data from sources	( )	( )	( )	( )	( )
Scheduler for coordinating data workflows	( )	( )	( )	( )	( )
Debugger for testing and tuning data and processing at any stage of the integration workflow	( )	( )	( )	( )	( )
Metadata capabilities: Create a view of integrated data (e.g. from customers, products, suppliers)	( )	( )	( )	( )	( )

and securely manage and deliver meta data for analytical use cases/tools					
Connectors to simplify access to data and event/messaging sources, social media, enterprise cloud applications, security services	( )	( )	( )	( )	( )
Mask or redact sensitive data	( )	( )	( )	( )	( )
Data profiling / assessment and quality capabilities	( )	( )	( )	( )	( )
Automated Data Quality rules	( )	( )	( )	( )	( )
Sharing and collaboration	( )	( )	( )	( )	( )
No code data transformations (this might be covered under graphical, drag-and-drop capabilities)	( )	( )	( )	( )	( )
Code-friendly data transformations (import/extend/export)	( )	( )	( )	( )	( )
Governance capabilities with audit trails and	( )	( )	( )	( )	( )

lineage					

Please indicate the importance of the following data engineering features for processing and transforming data

	Critical	**Very Important**	**Important**	**Somewhat Important**	**Not Important**
ETL (extract, transform, load) workflow	( )	( )	( )	( )	( )
ELT (extract, load, transform) workflow (use target platform for transformation processing)	( )	( )	( )	( )	( )
Process/Transform unstructured data (Text/Office documents, PDF files etc.)	( )	( )	( )	( )	( )
Federation/Virtualization (Ad hoc integration of data; i.e. to support Ad hoc analytics in the case of discovery and exploration and data sciences/modeling using machine learning and advanced analytic techniques such as prescriptive and predictive analytics)	( )	( )	( )	( )	( )

Capability to synchronize and combine transformations in the data integration workflow, including notifications and alerts (e.g. Business Activity Monitoring)	( )	( )	( )	( )	( )
Control and orchestrate data flows and data integration with master jobs	( )	( )	( )	( )	( )
Execution plan, time, and event-based scheduler for jobs	( )	( )	( )	( )	( )
Automate data quality error resolution and enforce rules	( )	( )	( )	( )	( )
Automated detection of anomalies, outliers, & duplicates	( )	( )	( )	( )	( )
Real time/streaming, trickle, increments/change capture	( )	( )	( )	( )	( )
Job definition without coding/scripting for map, aggregate, sort, enrich, and merge data operations	( )	( )	( )	( )	( )
Containerization and orchestration e.g. Kubernetes, Docker	( )	( )	( )	( )	( )
Support for Kafka	( )	( )	( )	( )	( )

Support for Apache big data services (Flume, Spark)	( )	( )	( )	( )	( )
Embed advanced analytics models in data transformation workflow (e.g. integrate models from R, Python, Scala and Weka to operationalize analytic models within data workflow.	( )	( )	( )	( )	( )
Alerting/Job monitoring	( )	( )	( )	( )	( )
Ability to unnest data (e.g. json/xml parsing)	( )	( )	( )	( )	( )
Ability to normalize, standardize & enrich data	( )	( )	( )	( )	( )
Support for cutting, merging & replacing of values	( )	( )	( )	( )	( )
Ability to aggregate & group data	( )	( )	( )	( )	( )
Ability to pivot (convert table to matrix) & reshape (convert matrix to table) data	( )	( )	( )	( )	( )
Ability to derive new data features from existing data (text extraction, math expressions, date	( )	( )	( )	( )	( )

expressions, etc.)					
Ability to manipulate the order of data transformation steps	( )	( )	( )	( )	( )
Session-ize log or event data	( )	( )	( )	( )	( )
Window and time series functions	( )	( )	( )	( )	( )
Custom user defined functions	( )	( )	( )	( )	( )
End user data prep capabilities	( )	( )	( )	( )	( )
Automated profiling for integrity constraints, formatting patterns, and dependencies between columns and datasets.	( )	( )	( )	( )	( )

Which of the following advanced transformations are important to you?

	Critical	Very Important	Important	Somewhat Important	Not Important	Don't Know
Multi-dimensional/OLAP	( )	( )	( )	( )	( )	( )
Path/Link Analysis	( )	( )	( )	( )	( )	( )
Pattern Matching	( )	( )	( )	( )	( )	( )
Aggregations	( )	( )	( )	( )	( )	( )
Statistical Analysis, R	( )	( )	( )	( )	( )	( )
Spark	( )	( )	( )	( )	( )	( )
Graph	( )	( )	( )	( )	( )	( )
Search Analytics	( )	( )	( )	( )	( )	( )
Text Analysis	( )	( )	( )	( )	( )	( )
Ranking/Scoring	( )	( )	( )	( )	( )	( )
User-defined Functions	( )	( )	( )	( )	( )	( )
Machine Learning	( )	( )	( )	( )	( )	( )
Sentiment Analysis	( )	( )	( )	( )	( )	( )

What are the top data sources, targets and interfaces for use as a part of data engineering?

	Source	**Target**	**Native interface (e.g., bulk loader)**	**Third party ODBC; JDBC**
Relational (Oracle, SQL Server)	[ ]	[ ]	[ ]	[ ]
Graph (e.g., Neo4J, Tiger graph)	[ ]	[ ]	[ ]	[ ]
Object stores (e.g., Amazon S3)	[ ]	[ ]	[ ]	[ ]
Files (e.g., Excel, CSV, Log files, JSON)	[ ]	[ ]	[ ]	[ ]
Analytical databases (e.g., Snowflake, Exasol, Incorta, AtScale)	[ ]	[ ]	[ ]	[ ]
NoSQL (e.g., Mongo, Couchbase)	[ ]	[ ]	[ ]	[ ]

Hadoop-based (e.g., Cloudera)	[ ]	[ ]	[ ]	[ ]
Specialty data platforms (e.g., SAP Hana, Palantir)	[ ]	[ ]	[ ]	[ ]
Applications (e.g., Salesforce, Workday, Oracle, SAP, Infor)	[ ]	[ ]	[ ]	[ ]

Where should data engineering functionality reside?

	Critical	**Very Important**	**Important**	**Somewhat Important**	**Not Important**
On Premises	( )	( )	( )	( )	( )
Private Cloud	( )	( )	( )	( )	( )
Public Cloud (SaaS)	( )	( )	( )	( )	( )
Hybrid (on premises and cloud)	( )	( )	( )	( )	( )