February 28, 2019

Dresner Advisory Services, LLC

2019 Edition

Data Preparation

Wisdom of Crowds® Series

Licensed to Trifacta

Disclaimer:

This report is for informational purposes only. You should make vendor and product selections based on multiple information sources, face-to-face meetings, customer reference checking, product demonstrations, and proof-of-concept applications.

The information contained in this Wisdom of Crowds® market study report is a summary of the opinions expressed in the online responses of individuals that 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 this report, the study results, or for any damages incurred or alleged to be incurred by any of the companies included in the report as a result of the report's 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.

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

Data Preparation Defined

Data Preparation is a capability for a variety of users—both business and IT—to model, prepare, and combine data prior to analysis.

Contents

Introduction

In 2019, we celebrate the 12th anniversary of Dresner Advisory Services! Our thanks to all of you for your continued support and ongoing encouragement. Since our founding in 2007, we have worked hard to set the “bar” high—challenging ourselves to innovate and lead the market—offering ever greater value with each successive year.

We are also pleased that our second annual conference, Real Business Intelligence, held June 27-28, 2018 on the MIT campus in Cambridge, Massachusetts, was a great success! Unlike other events, we designed Real Business Intelligence as an immersive thought leadership event focused on strategies for success with information management, business intelligence, analytics, and performance management. Our 2019 event will return to the campus of MIT and is scheduled for May 14-15.

We renamed our fifth edition of this report from “End User Data Preparation” to simply “Data Preparation.” This reflects a shift within the data preparation market to include business professionals—as well as IT—as key users. Data preparation is a topic that resonates strongly with organizations—and especially with power users and analysts that, historically, have been relegated to using whatever tools were available for the purpose—regardless of limitations.

An important step towards the ongoing trend of user empowerment and self-service business intelligence, data preparation is driving an increasing amount of investment on both demand and supply sides of the equation.

We hope you enjoy this report!

Best,

Chief Research Officer
Dresner Advisory Services

About Howard Dresner and Dresner Advisory Services

The Dresner Advisory Services Data preparation 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" Business Intelligence Market study
- Advanced and Predictive Analytics
- Analytical Data Infrastructure
- Business Intelligence Competency Center
- Cloud Computing and Business Intelligence
- Collective Insights®
- Embedded Business Intelligence
- IoT Intelligence®
- Location Intelligence

Howard conducts a weekly Twitter "tweetchat" on Fridays at 1:00 p.m. ET. During these live events the #BIWisdom "tribe" discusses a wide range of business intelligence topics.

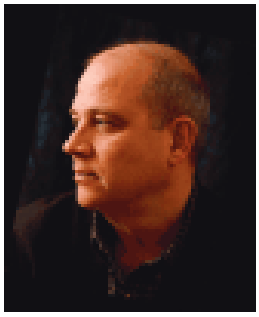
You can find more information about Dresner Advisory Services at www.dresneradvisory.com.

About Jim Ericson

Jim Ericson is a research director with Dresner Advisory Services.

Jim has served as a consultant and journalist who studies end-user management practices and industry trending in the data and information management fields.

From 2004 to 2013 he was the editorial director at *Information Management* magazine (formerly *DM Review*), where he created architectures for user and industry coverage for hundreds of contributors across the breadth of the data and information management industry.



As lead writer he interviewed and profiled more than 100 CIOs, CTOs, and program directors in a 2010-2012 program called “25 Top Information Managers.” His related feature articles earned ASBPE national bronze and multiple Mid-Atlantic region gold and silver awards for Technical Article and for Case History feature writing.

A panelist, interviewer, blogger, community liaison, conference co-chair, and speaker in the data-management community, he also sponsored and co-hosted a weekly podcast in continuous production for more than five years.

Jim’s earlier background as senior morning news producer at NBC/Mutual Radio Networks and as managing editor of MSNBC’s first Washington, D.C. online news bureau cemented his understanding of fact-finding, topical reporting, and serving broad audiences.

Findings and Analysis

In this report, we present the deliverables for our Data Preparation Market Study based upon data collection from July through October 2018.

Focus of Research

In this study, we address key data preparation issues including:

- Perceptions and intentions surrounding data preparation
- End-user requirements and features:
    - Usability features
    - Integration features
    - Manipulation features
    - Output options
    - Deployment options
- Industry support for data preparation
- User requirements versus industry capabilities
- Vendor ratings

Benefits of the Study

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

Consumer Guide

As an objective source of industry research, consumers use the Dresner Advisory Services Data Preparation Market Study to understand how their peers leverage and invest in data preparation 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 Preparation Market Study in several important ways:

External Awareness

- Build awareness for business intelligence markets and supplier brands, citing the Dresner Advisory Services Data Preparation Market Study trends and vendor performance
- Gain lead and demand generation for supplier offerings through association with the Dresner Advisory Services Data Preparation 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 Preparation Market Study
- Better understand customer priorities, concerns, and issues
- Identify competitive pressures and opportunities

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.

Data Quality

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

Executive Summary

- Data preparation ranks 14th, in the top half of 33 BI technologies and Initiatives under our study (p. 18).
- Sixty-three percent of all respondents say data preparation is either “critical” or “very important.” Importance declines year over year, possibly signaling a watershed of user sentiment. Industry sentiment is very high and may have peaked, but remains ahead of user sentiment (p. 79).
- Seventy-eight percent say their current data preparation approach is highly or very effective, and confidence grew over time. Success with data preparation correlates to success with BI (pp. 25-31).
- Sixty-six percent "constantly" or "frequently" make use of data preparation. R&D respondents are the most common users (pp. 32-37).
- Fifty-nine percent of respondents "constantly," "frequently," or "occasionally" enrich data preparation with third-party data. Marketing/Sales are the most frequent users (pp. 38-43).
- Sixty percent or more say a broad range of data preparation usability features are all at least "somewhat important" (pp. 44-49). Industry support is robust and aligned with user needs (p. 81).
- The top two data preparation integration features are "access to file formats" and "ability to combine data across multiple data sets and sources" (pp. 50-55). Industry support answers all current user expectations (p. 82).
- The top two “critical” data preparation manipulation features are "ability to aggregate and group data" and "ability to pivot data" (pp. 56-61). Industry support in 2019 is well ahead of user needs (p. 83).
- The most important data preparation output is to Excel and CSV, followed by output to traditional relational databases. Azure gained the most user interest in 2019 (pp. 62-67). Industry support is strong and addresses future user needs (p. 83).
- The two most popular data preparation deployment features are "schedule a process to run on a time-based or trigger-based event" and "ability to schedule execution/replay of data transformation processing " (pp. 68-73). Industry support is strong, with some future investment planned (p. 84).
- Respondents prefer on-premises deployment of data preparation capabilities to private or public cloud deployment (pp. 74-78). For the first time, industry respondents currently support more cloud-based than on-premises products. (pp. 85-86).
- Data preparation vendor ratings are shown on p. 87.

Study Demographics

Our sample includes a cross-section of data across geographies, functions, organization sizes, and vertical industries. We believe that, unlike other industry research, we offer a more characteristic sample and better indicator of true market dynamics.

Geography

Survey respondents represent a mix of global geographies. Sixty-seven percent represent North America (including five Canadian provinces and the majority of U.S. states). Twenty-five percent work in EMEA; the remainder represents Asia Pacific and Latin America (fig. 1).

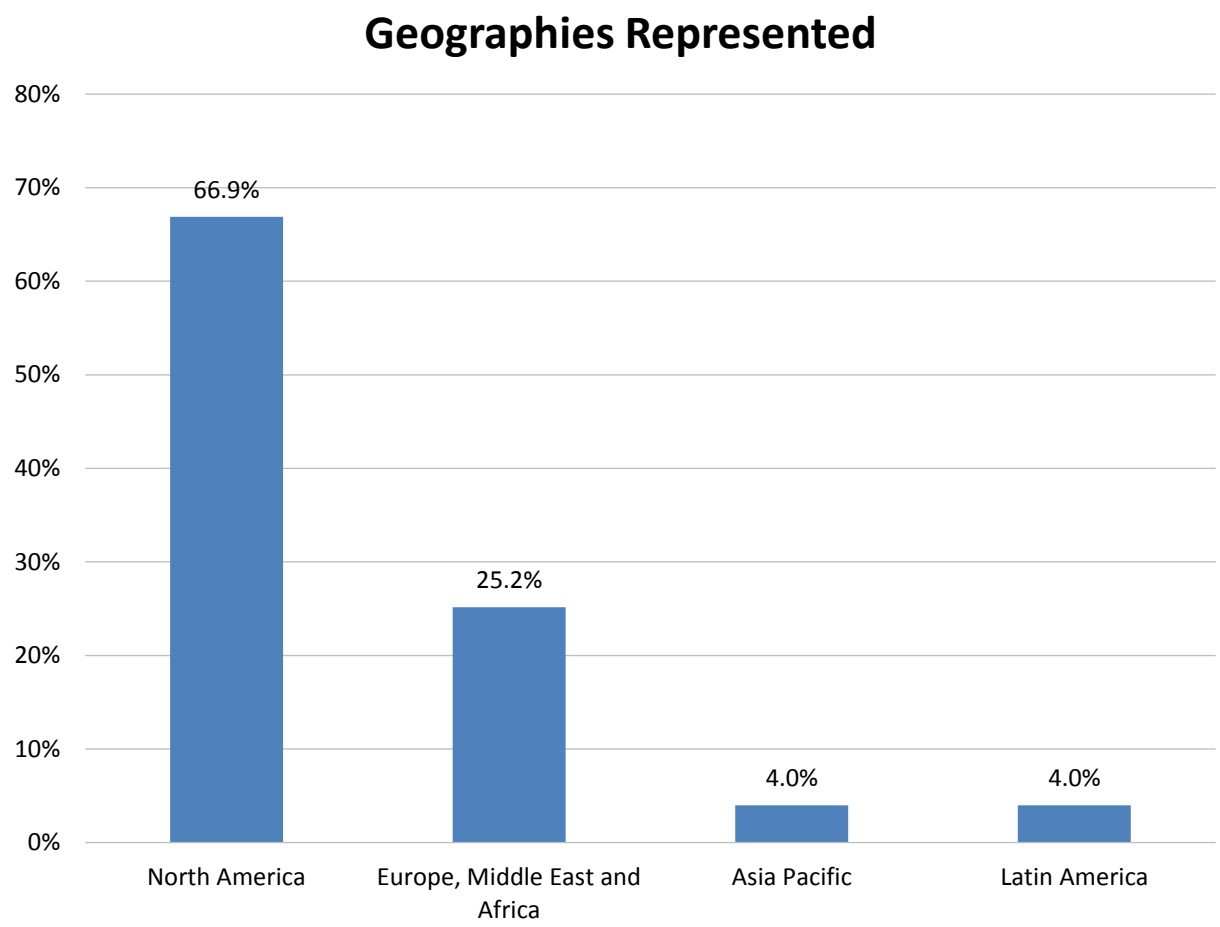


Figure 1 – Geographies represented

Functions

Information Technology accounts for the largest group of respondents (35 percent) by function. About 19 percent come from the Business Intelligence Competency Center (BICC). Executive Management and Finance are the next most represented (fig. 2).

Tabulating results by function enables us to compare and contrast the plans and priorities of different departments within organizations.

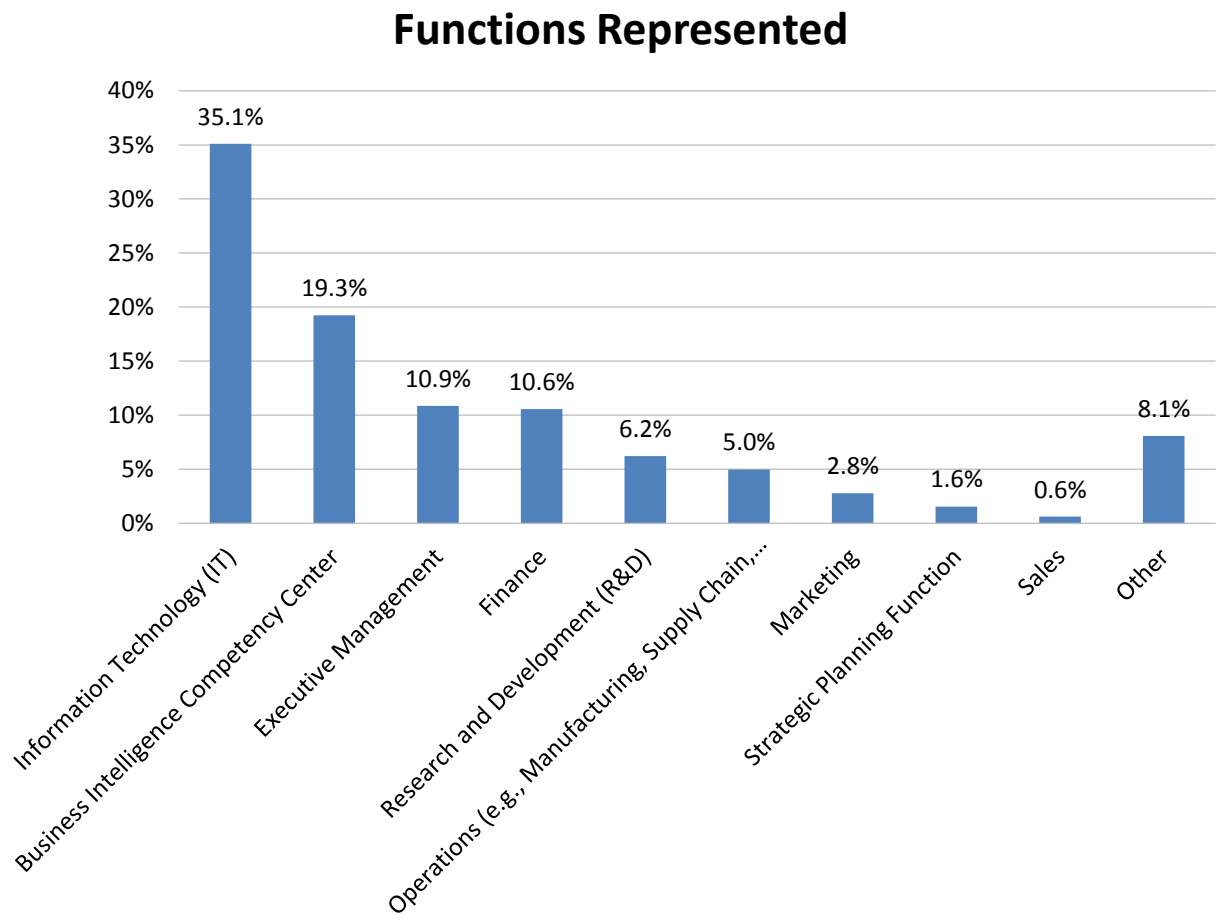


Figure 2 – Functions represented

Vertical Industries

Survey participants represent a wide range of vertical industries, led by Financial Services (12 percent), Technology (12 percent), and Healthcare (10 percent) (fig.3). Consulting, Higher Education, and Retail/Wholesale are the next most represented. We allow and encourage the participation of consultants, who often have deeper industry knowledge than their customer counterparts. Third-party relationships give us insight into the partner ecosystem for BI vendors.

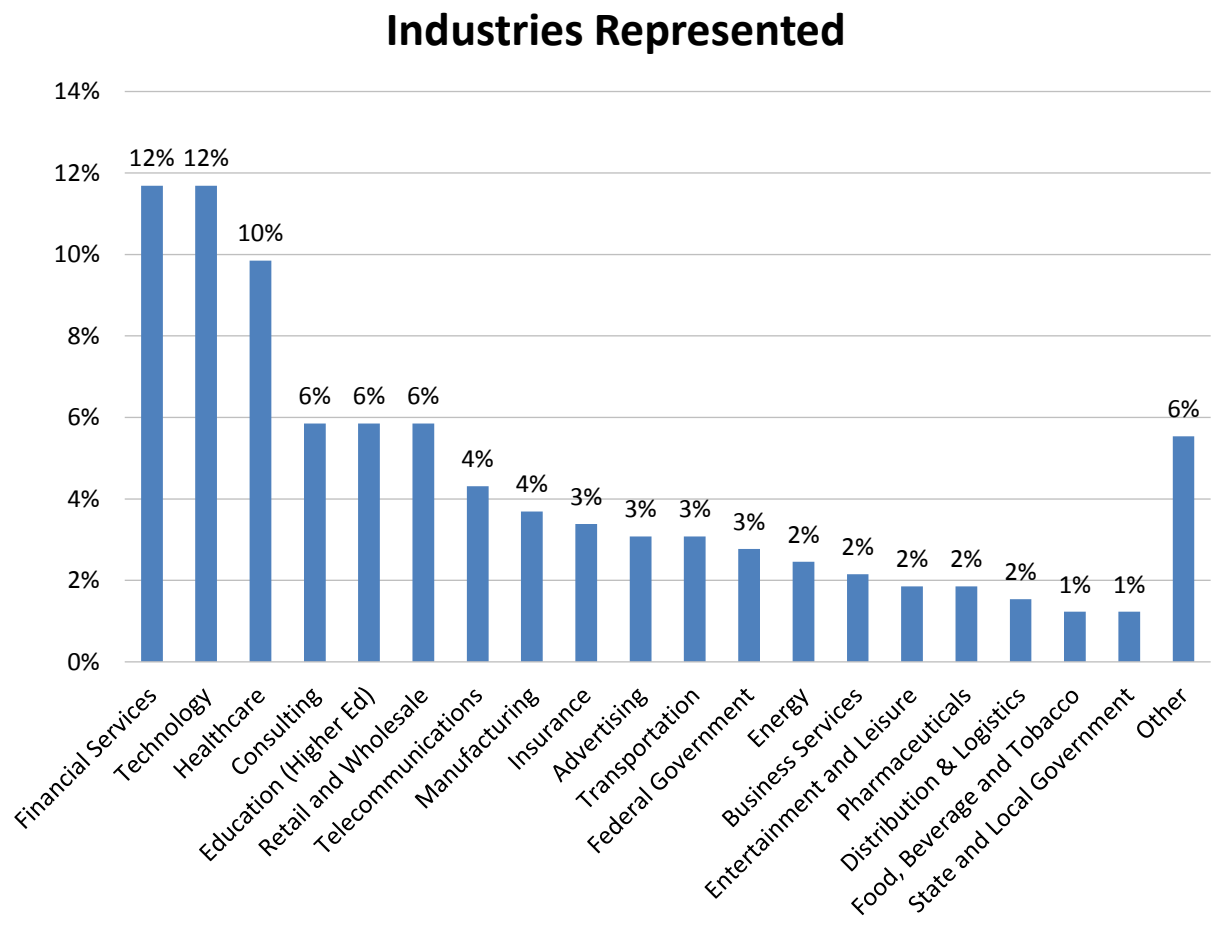


Figure 3 – Industries represented

Organization Size

Our survey sample includes a mix of small, medium, and large organizations (fig. 4). In 2019, small organizations (1-100 employees) account for about 25 percent of the sample, and mid-sized organizations (101-1,001 employees) account for 28 percent of the sample. Large organizations (>1,000 employees) account for the remaining respondents, with very large organizations (>5,000 employees) accounting for 22 percent.

Segmenting respondents by organization size helps us identify differences in behavior, attitudes, and planning often related to headcount.

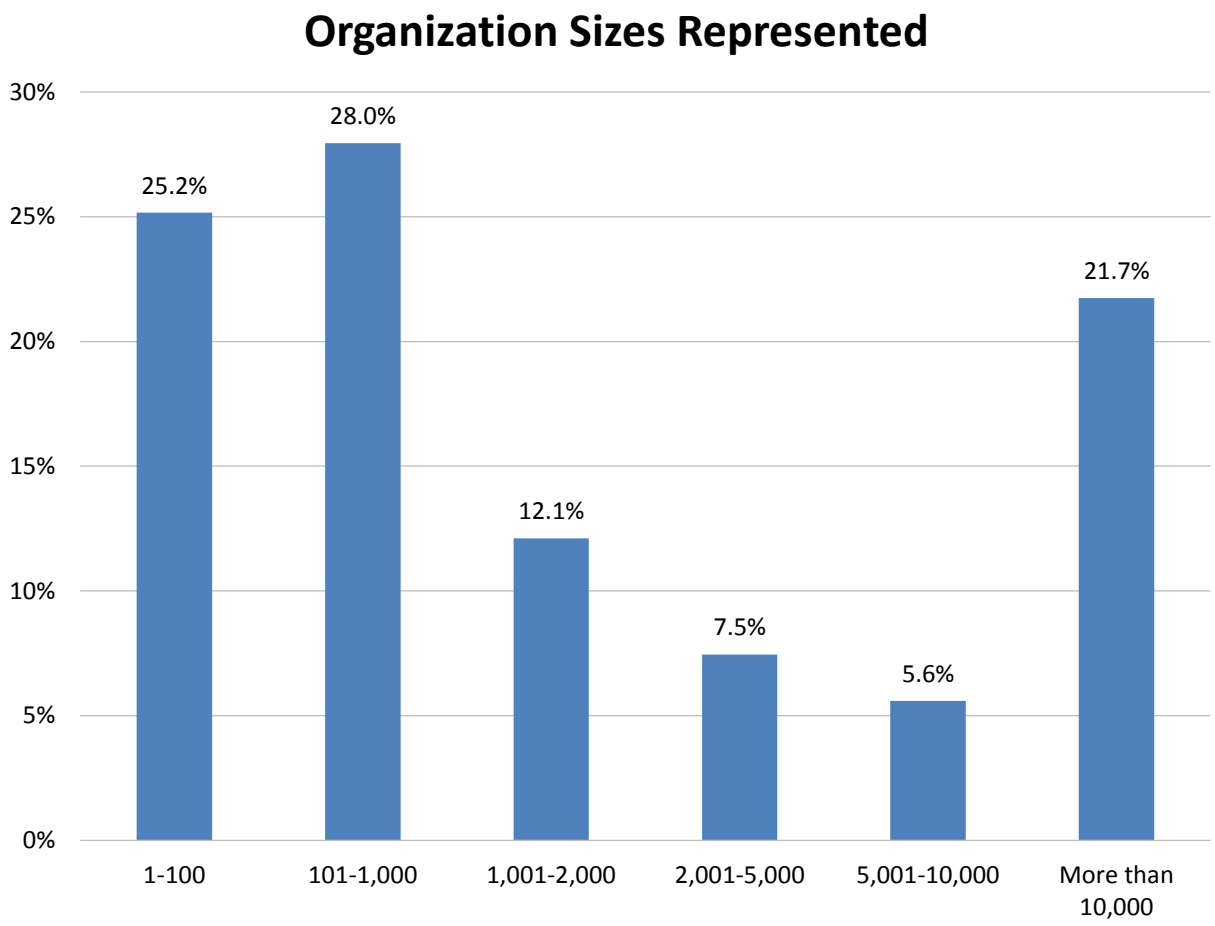


Figure 4 – Organization sizes represented

Analysis of Findings

In 2019, our fifth annual Data Preparation Market Study examines the nature of data preparation, exploring user sentiment and perceptions, the nature of current implementations, and plans for the future.

Importance of Data Preparation

Among technologies and initiatives strategic to business intelligence in 2019, data preparation (aka blending) ranks 14th, in the top half of 33 topics we currently study and in the same relative position as in our 2018 study (fig. 5). Thus, data preparation importance trails traditional topics including reporting, dashboards, advanced visualization, and end-user self-service. But it also ranks well ahead of other familiar topics including cloud computing, big data, and the Internet of Things. We believe the relative strategic importance users attach to data preparation underscores the value attached to user empowerment and self-service generally.

Figure 5 – Technologies and initiatives strategic to business intelligence

In this study of data preparation, we find that respondents' perceived importance of data preparation is very high and in line with high user demand for self-service business intelligence and user autonomy (fig. 6). Sixty-three percent of all respondents say data preparation is either "critical" or "very important." About 87 percent of respondents say data preparation is, at minimum, "important." Just 1 percent says data preparation is "not important."

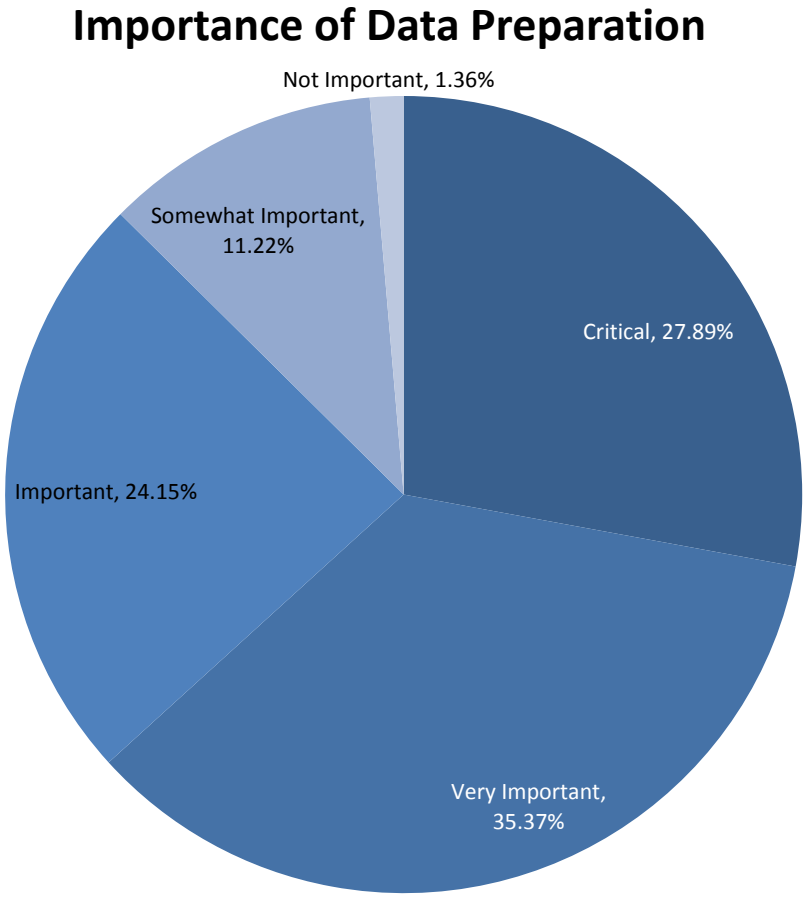


Figure 6 – Importance of data preparation

Across five years of data, respondents' perceived importance of data preparation remains consistently important with weighted mean values between 3.8 and 4.0, or near "very important (fig. 7). Year-over-year weighted mean sentiment decline however, from a 2018 high of 3.97 to 3.77 in 2019. Most noticeably, estimations of "critical" importance declined to 28 percent compared to 39 percent in the previous year, perhaps signaling a watershed in perceived importance. Still, 88 percent of respondents say data preparation is, at minimum, "important" in 2019, very close to the 89 percent who said the same in 2018.

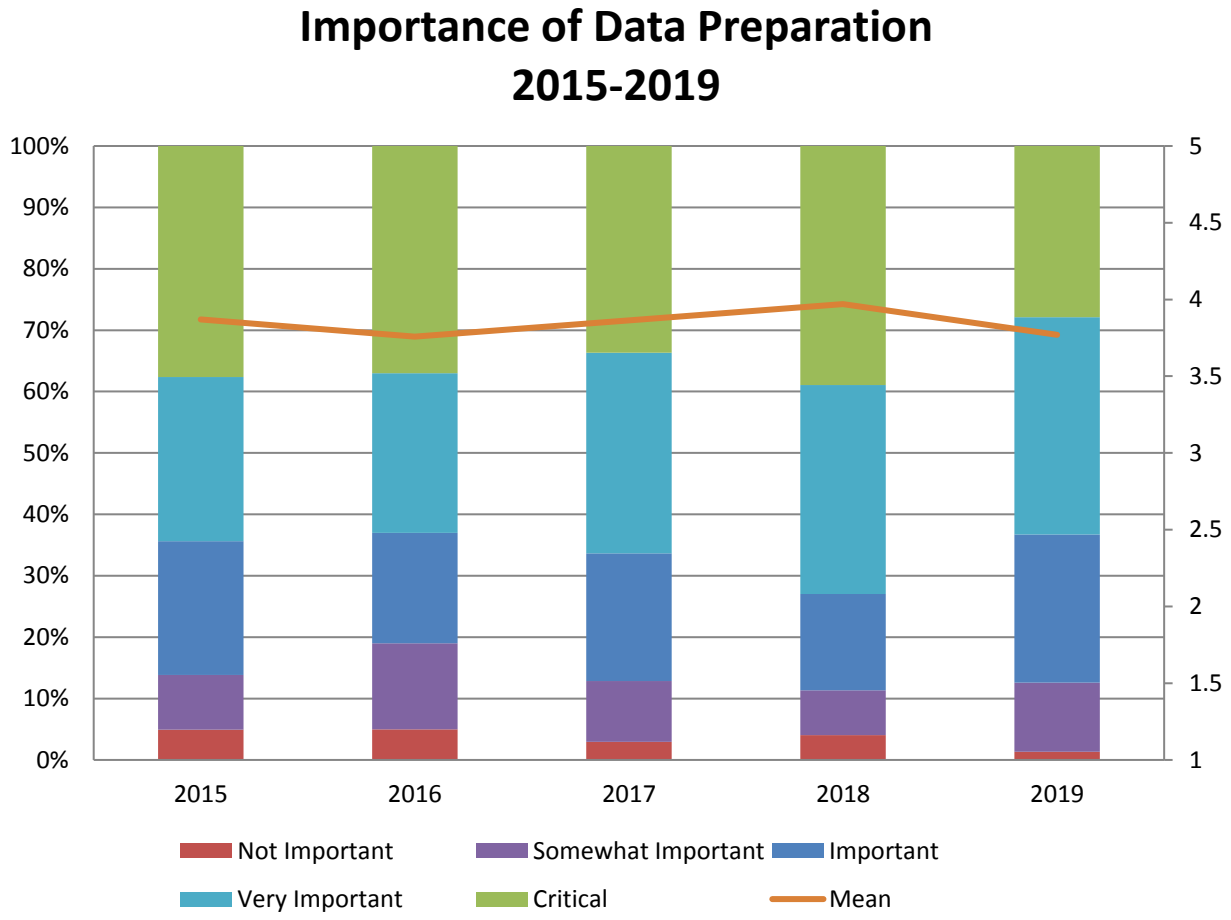


Figure 7 – Importance of data preparation 2015-2019

In 2019, Executive Management respondents report the highest "critical" and overall positive sentiment toward the importance of data preparation (fig. 8). Operations and Marketing/Sales report the next highest interest, and between 67 and 70 percent of all three functions say data preparation is, at minimum, "important." Interest wanes, though not precipitously, across R&D, Finance, BICC, and IT respondents. Overall weighted mean scores for all functions are between 3.6 and 4.1. Mostly average or lower interest in IT, BICC and R&D appear to indicate that data prep does not present an undue development or deployment hurdle in organizations.

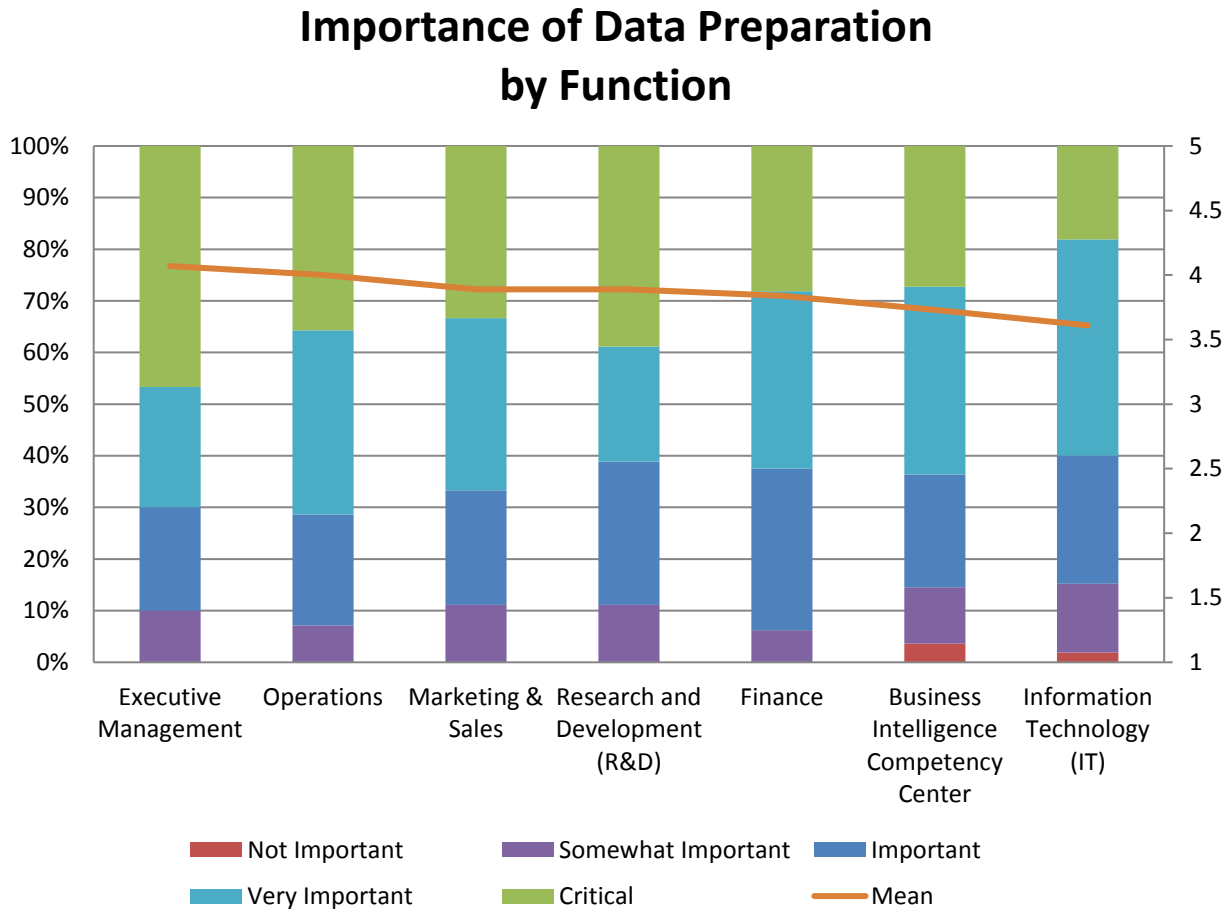


Figure 8 – Importance of data preparation by function

By geography, respondents in Asia Pacific and Latin America assign the greatest importance to data preparation in 2019 (fig. 9). Weighted mean importance in the aforementioned regions is near 4.0, compared to about 3.8 in North America and 3.6 in EMEA. Skepticism is low, however, across all geographies, with just 9-13 percent reporting that data preparation is only "somewhat important" or "not important."

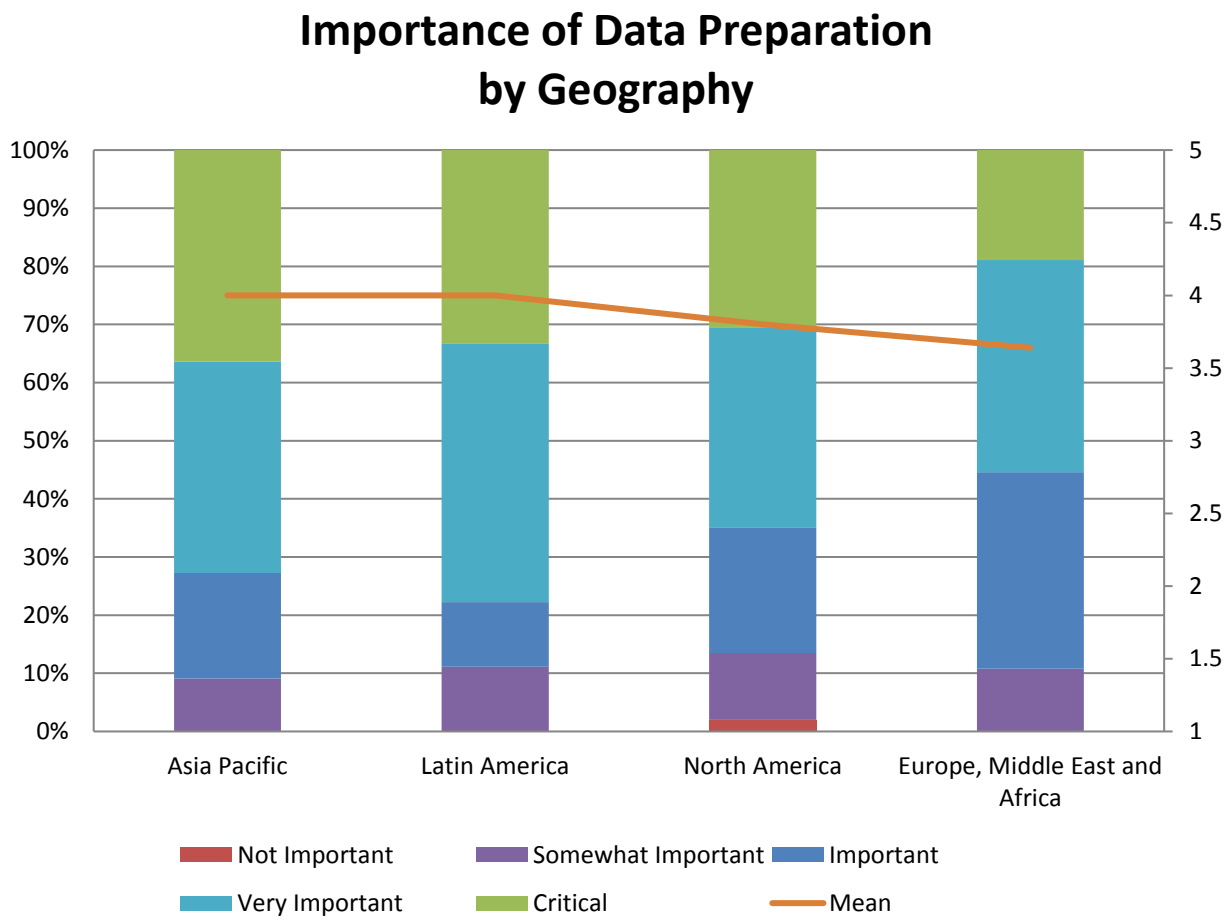


Figure 9 – Importance of data preparation by geography

In 2019, the importance of data preparation is similar across organizations of different size with weighted mean scores narrowly bunched between 3.7 and 3.8 (fig. 10). This is somewhat a departure from earlier studies that showed the importance of data preparation generally increasing with organization size. In 2019, "critical" sentiment is actually lowest at the largest organizations (25 percent) compared to the smallest (30 percent). Overall scores in the range of near "very important across the board" support our expectation that data preparation can be highly important and relevant to organizations of any size.

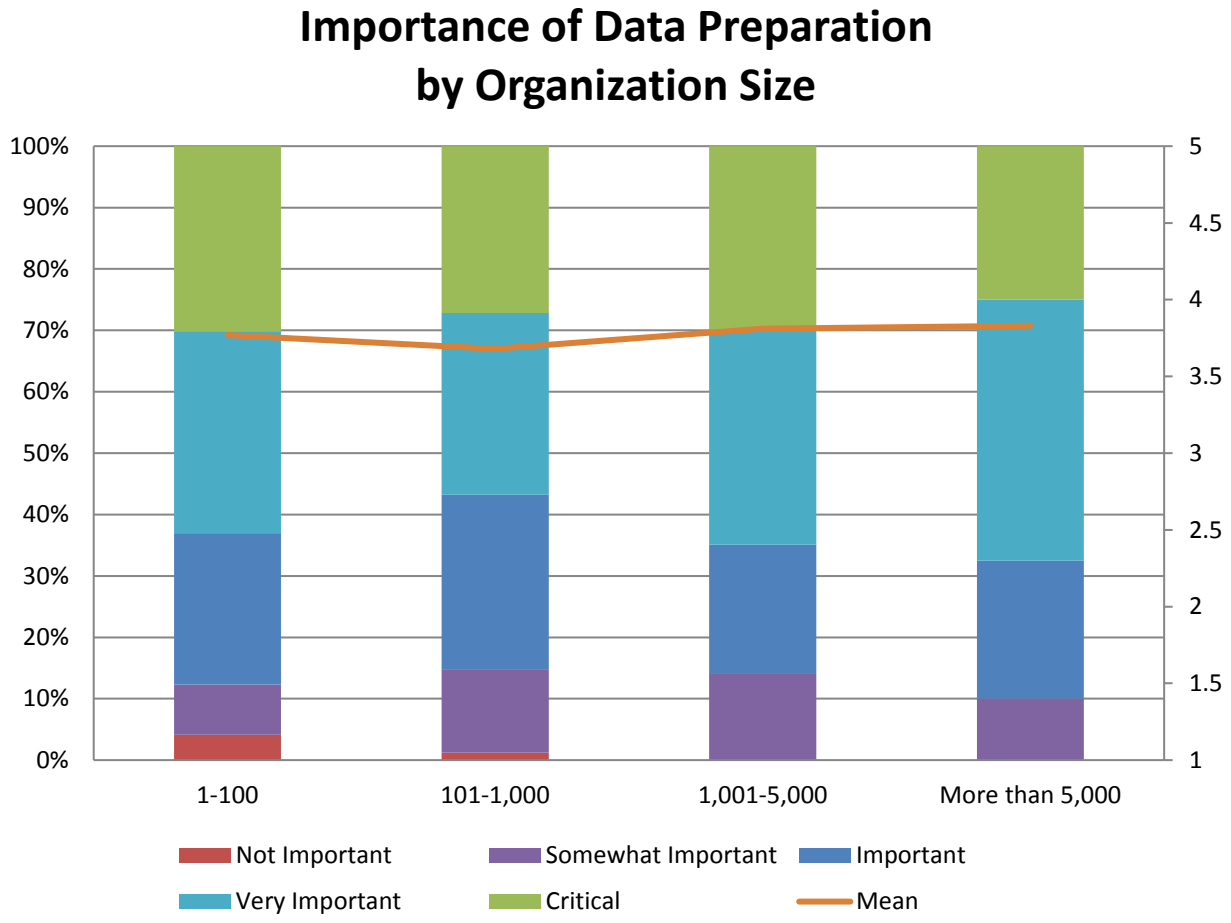


Figure 10 – Importance of data preparation by organization size

Compared to other measures, the importance of data preparation varies noticeably by industry (fig. 11). In 2019, Retail/Wholesale respondents report both the highest adjusted mean importance (4.3) and the most "critical" scores (47 percent). At the other end of the spectrum, Healthcare respondents' weighted mean (3.4) and "critical" scores (18 percent) are the lowest by industry. All other industries reported weighted mean importance between 3.7 and 4.0, or in the range of "very important."

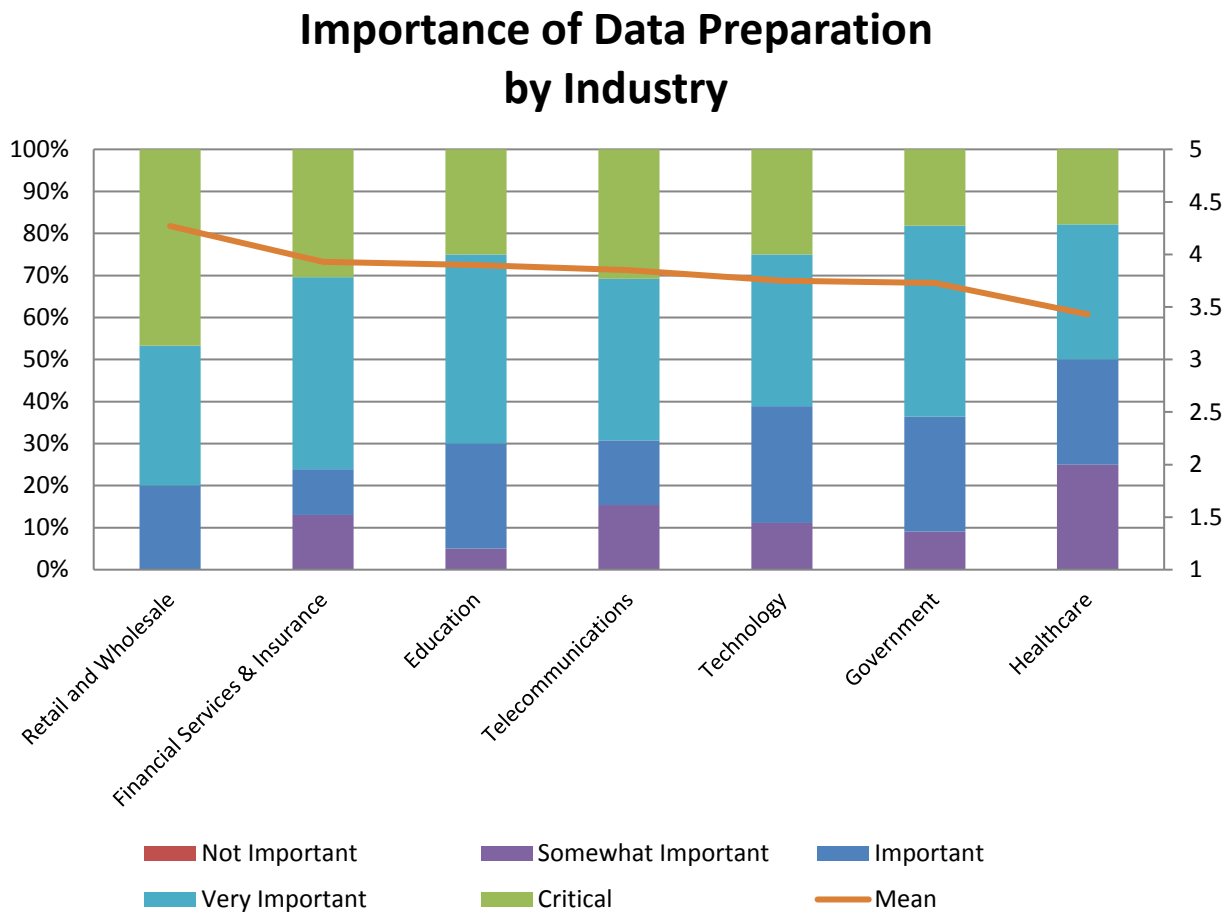


Figure 11 – Importance of data preparation by industry

Effectiveness of Current Approach to Data preparation

In 2019, the majority of organizations (59 percent) say their current data preparation approach is "somewhat effective" (fig. 12). Combined with those who report "highly effective" data preparation, the figure exceeds three-quarters (78 percent) of respondents. The remaining respondents, about 22 percent, report only "somewhat ineffective" or "totally ineffective" approach to data preparation. This positive overall response implies very positive interactions and growing maturity of data preparation, likely in the context of increasing self-service and user autonomy.

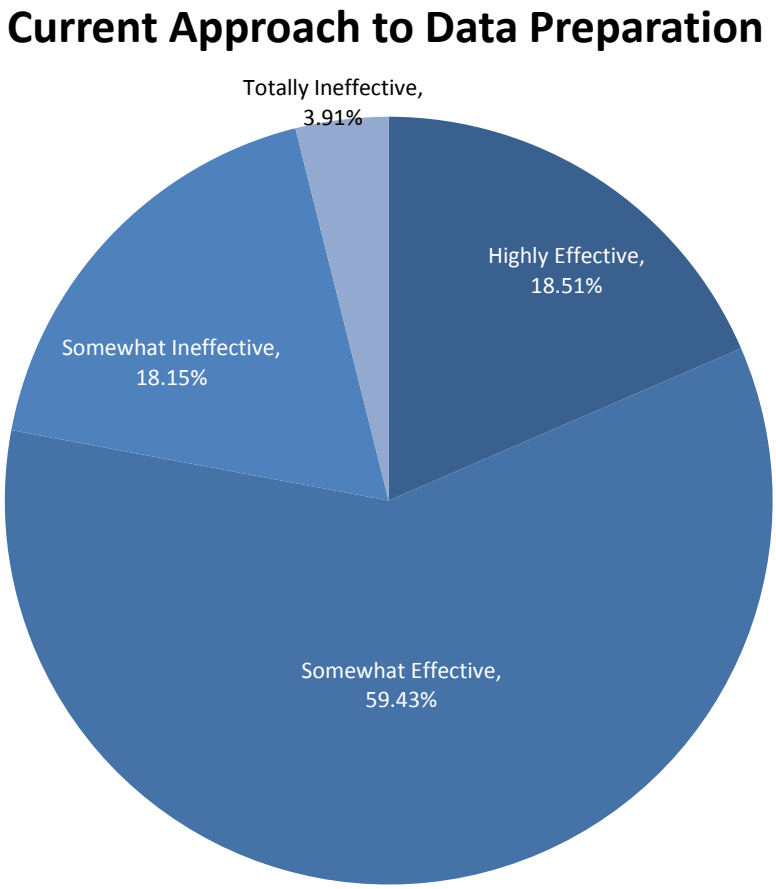


Figure 12 – Current approach to data preparation

Across five years of data collection, perceptions of data preparation effectiveness steadily improve (fig. 13). The number that report "highly effective" use declines slightly in 2019 but is offset by the increase in "somewhat effective" organizations. As positive scores grew over time, reports of "somewhat ineffective" or "totally ineffective" data preparation decline proportionately to all-time lows in 2019.

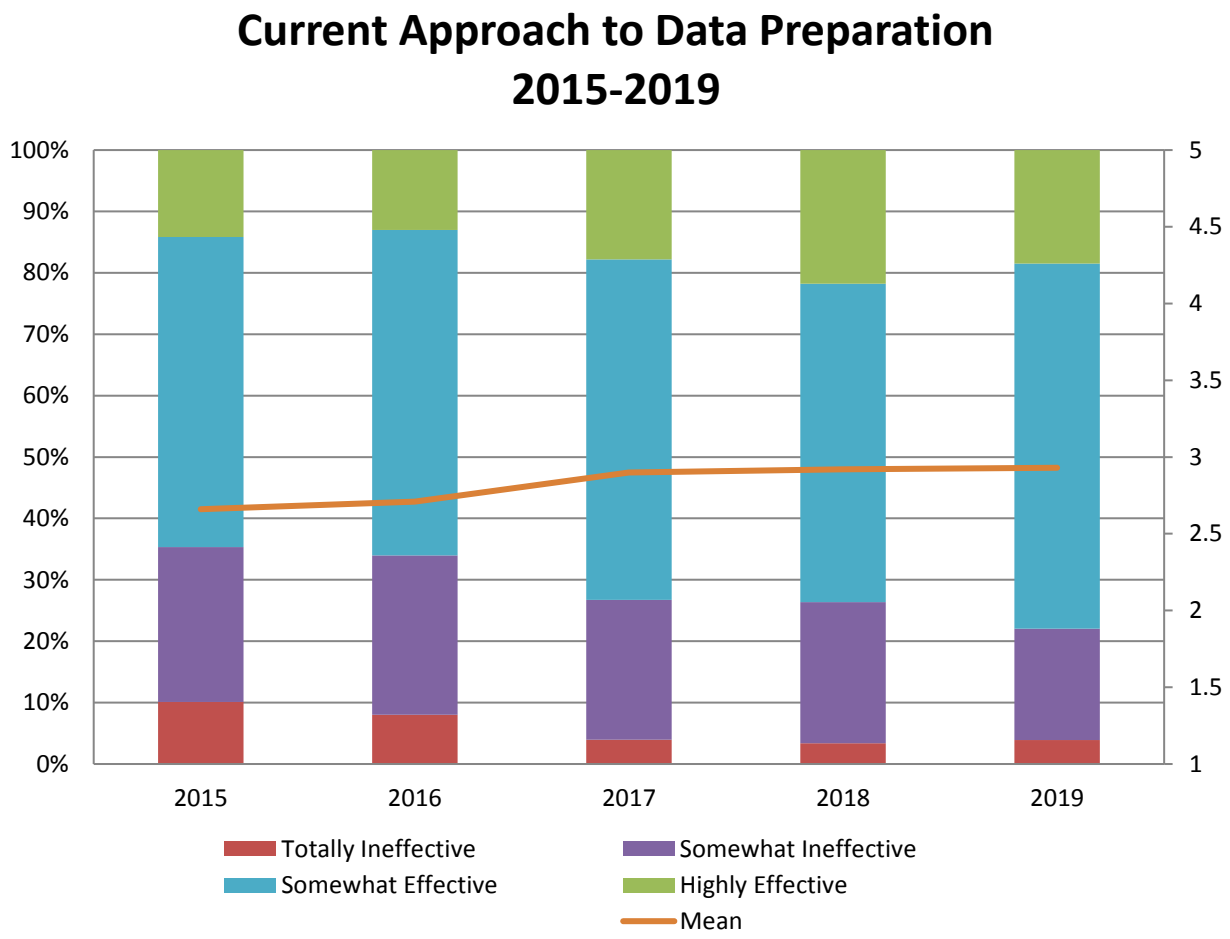


Figure 13 – Current approach to data preparation 2015-2019

All functions report results in the range of "somewhat effective" weighted mean satisfaction with data preparation in 2019 (fig. 14). Eighty-eight percent of respondents in Marketing/Sales say their approach to data preparation is "highly effective" or "somewhat effective." More than 80 percent of R&D, BICC, and Executive Management respondents agree. "Somewhat ineffective" or lower scores are marginally higher in operations, IT, and Finance.

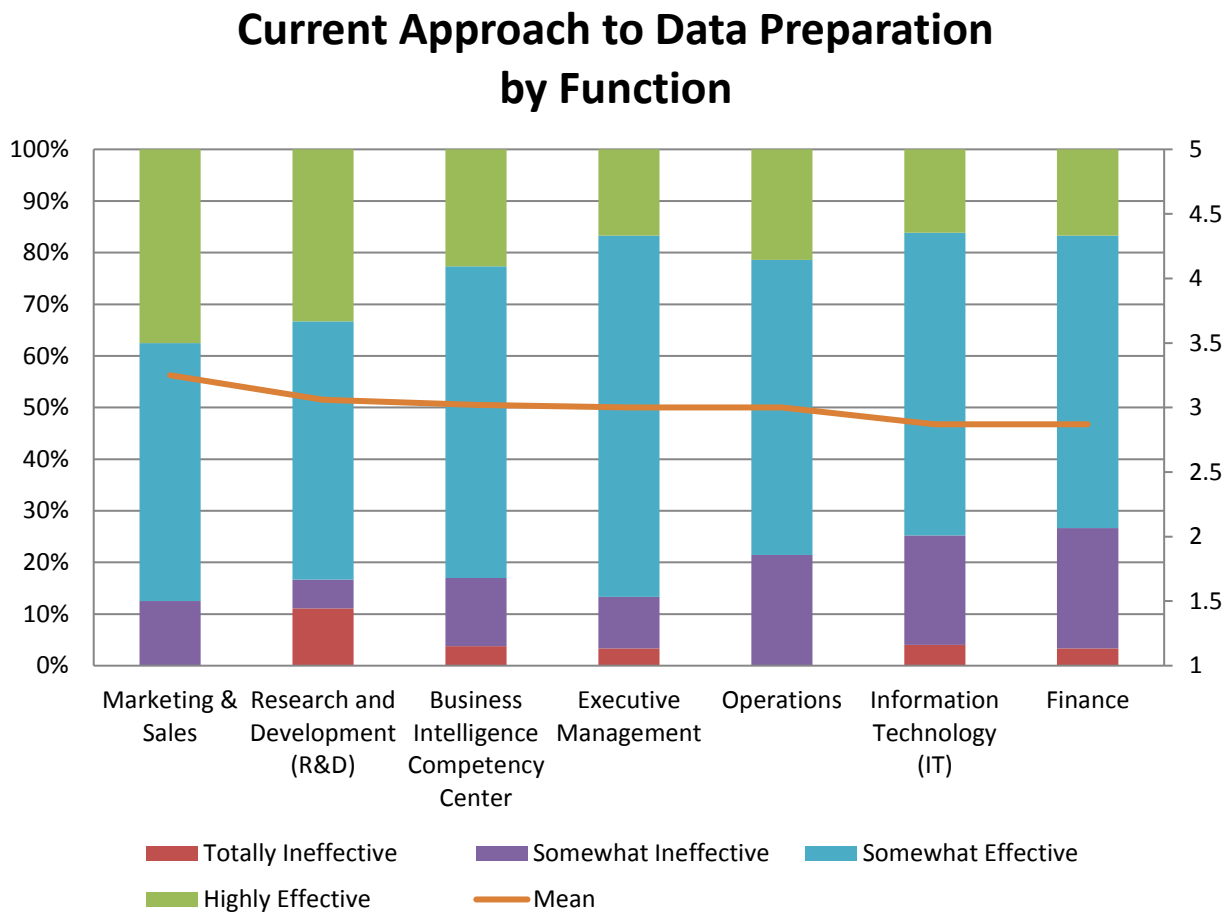


Figure 14 – Current approach to data preparation by function

The perceived effectiveness of data preparation in 2019 is highest among Asia-Pacific (weighted mean = 3.2) and North American (3.0) respondents, compared to 2.8 in other regions (fig. 15). Combined "highly effective" and "somewhat effective" responses account for 90 percent of respondents in Asia Pacific compared to 79 percent or fewer in other geographies.

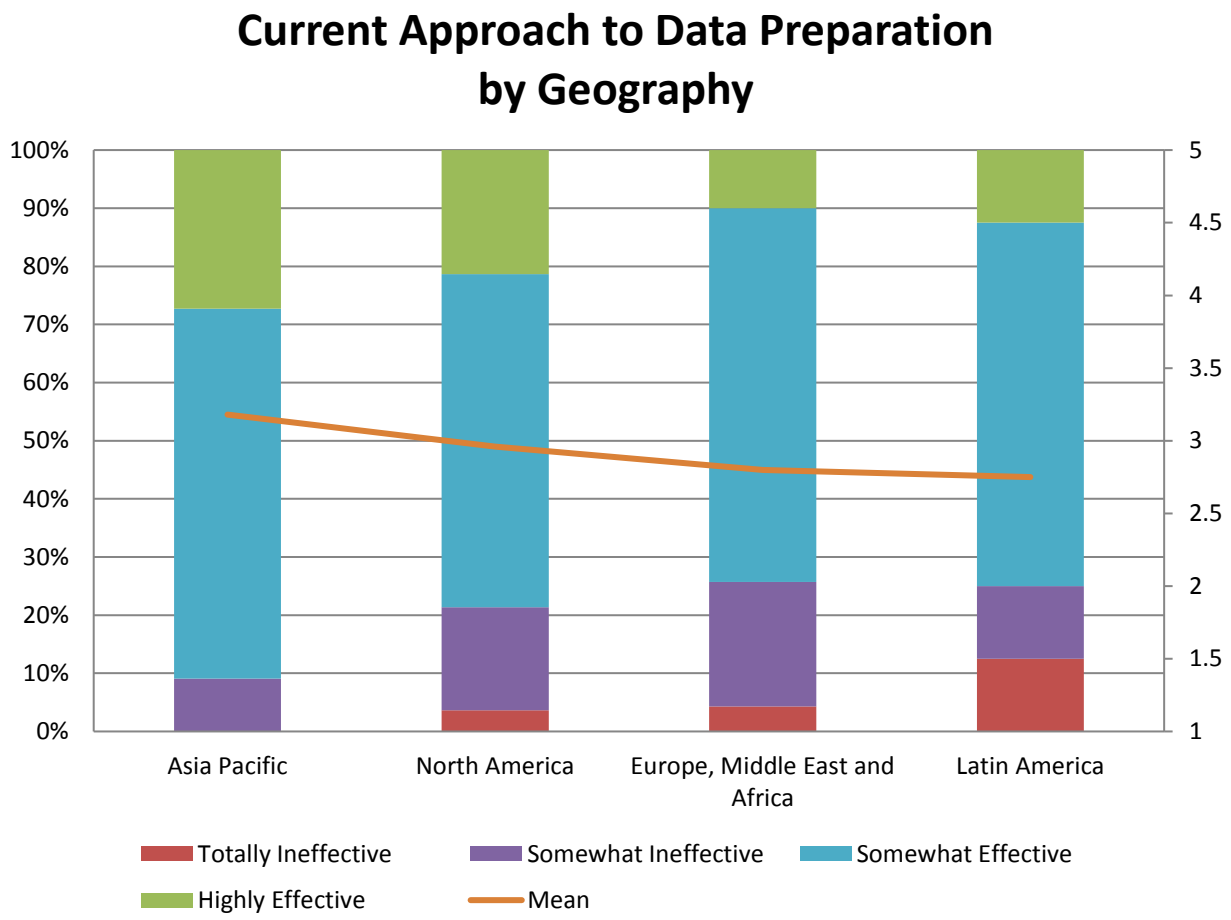


Figure 15 – Current approach to data preparation by geography

Weighted mean scores for data preparation effectiveness in 2019 fall in a narrow range (2.9-3.0) for organizations of different size (fig. 16). However, we find more "highly effective" data preparation in about 20 percent of small and mid-sized organizations, compared to 13 percent at large organizations and 17 percent at organizations with more than 5,000 employees. Despite these differences, organizations of different size all report their current approach in the range of "somewhat effective."

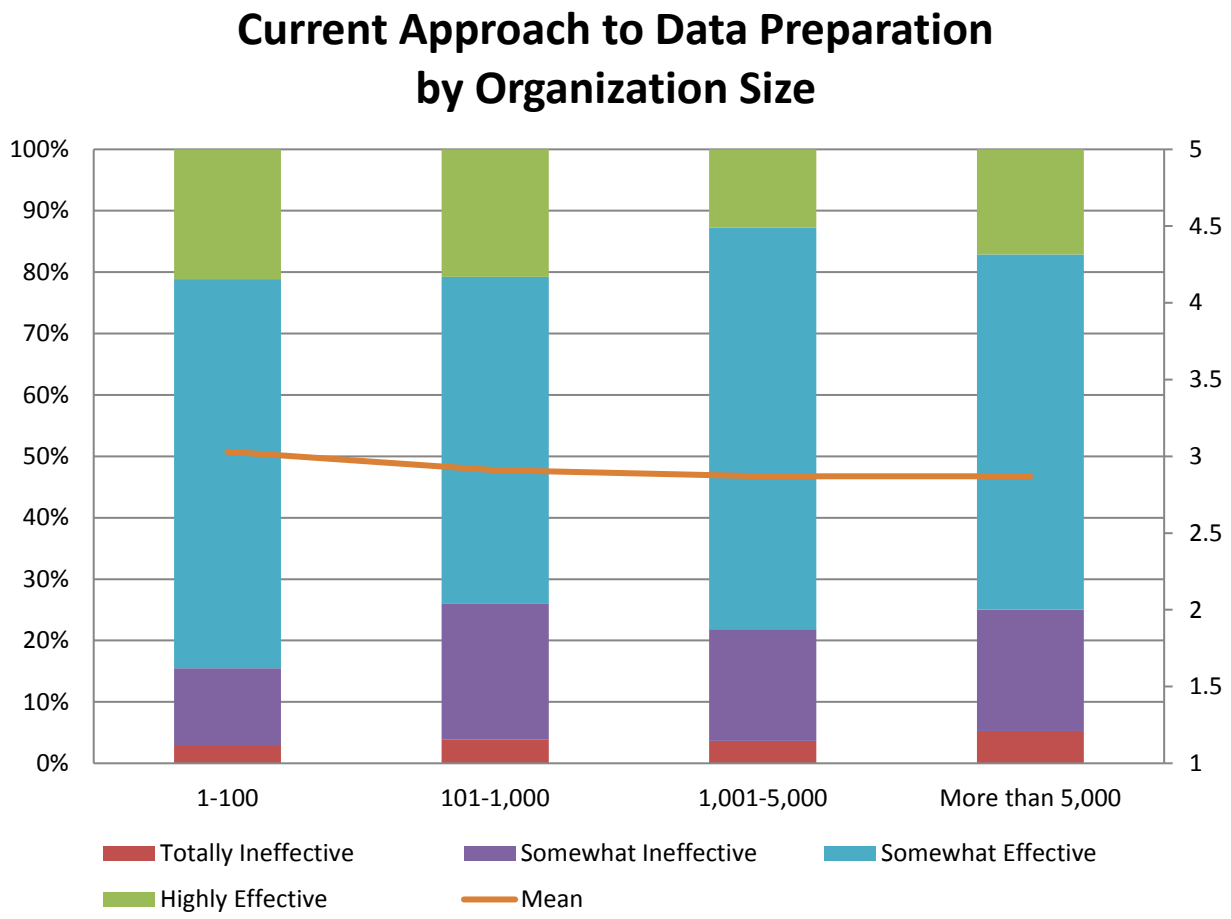


Figure 16 – Current approach to data preparation by organization size

By industry, the perceived effectiveness of data preparation is highest in Financial Services (weighted mean = 3.3), followed by Technology (3.0) and Telecommunications (2.9) organizations (fig. 17). Retail/Wholesale, Higher Education, and Government make up a next tier of effectiveness with weighted mean scores between 2.7 and 2.8. As is the case with data preparation importance (fig. 11, p. 24), Healthcare respondents report data preparation effectiveness lowest, where close to half (47 percent) of respondents report "somewhat effective" or lower effectiveness.

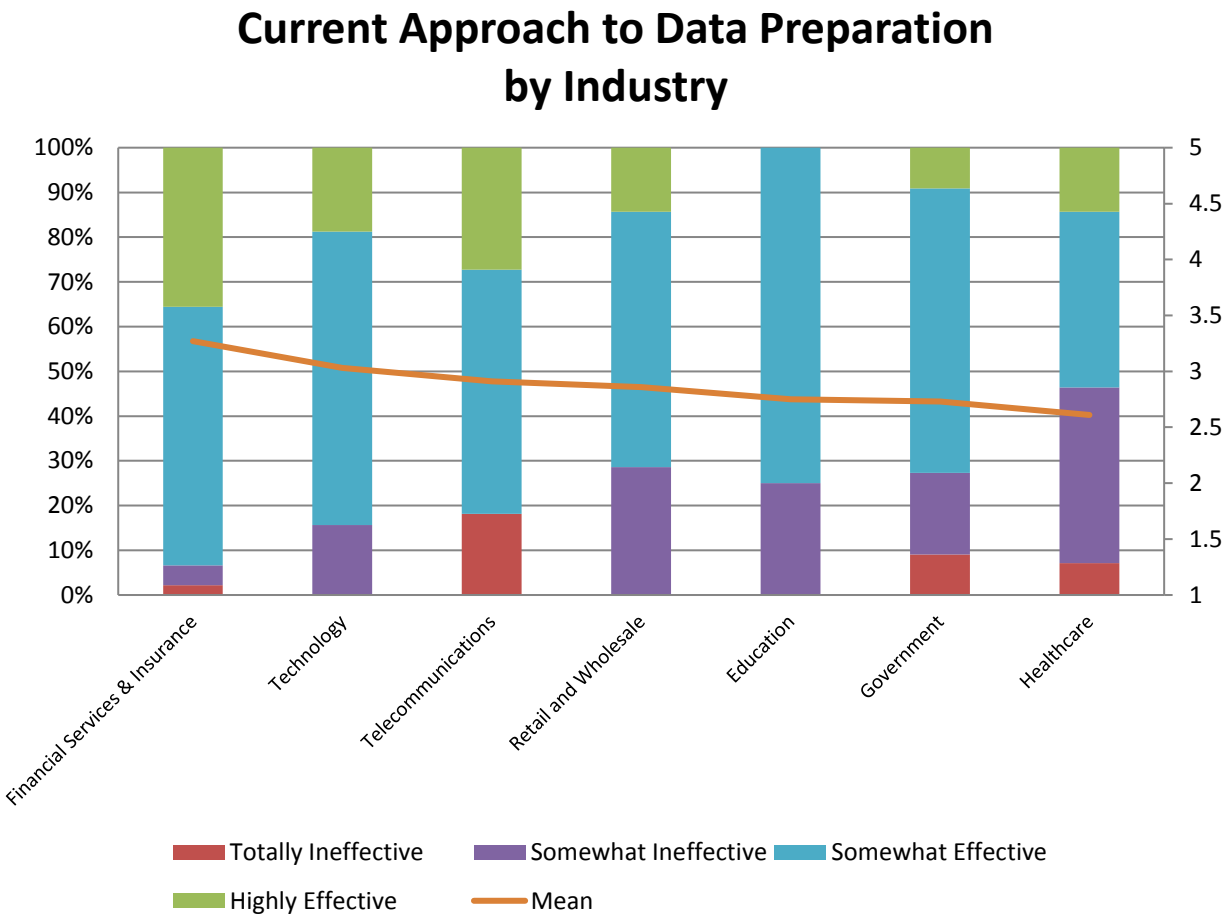


Figure 17 – Current approach to data preparation by industry

Success with data preparation correlates to success with business intelligence (fig. 18). We would expect this since data preparation can be considered a core activity of any BI undertaking beyond the most rudimentary. In 2019, 85 percent of organizations that are "completely successful" or "somewhat successful" with BI also report "highly effective" or "somewhat effective" data preparation. At the opposite end of the spectrum, organizations that are "unsuccessful" with BI are less than 60 percent likely to be "somewhat effective" with data preparation. No "unsuccessful" BI programs are associated with highly effective data preparation.

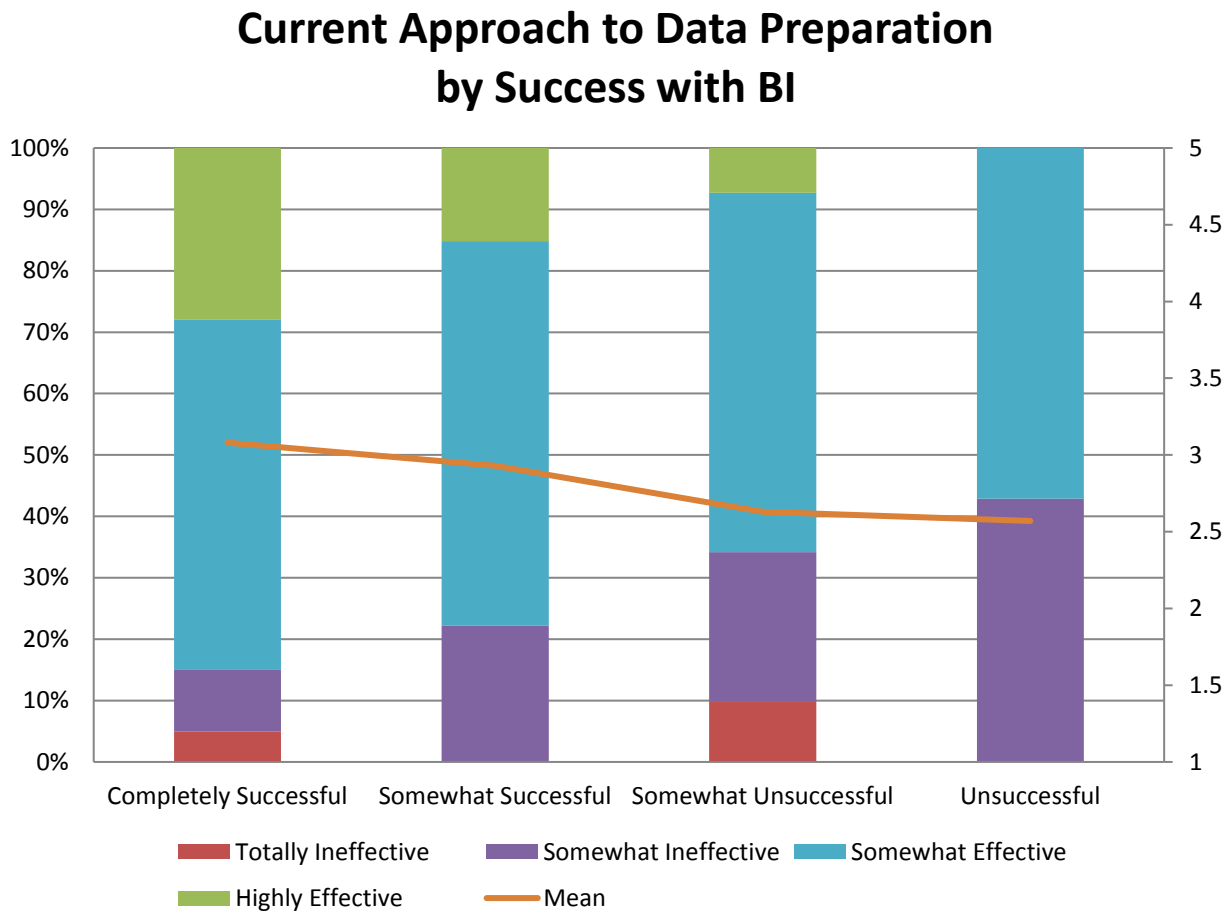


Figure 18 – Current approach to data preparation by success with BI

Frequency of Data preparation

Sixty-six percent of respondents say they "constantly" or "frequently" make use of data preparation (fig. 19). We cannot distinguish whether end-user efforts are unique or repeated practices, but overall usage of data preparation appears to be high, with another 26 percent reporting at least "occasional" data preparation activity. Only about 8 percent of respondents "rarely" or "never" perform data preparation.

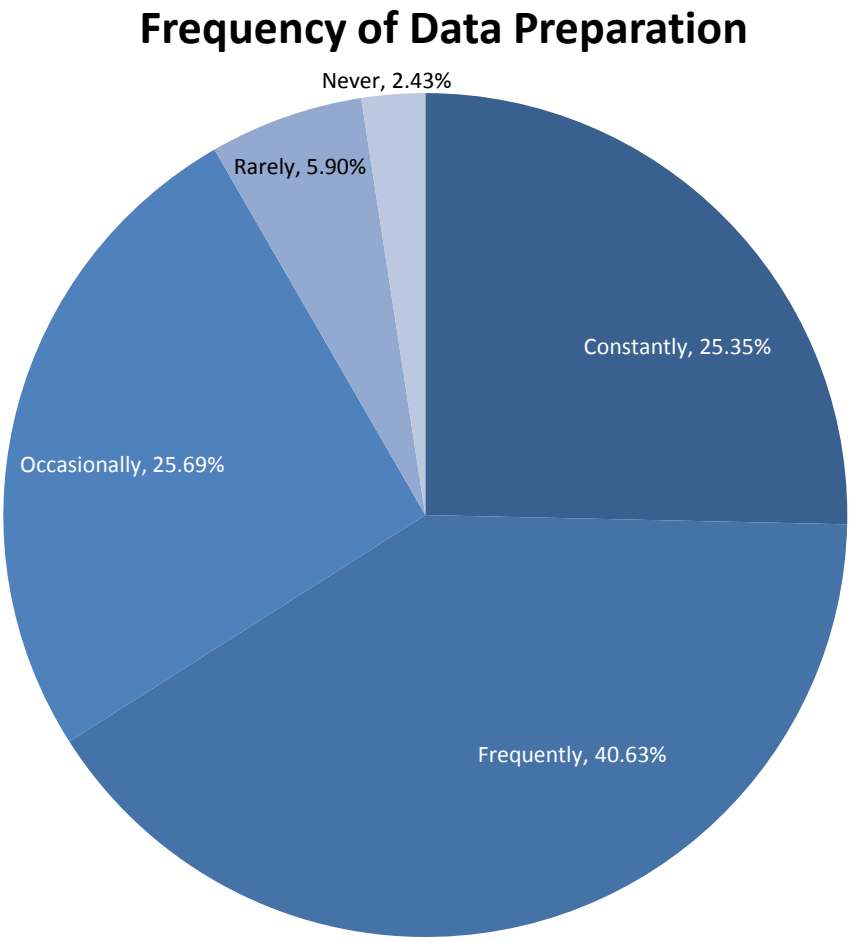


Figure 19 – Frequency of data preparation

Across the five years of our focused data preparation study, respondents report frequency of use as mostly unchanged (fig. 20). Weighted mean importance remains consistently high in a range of 3.8 to 3.9, or near a value of "frequently." In 2019, "constant" use is steady year over year, while "frequent" use declines (from 46-40 percent) compared to 2018.

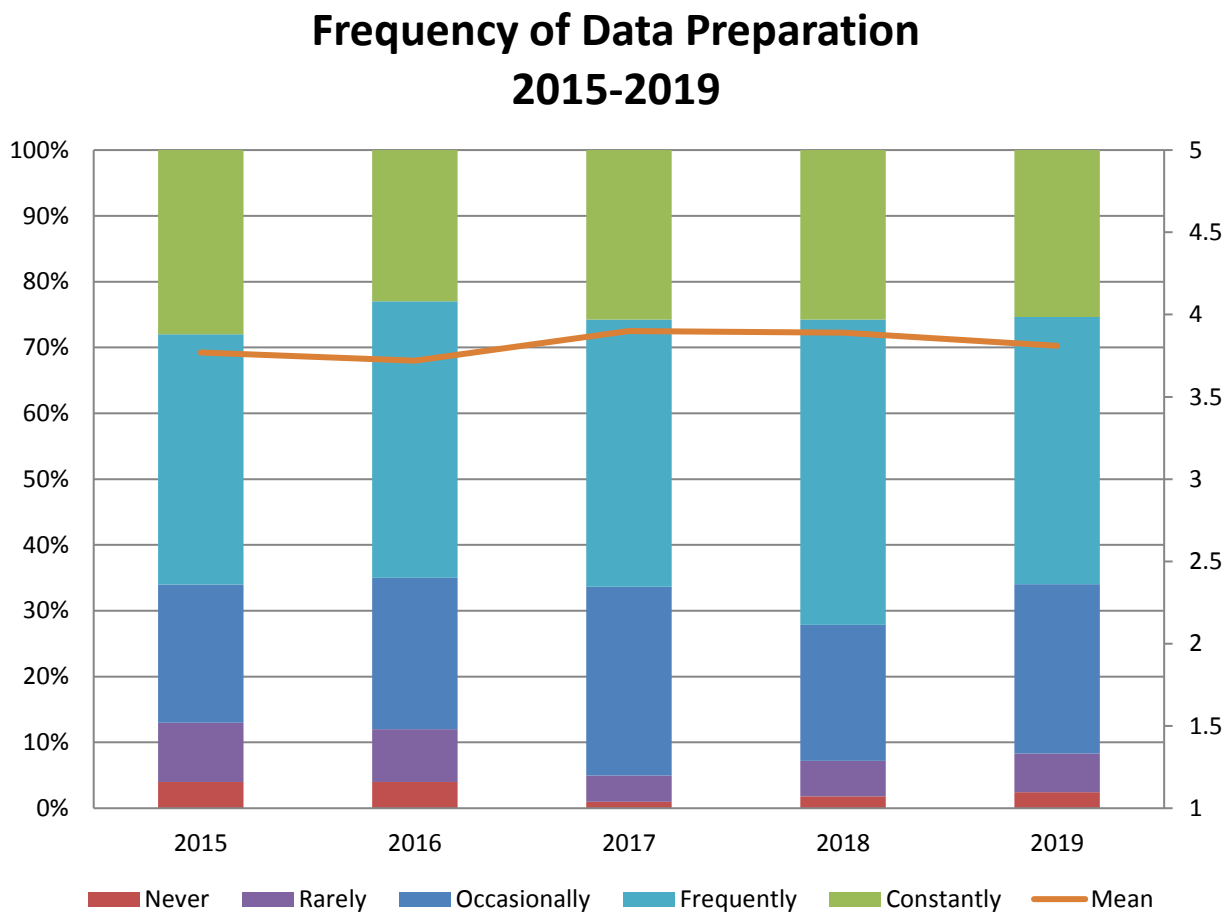


Figure 20 – Frequency of data preparation 2015-2019

We would expect that issues of business performance and revenue would drive the frequency of use of data preparation. In 2019, however, we see a possibly interesting union between R&D activity (>40 percent "constant" users) and the use of data preparation (fig. 21). Finance and Marketing/Sales, which report the most frequent use in 2019, are the next most active users by function. Also interesting, close to 40 percent of Executive Management respondents are "constant" users. BICC and IT represent the least frequent users by function, possibly indicating that data preparation is easily acquired and adopted without much internal intervention.

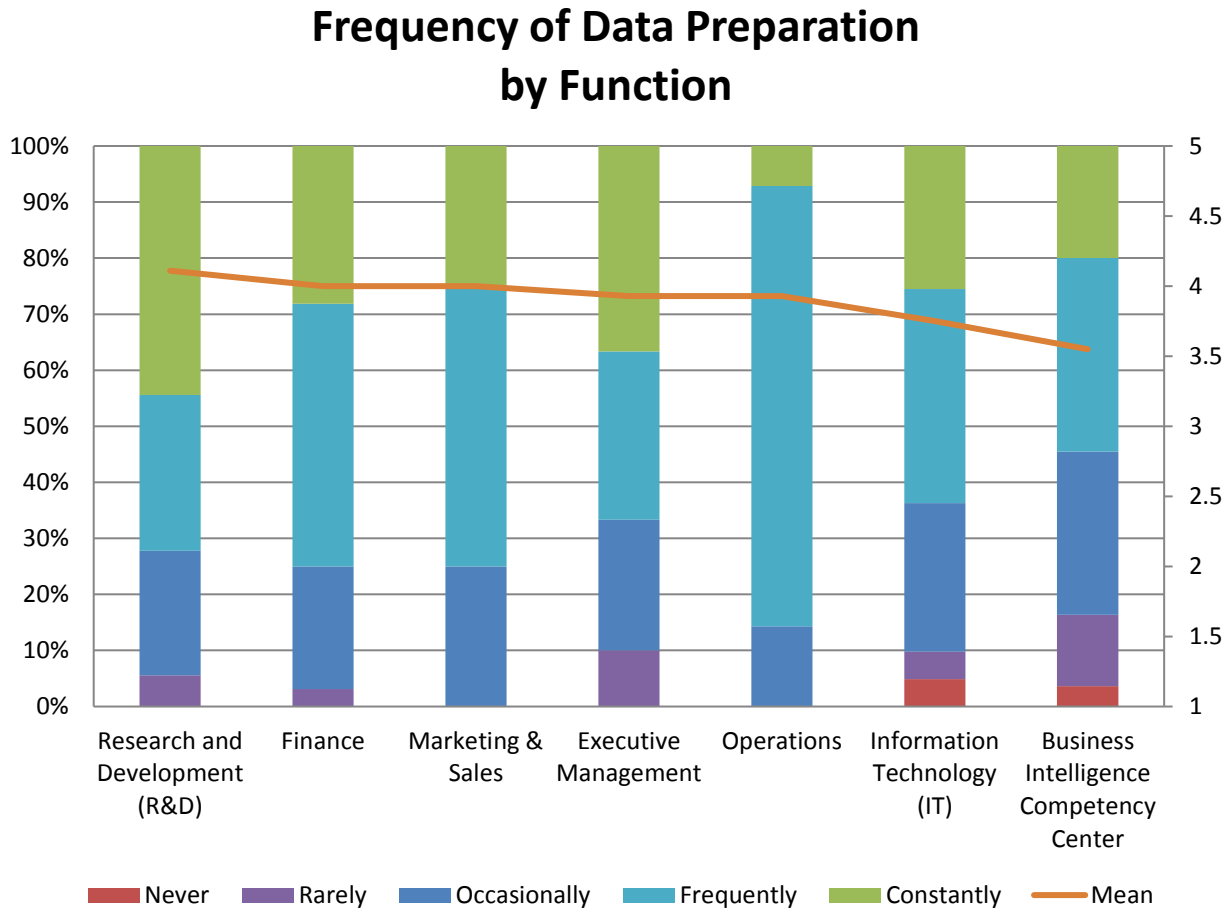


Figure 21 – Frequency of data preparation by function

In 2019, the number of "constant" data preparation users is highest in North America (30 percent) (fig. 22). Combined "constant" and "frequent" users are also highest in North America, followed by Latin America, EMEA and Asia Pacific. Fewer "constant" users, however, are reported in Latin America and in EMEA. Overall, more than 90 percent of respondents in all geographic regions report, at minimum, "occasional" use of data preparation.

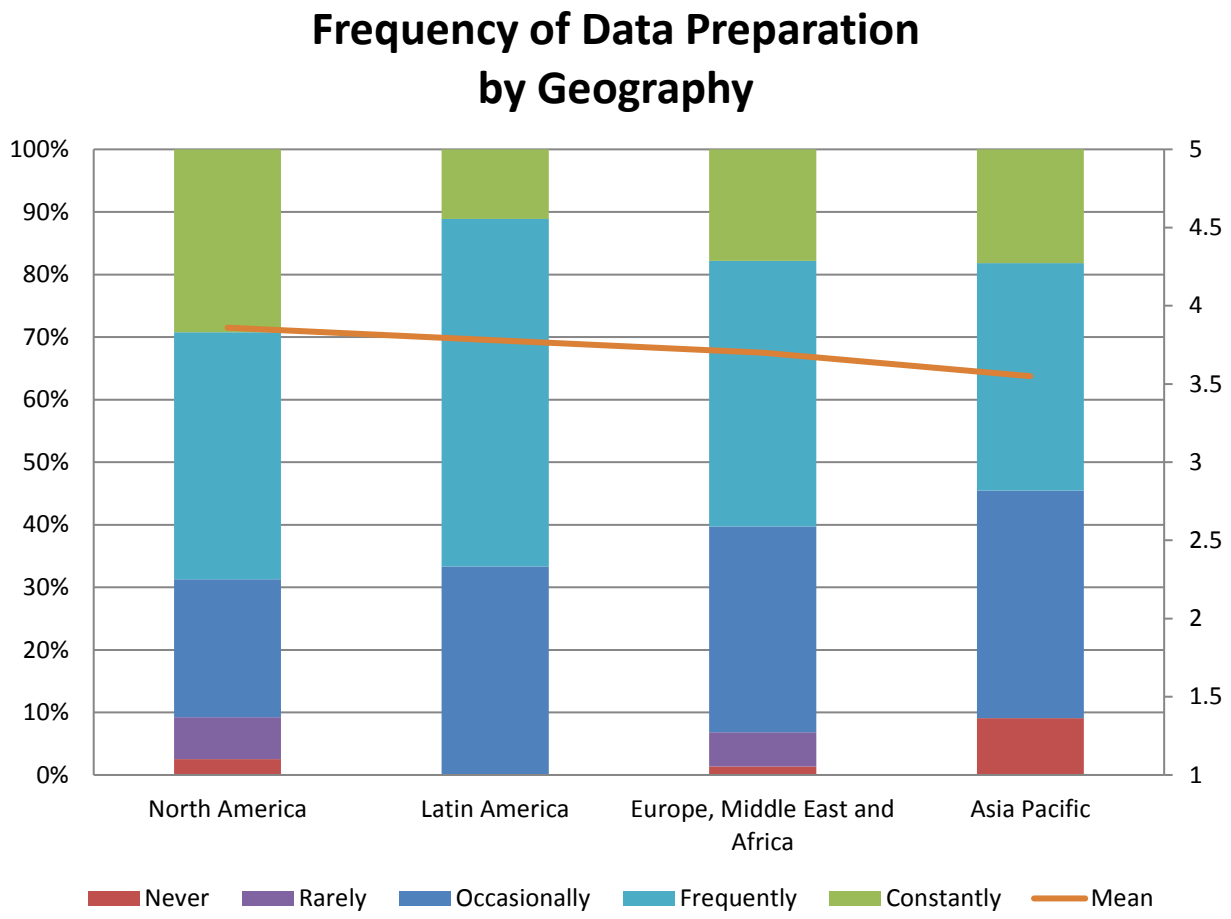


Figure 22 – Frequency of data preparation by geography

Mean frequency of data preparation in 2019 increases marginally with organization size (fig. 23). Combined "constant" and "frequent" usage is highest at large organizations (> 1,000 employees) and thereafter decreases with global headcount. However, "constant" users are most often found (30 percent) at small organizations (1-100 employees), a figure that partly reflects lower headcount. All organizations of any size report weighted mean frequency in a narrow range between 3.7 and 3.9.

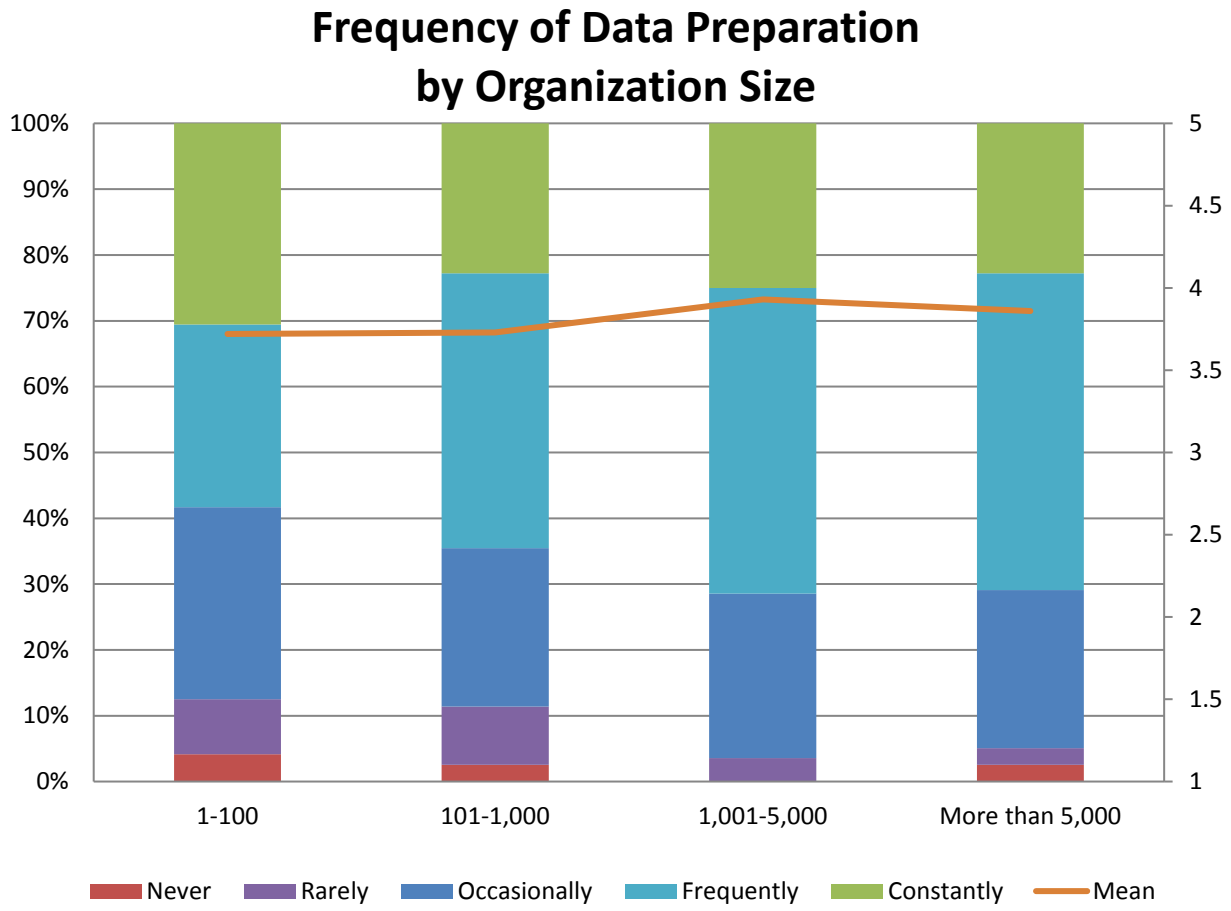


Figure 23 – Frequency of data preparation by organization size

Data preparation frequency in 2019 is markedly higher at Telecommunications organizations, where 64 percent of respondents report "constant" usage (fig. 24). Technology organizations report the next most "constant" users (40 percent). Outside Financial Services, all other industries say 20 percent or fewer are "constant" users, with the fewest (10 percent) in Higher Education. Fewer than 10 percent of organizations report "rare" or "never" usage of data preparation.

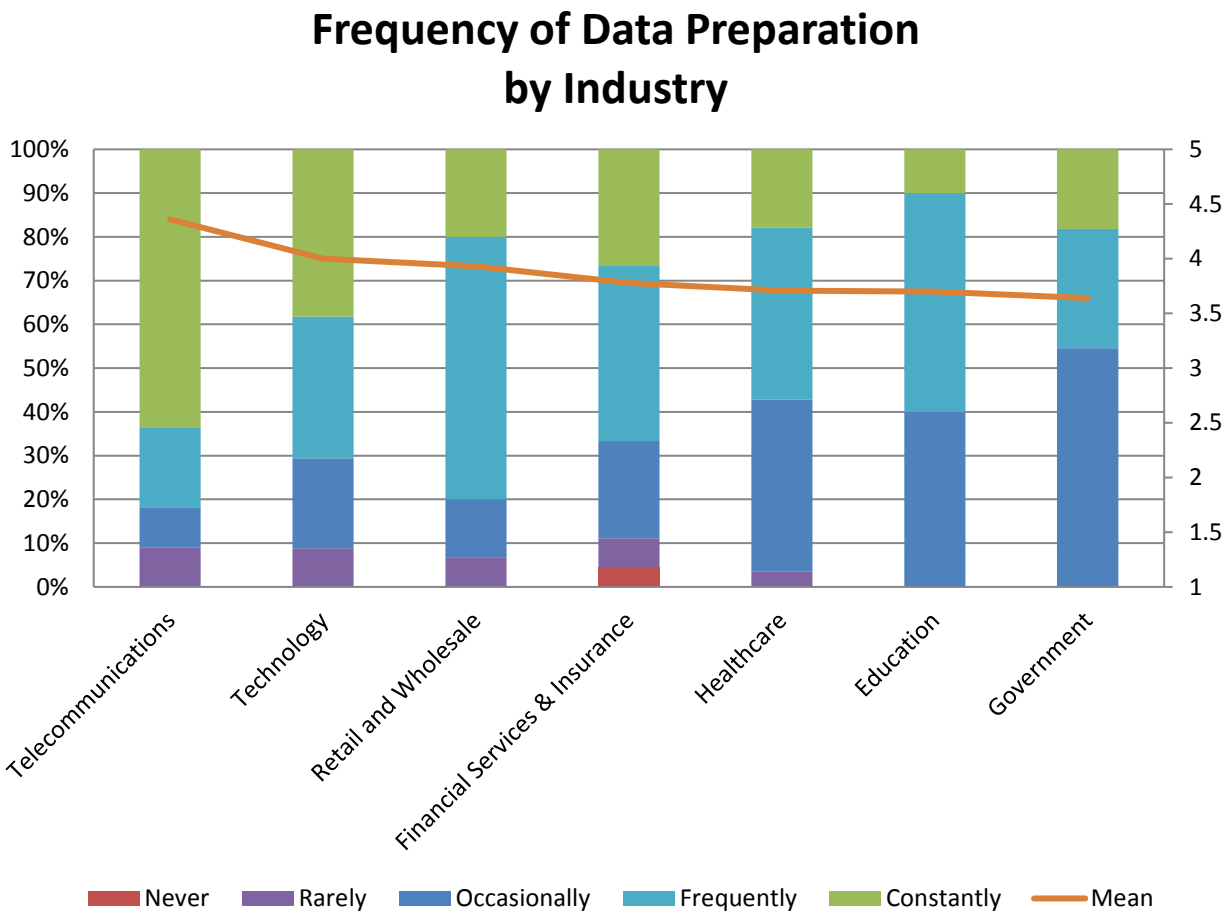


Figure 24 – Frequency of data preparation by industry

Frequency of Data Preparation Enrichment with Third-Party Data

Less than a quarter (24 percent) of respondents indicate they "constantly" or "frequently" enrich data preparation with third-party data (fig. 25). "Frequent" or "occasional" enrichment is the most likely level of activity (54 percent). This might suggest that the select 5 percent of "constant" users of non-proprietary data are found in unique roles and industries, with specific use cases.

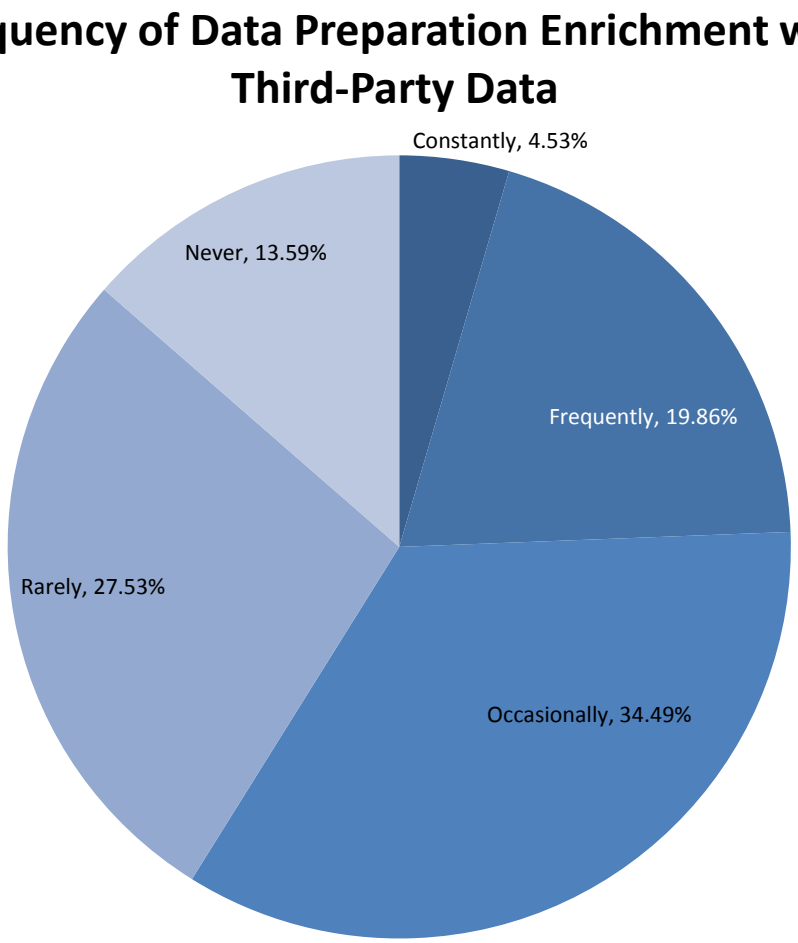


Figure 25 – Frequency of data preparation enrichment with third-party data

Across five years of data, respondents report mostly consistent frequency of third-party data use in conjunction with data preparation (fig. 26). By weighted mean, third-party data enrichment remains in a narrow range between 2.7 and 2.9. While we might have expected third-party data enrichment use to grow in this time span, 2019 activity actually decreases to 2.74 by weighted mean, clearly indicating that organizations continue to grapple most often with internal data.

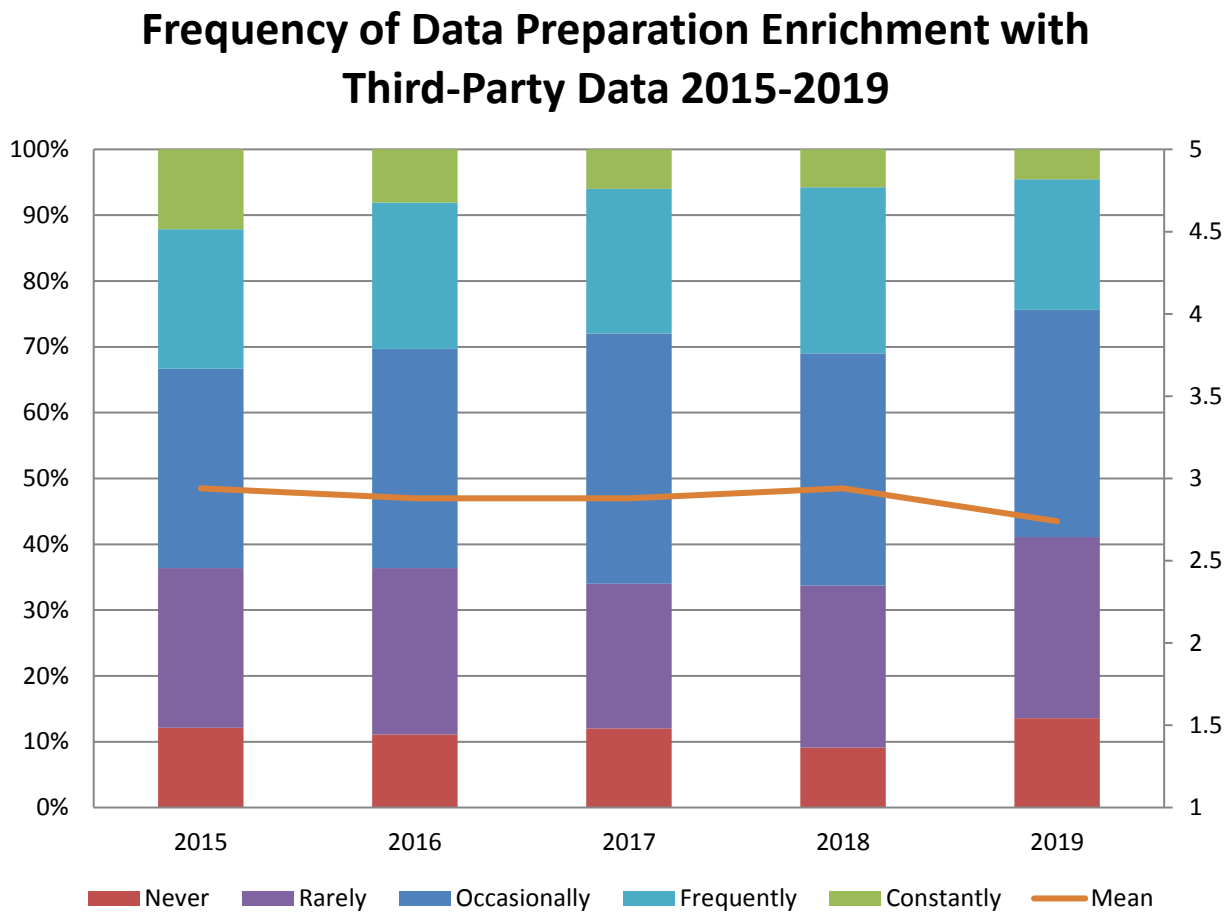


Figure 26 – Frequency of data preparation enrichment with third-party data 2015-2019

By function, Marketing/Sales holds a telling dominance in the frequency of enriching data preparation with third-party data in 2019 (fig. 27). We would expect that at least some of this enrichment is provided through third-party hosted applications as well. Beyond the 43 percent of Marketing/Sales respondents that say they are "constant" users, frequency declines rapidly. The group that is the next most likely to at least "frequently" use data preparation is in Finance, followed by the BICC. Respondents in operations are more than 60 percent likely to "never" or "rarely" enrich data preparation with third-party data.

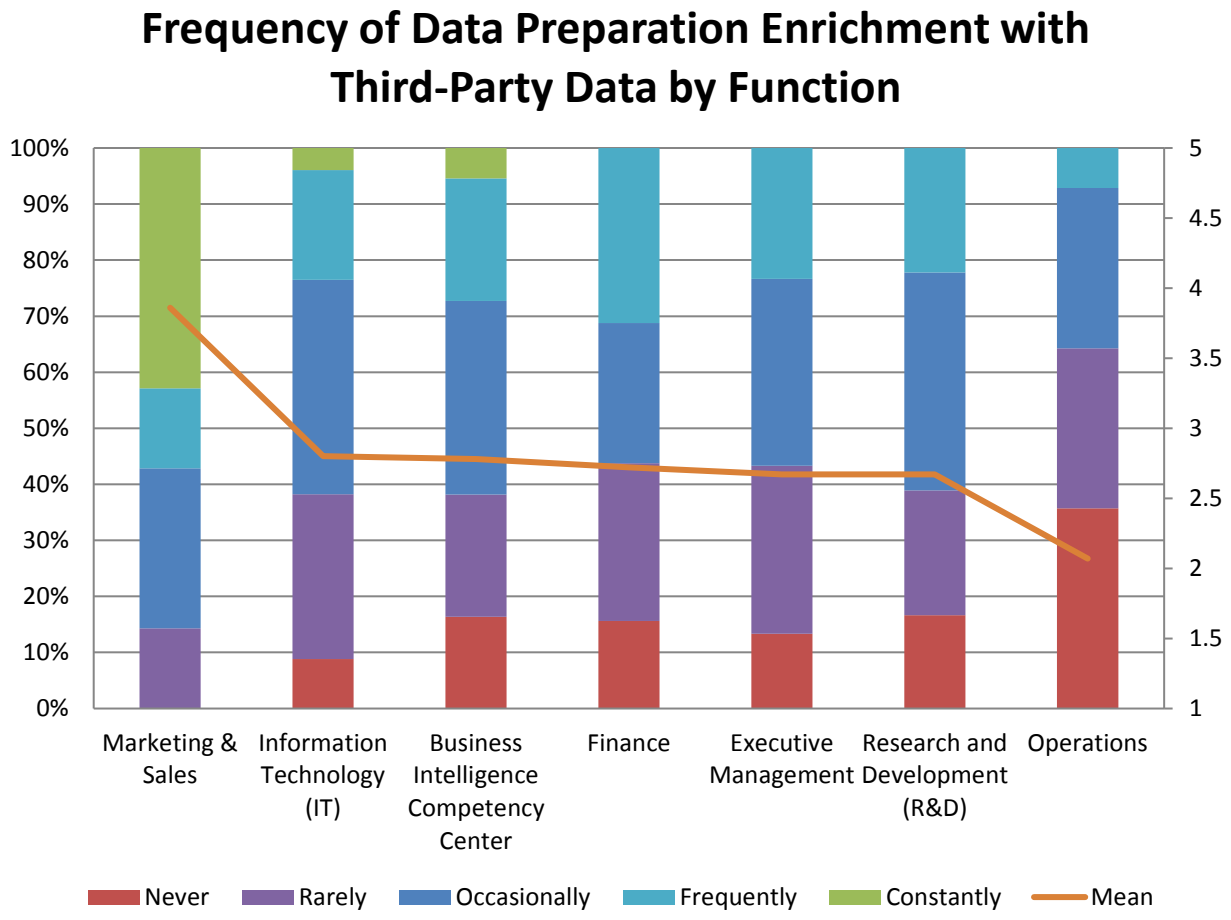


Figure 27 – Frequency of data preparation enrichment with third-party data by function

Interest in third-party data enrichment of data preparation is greatest in Asia Pacific, followed by North America and EMEA (fig. 28). Respondents in Asia Pacific are more than 70 percent likely to be, at minimum, "occasional" users, compared to about 60 percent in North America and 52 percent in EMEA. Mean global sentiment toward third-party data enrichment use is again generally between 2.4 and 3.0, in the range of "occasional" use or less.

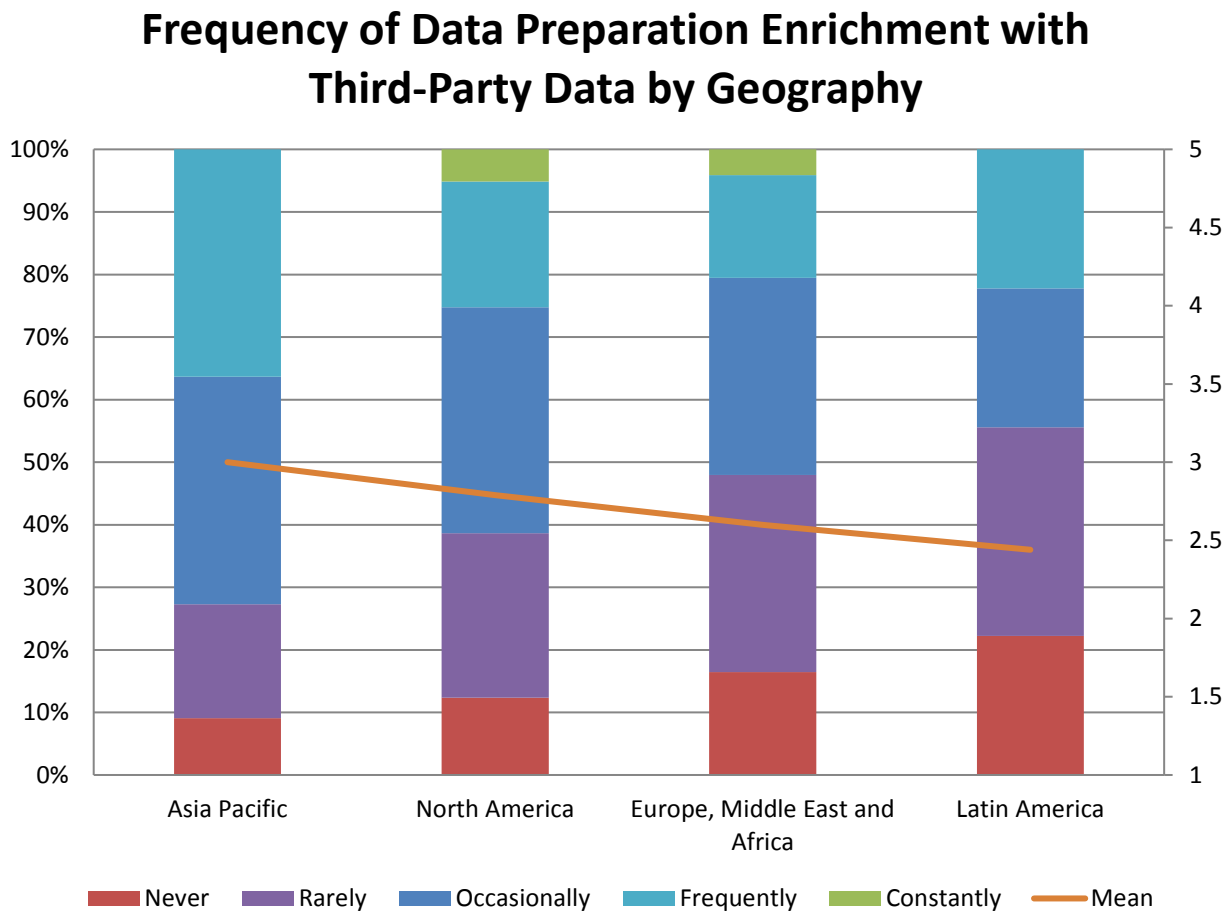


Figure 28 – Frequency of data preparation enrichment with third-party data by geography

Use of third-party data enrichment in data preparation is not steady or linear across organizations of different sizes (fig. 29). Very large organizations (>5,000 employees) are most likely (33 percent) "constant" or "frequent" users of third-party data. Mid-sized organizations (101-1,000 employees) are the next most likely "constant" or "frequent" users. Weighed mean levels of third-party data use are in a range between 2.6 and 2.9, which indicates "occasional" or slightly less use.

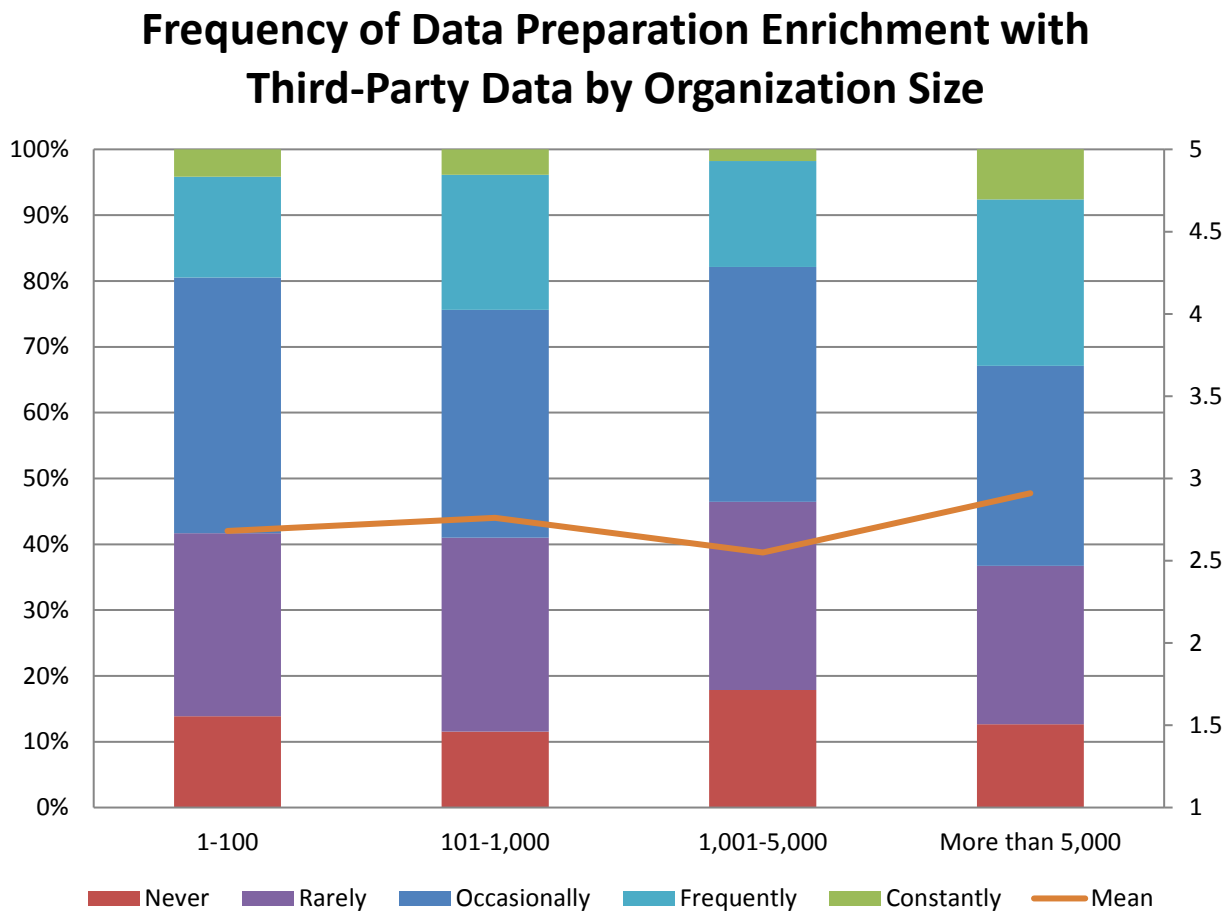


Figure 29 – Frequency of data preparation enrichment with third-party data by organization size

By industry, highly transactional and fraud-sensitive Financial Services / Insurance organization respondents are narrowly the most likely frequent users of third-party data enrichment (fig. 30). Retail/Wholesale respondents are almost as engaged with third-party data, perhaps in areas that include census, weather, social media, and syndicated sources. Thereafter, "frequent" use declines across industries, though at least "occasional" usage is reported by more than one-half or well more of respondents in all industries.

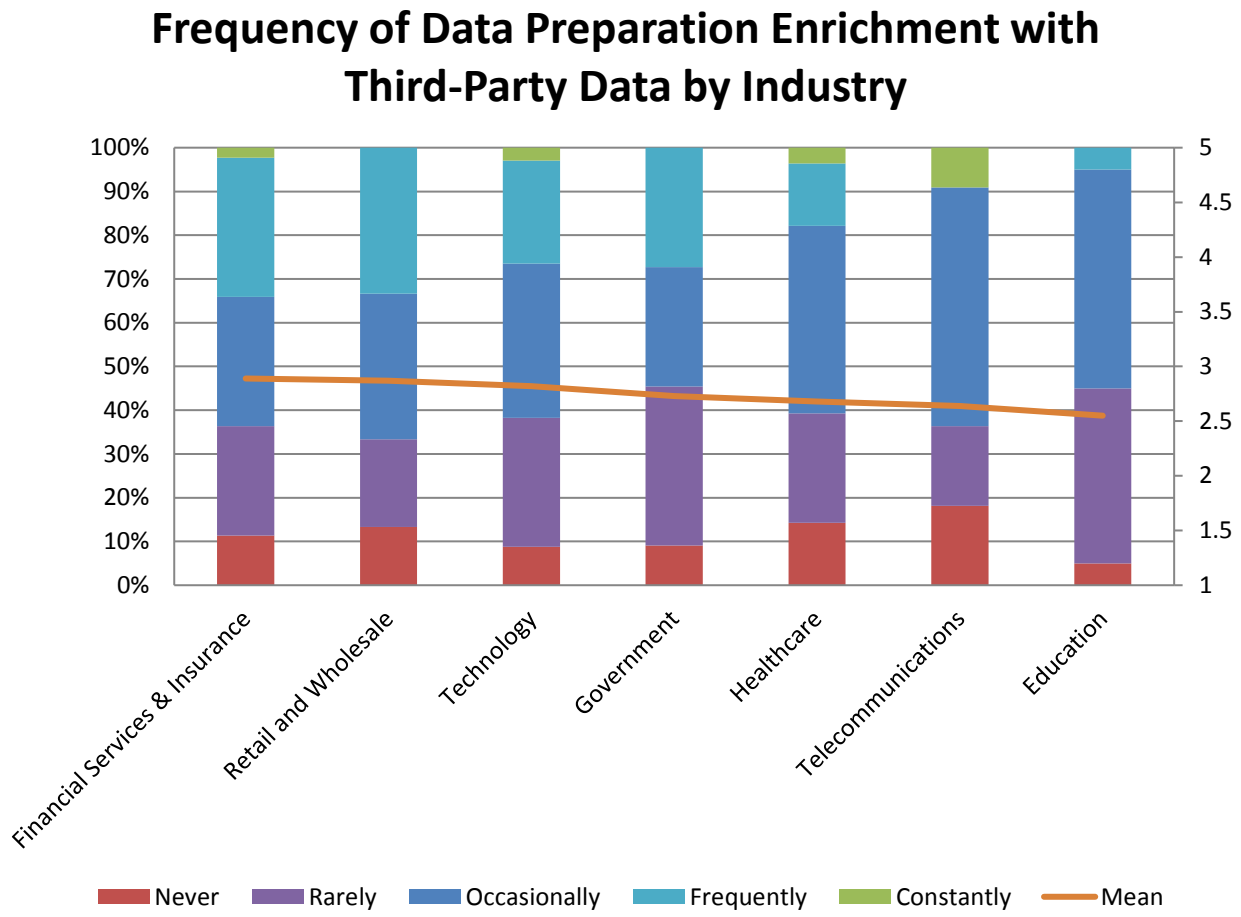


Figure 30 – Frequency of data preparation enrichment with third-party data by industry

Data preparation Usability Features

In 2019, there is strong interest in a wide range of data preparation usability features, all of which are at least "somewhat important" to 60 percent or far more respondents (fig. 31). We believe this reflects good understanding of needs and high expectations for basic to advanced data preparation features. The most important of these are utilitarian save-and-preview capabilities, closely followed by automated detection and more advanced capabilities. We note that machine learning, highly touted in 2019 is presently the least-required usability feature for data preparation.

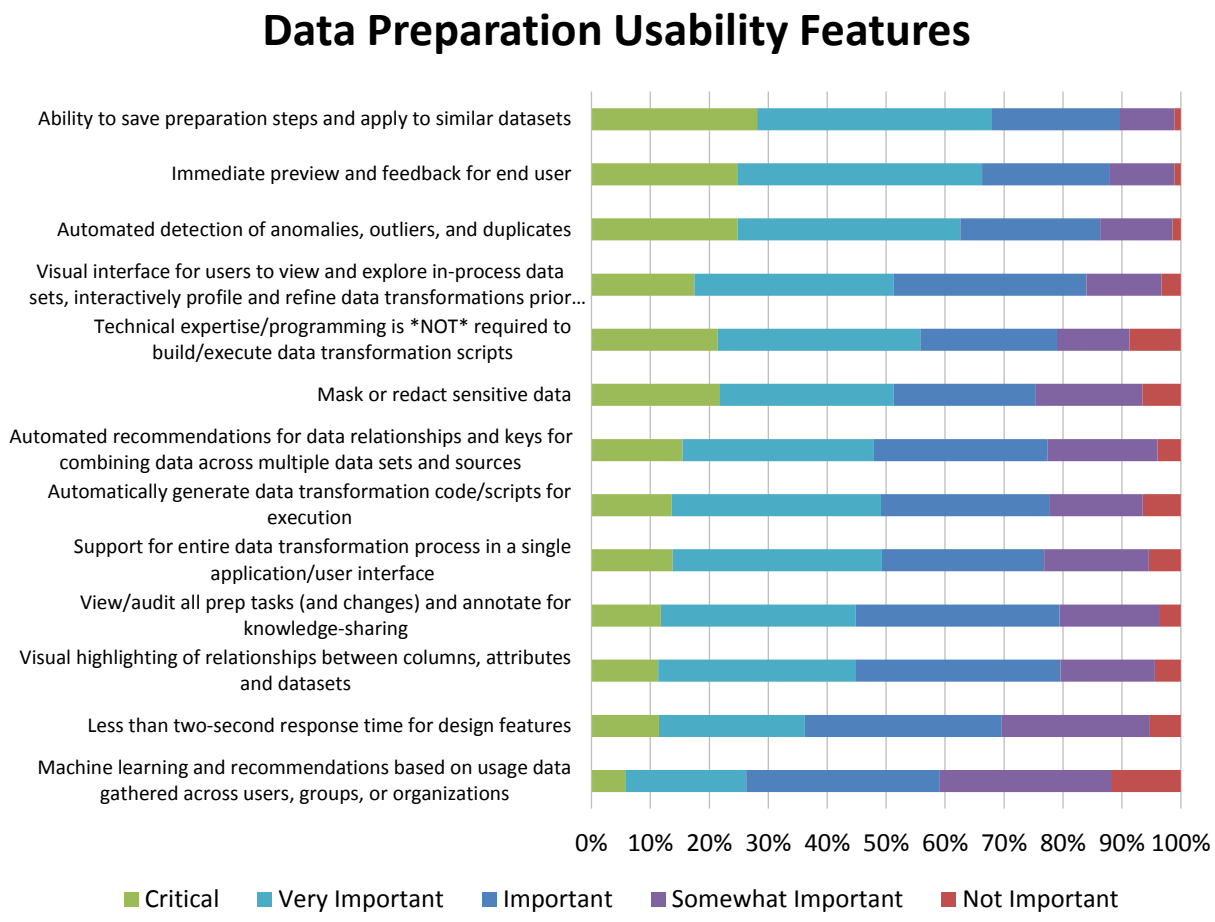


Figure 31 – Data preparation usability features

Across five years of study, attitudes toward data preparation features have fluctuated somewhat by rank (fig. 32). In 2019, "immediate preview and feedback" retains the top ranking, while "visual interface" falls behind "automated detection of anomalies" as a feature requirement. Many usability features decline in importance year over year, though all but "machine learning" post weighted mean scores of 3.0 to 3.8, between "important" and "very important."

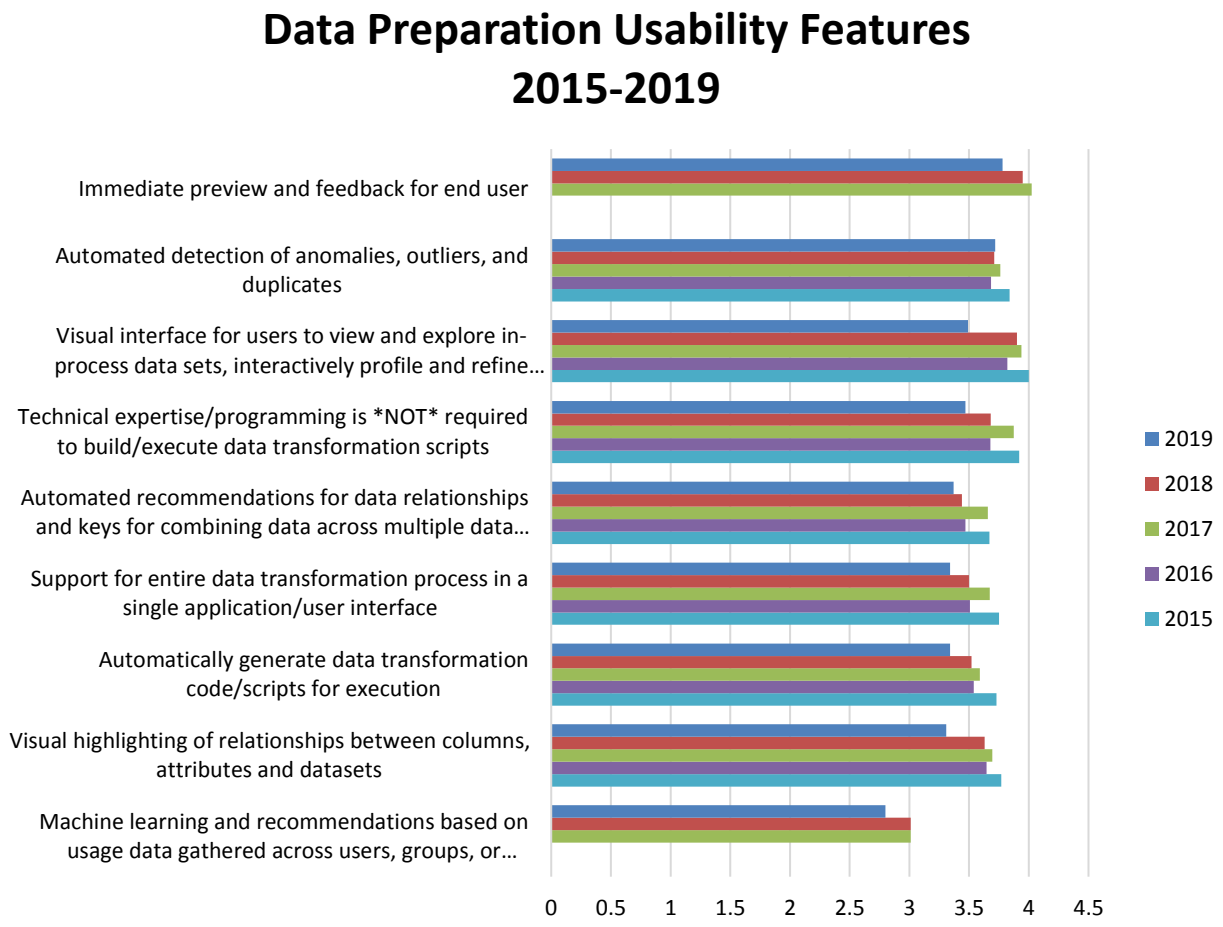


Figure 32 – Data preparation usability features 2015-2019

By function, peak interest in usability features varies rather widely by function (fig. 33). Respondents in Finance report the most interest in "immediate preview and feedback," "support for entire data transformation process in a single application," and "technical expertise not required," along with several lower priorities that include "view/audit all prep tasks," "visual highlighting of relationships," and "less than two-second response time." BICC respondents focus on "automated recommendations for data relationships" and "automatically generate data transformation." R&D respondents are most interested in "ability to save preparation steps and apply to similar databases" and "mark or redact sensitive data."

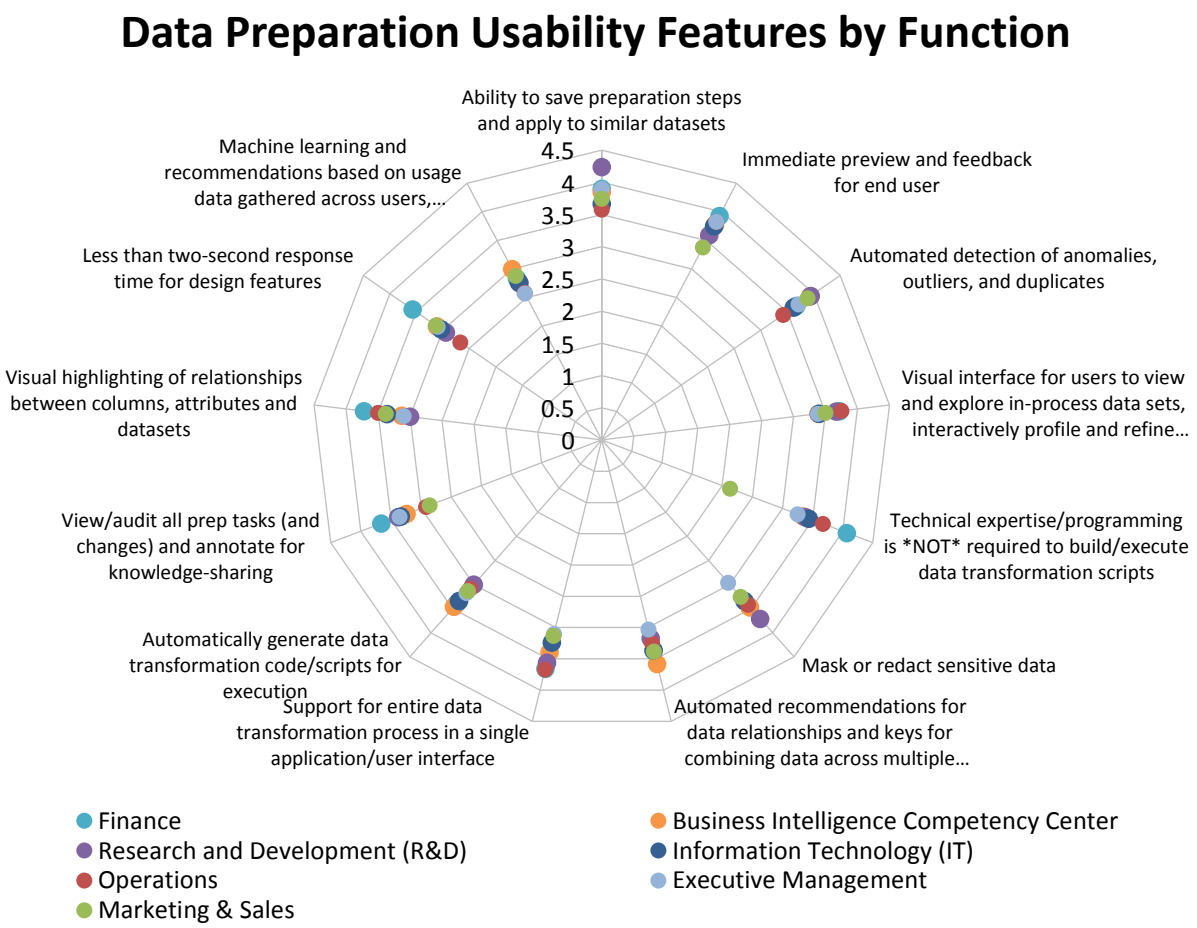


Figure 33 – Data preparation usability features by function

Interest in data preparation features varies by geographical region and is most often led by respondents in Asia Pacific and North America (fig. 34). The strongest comparative interest among Asia-Pacific respondents is in "automated recommendations for data relationships," "automated detection of anomalies," "visual interface," and "technical expertise not required." EMEA interest in usability features is lowest in nearly every case.

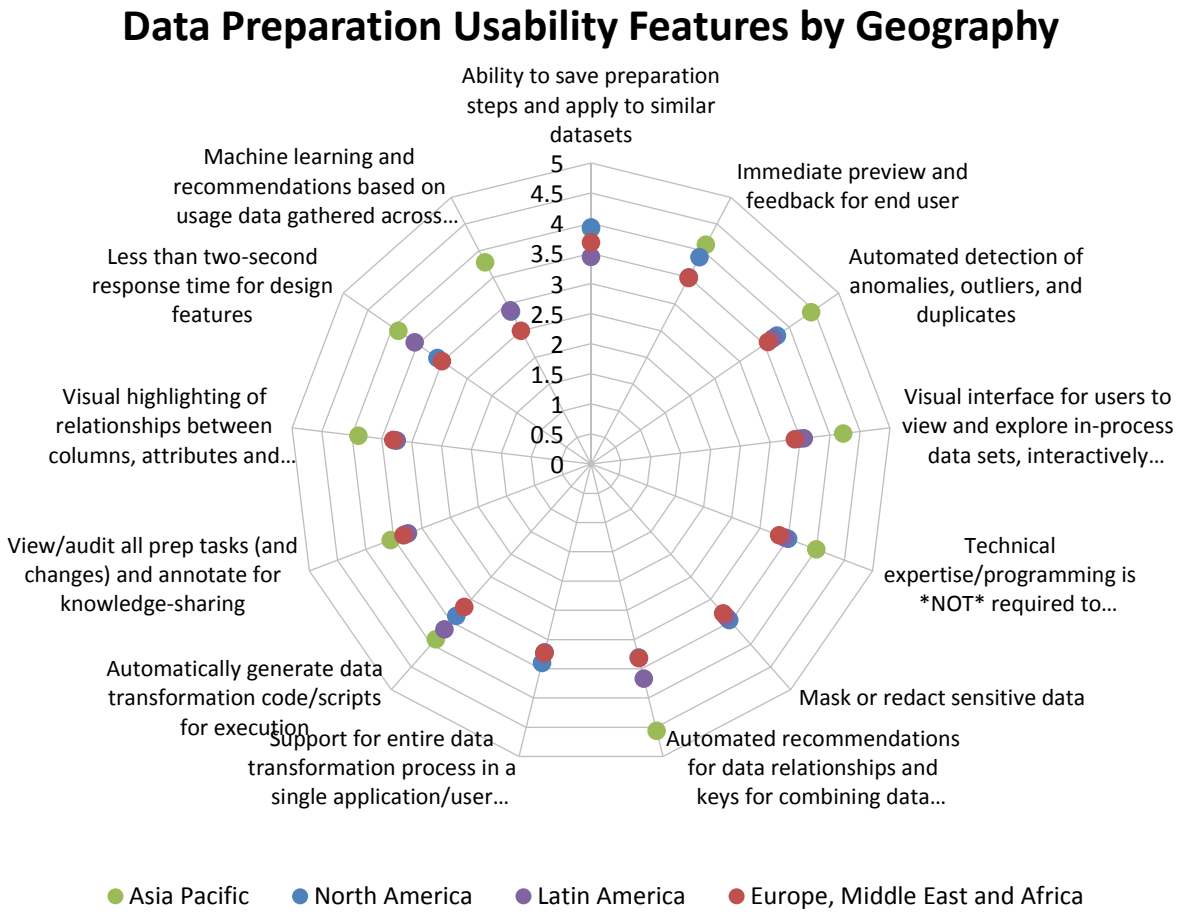


Figure 34 – Data preparation usability features by geography

Compared to other measures, interest in data preparation usability features clusters across organizations of different sizes (fig. 35). Interest in usability features most often increases with organization headcount. Very large organizations (>5,000 employees) lead interest in every usability feature category we sample. Surprisingly, small organizations (1-100 employees) report somewhat lower interest in usability features. Mid-sized organizations (101-1,000 employees) and large organizations with 1,001 to 5,000 employees mostly juggle middling interest across different usability features.

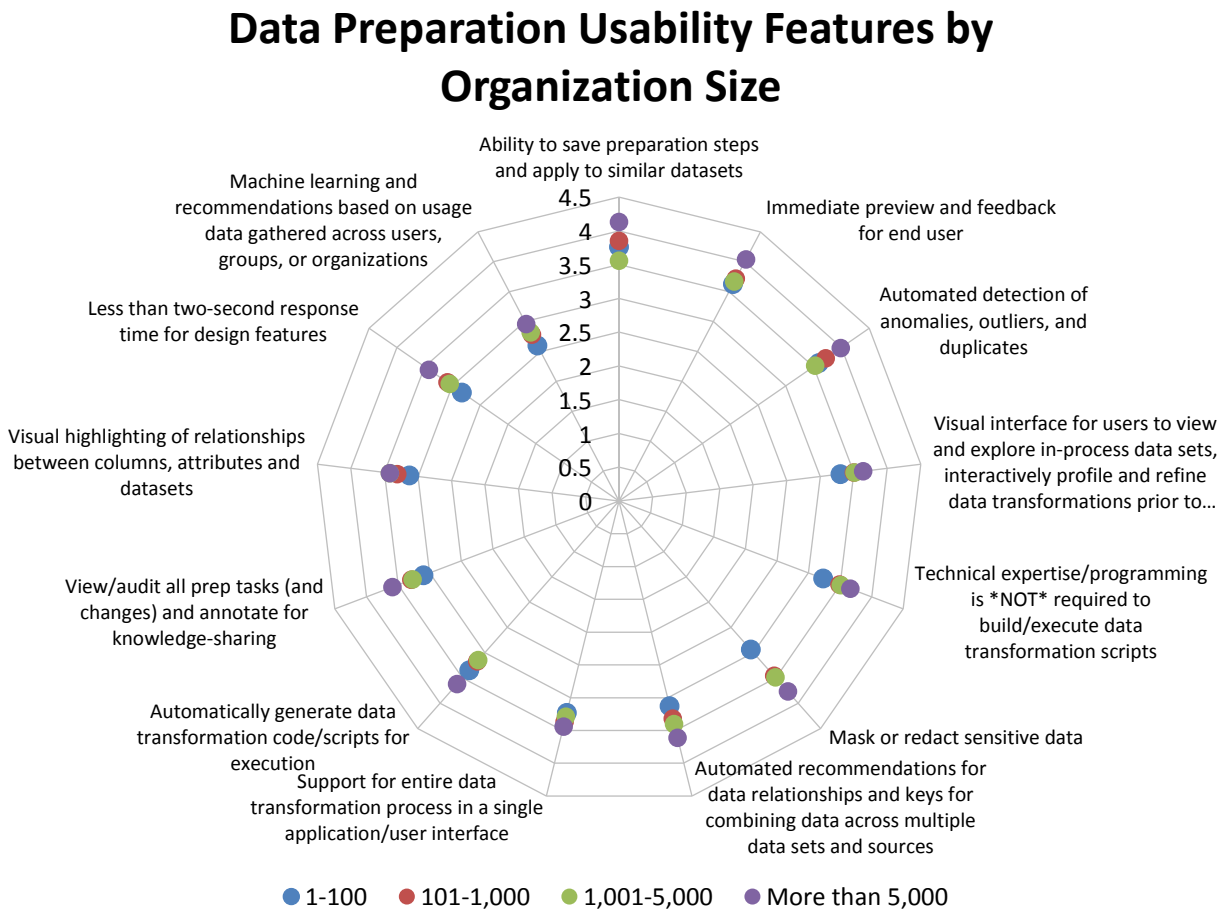


Figure 35 – Data preparation usability features by organization size

Compared to other measures, interest in data preparation usability features varies most dramatically by industry (fig. 36). For example, Government respondents most highly rank "immediate preview and feedback" and "visual interface" but show the least interest in the top feature, "ability to save preparation steps." Education respondents show the most interest in five usability areas that include "automated detection of anomalies" and "mask or redact sensitive data." Respondents in Retail/Wholesale lead interest in three features, including "automated recommendations for data relationships" and "support for entire data transformation process in a single application."

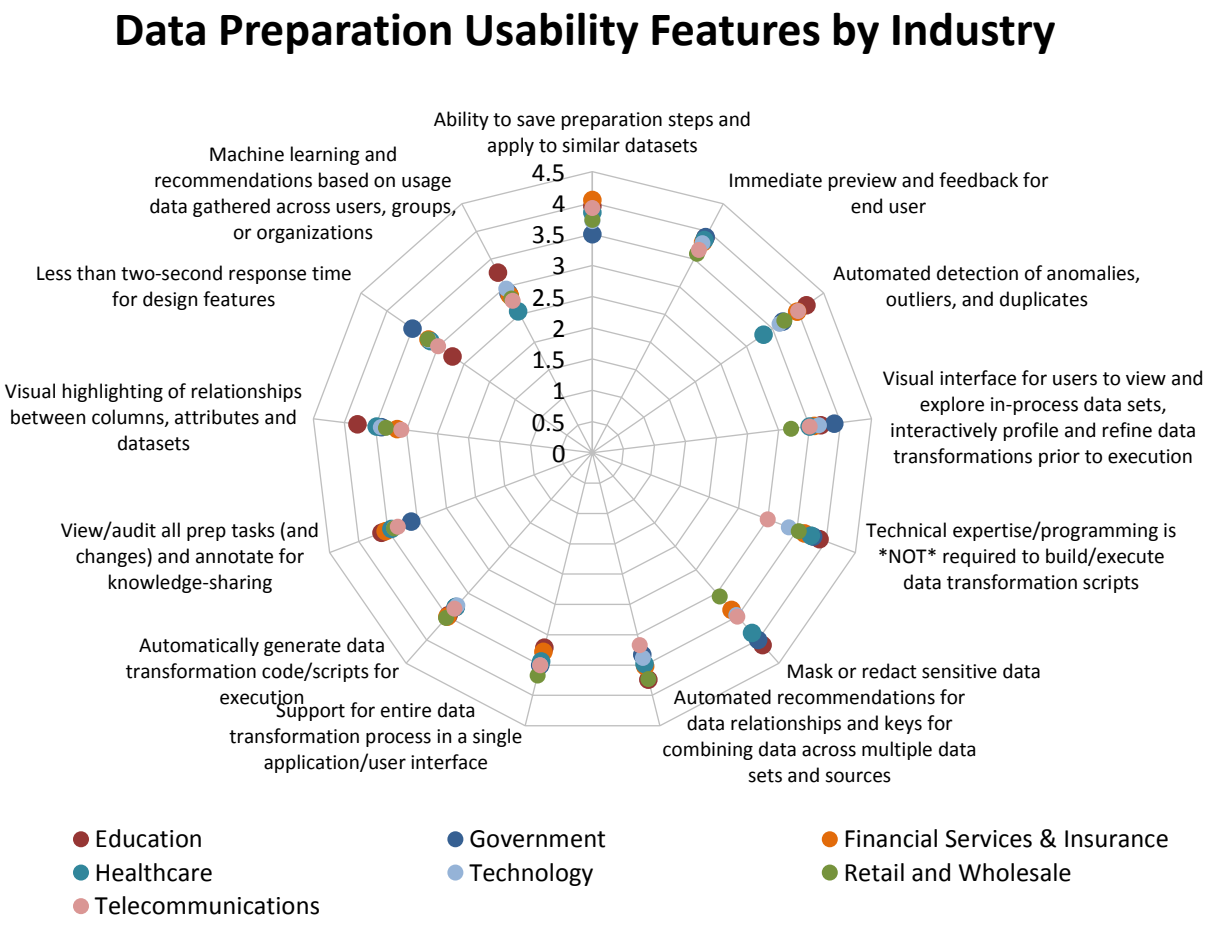


Figure 36 – Data preparation usability features by industry

Data Preparation Data Integration Features

Though not as pronounced as usability, demand for data preparation integration features is nonetheless strong in 2019 (fig. 37). The top two features, "access to file formats" and "ability to combine data across multiple data sets and sources," are "critical" to between 45-50 percent of respondents and at least "very important" to nearly 80 percent or more. "Access to traditional databases" is the third most popular feature, at least "very important" to nearly 70 percent of respondents. Notably, "access to big data" remains the least important data preparation integration feature among respondents.

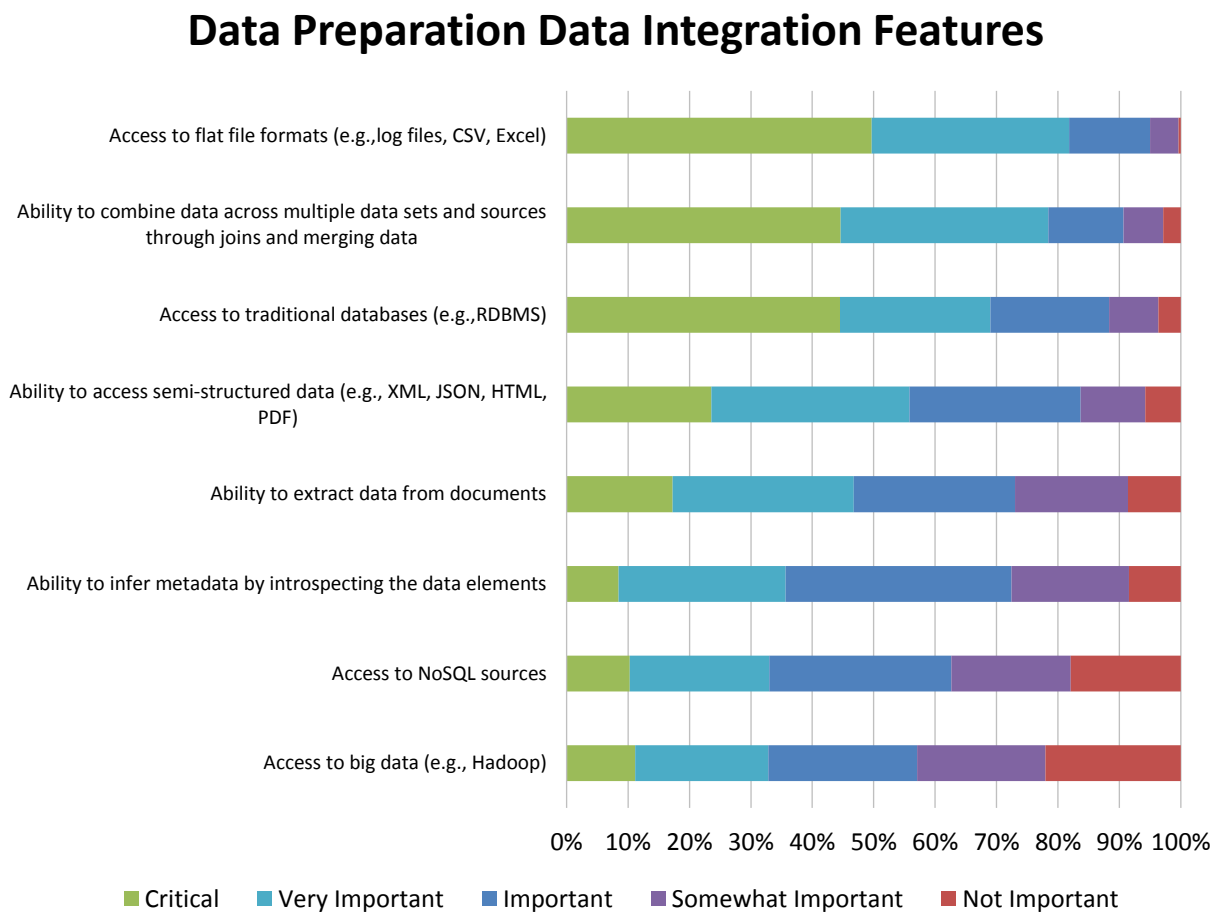


Figure 37 – Data preparation data integration features

Peak interest in the top three data preparation features is mostly sustained in the last three years of our five years of focused study (fig. 38). Interest in the top three features ("access to flat files," “ability to combine data across multiple data sets," and "access to traditional databases”) declines slightly year over year but remains at the "very important" level of 4.0 or higher. Interest in NoSQL and big data integration retreats to below criticality of 3.0, or a bit less than "important" to respondents.

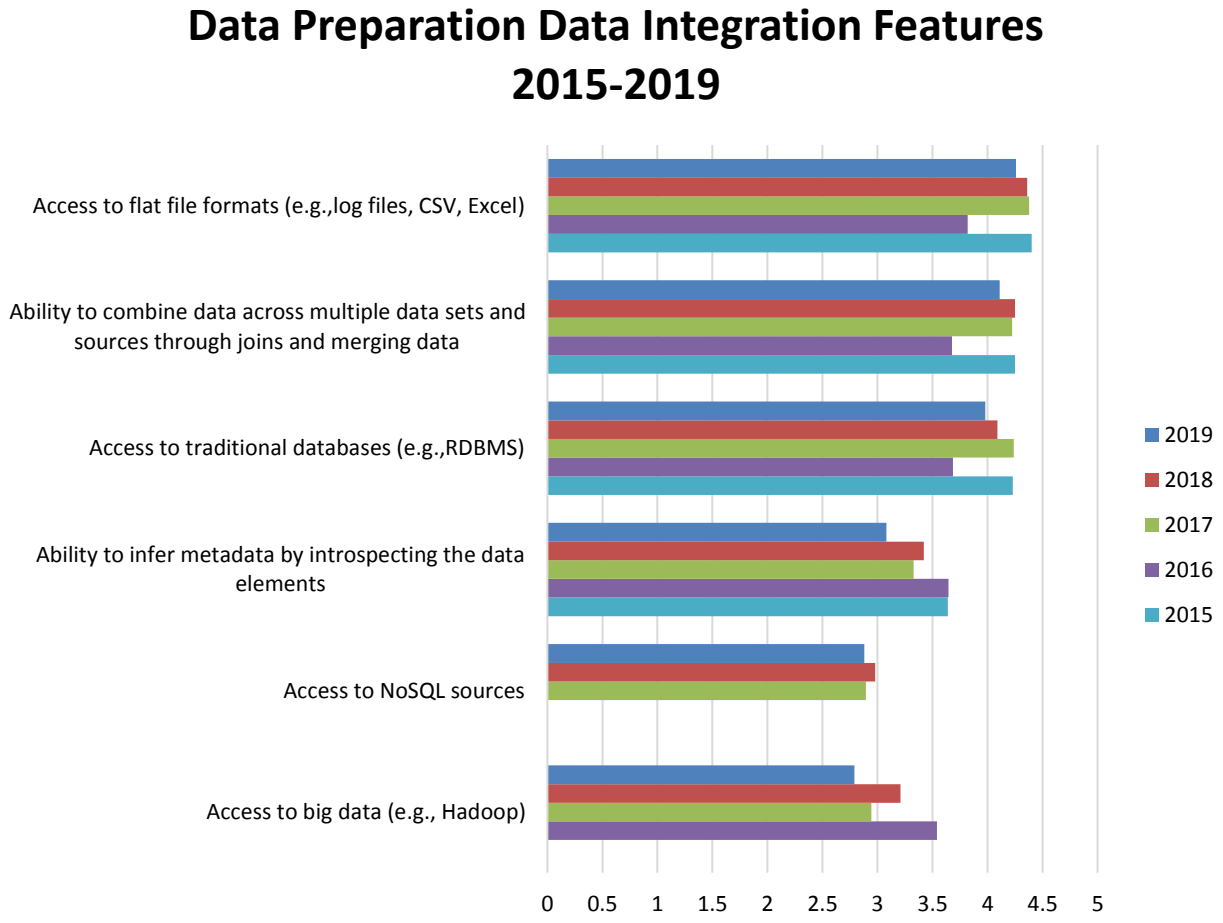


Figure 38 – Data preparation data integration features 2015-2019

Viewed by function, the highest interest in data preparation data integration features most often falls to Operations (fig. 39). Operational areas of peak interest include traditional and process-oriented flat file access, access to traditional databases and the ability to extract data from documents. In 2019 however, Operations is tellingly least interested in access to NoSQL or big data and less interested by comparison in "ability to combine data across multiple data sets." Integration features that do not fall to Operations are most likely to be addressed most often by R&D, BICC, or IT functions.

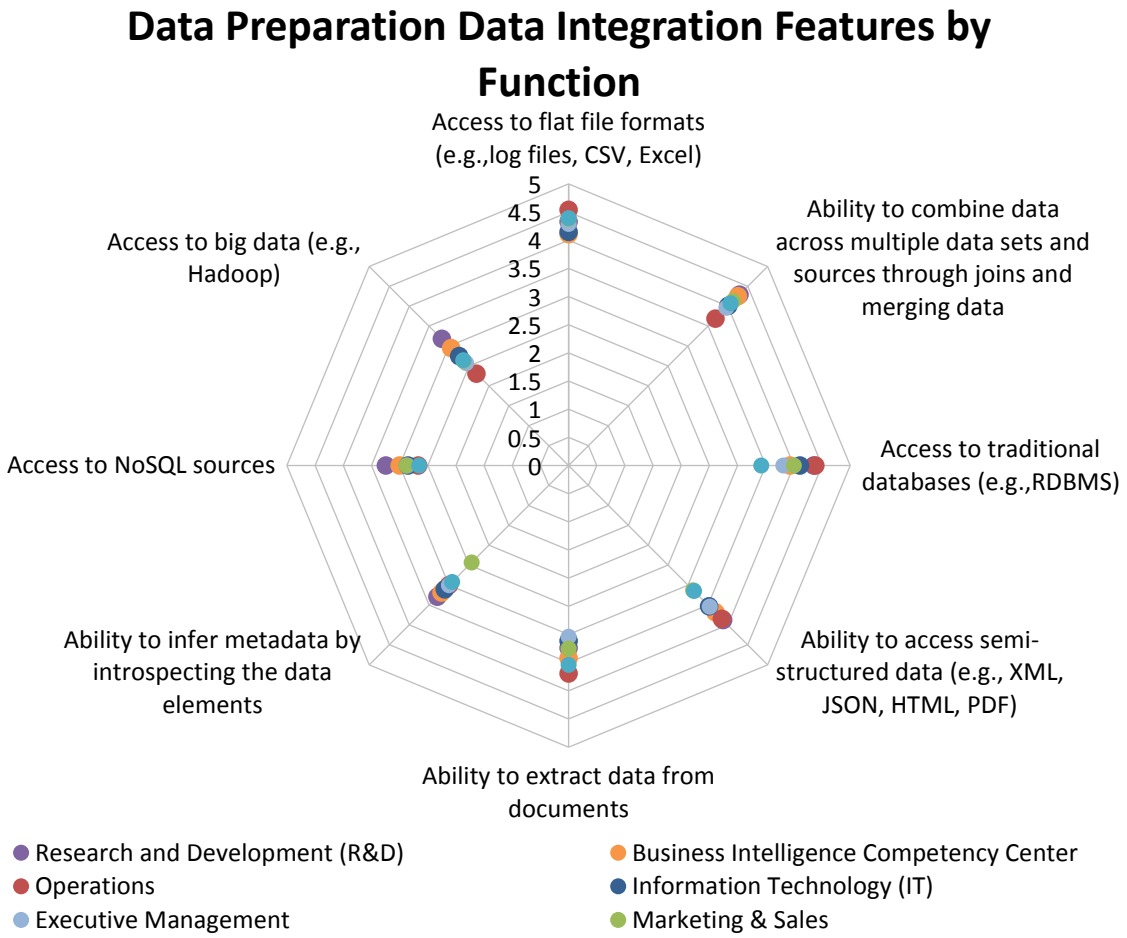


Figure 39 – Data preparation data integration features by function

By region, Asia-Pacific respondents report the highest interest in "ability to combine data across multiple datasets," “ability to access semi-structured data,” "ability to infer metadata," and "access to NoSQL" (fig. 40). North American respondents are slightly most interested in "access to flat file formats." EMEA respondents have the highest interest in "access to traditional databases" but are more circumspect regarding other features. Amid clustered interest, Latin American respondents are slightly most likely interested in "ability to extract data from documents."

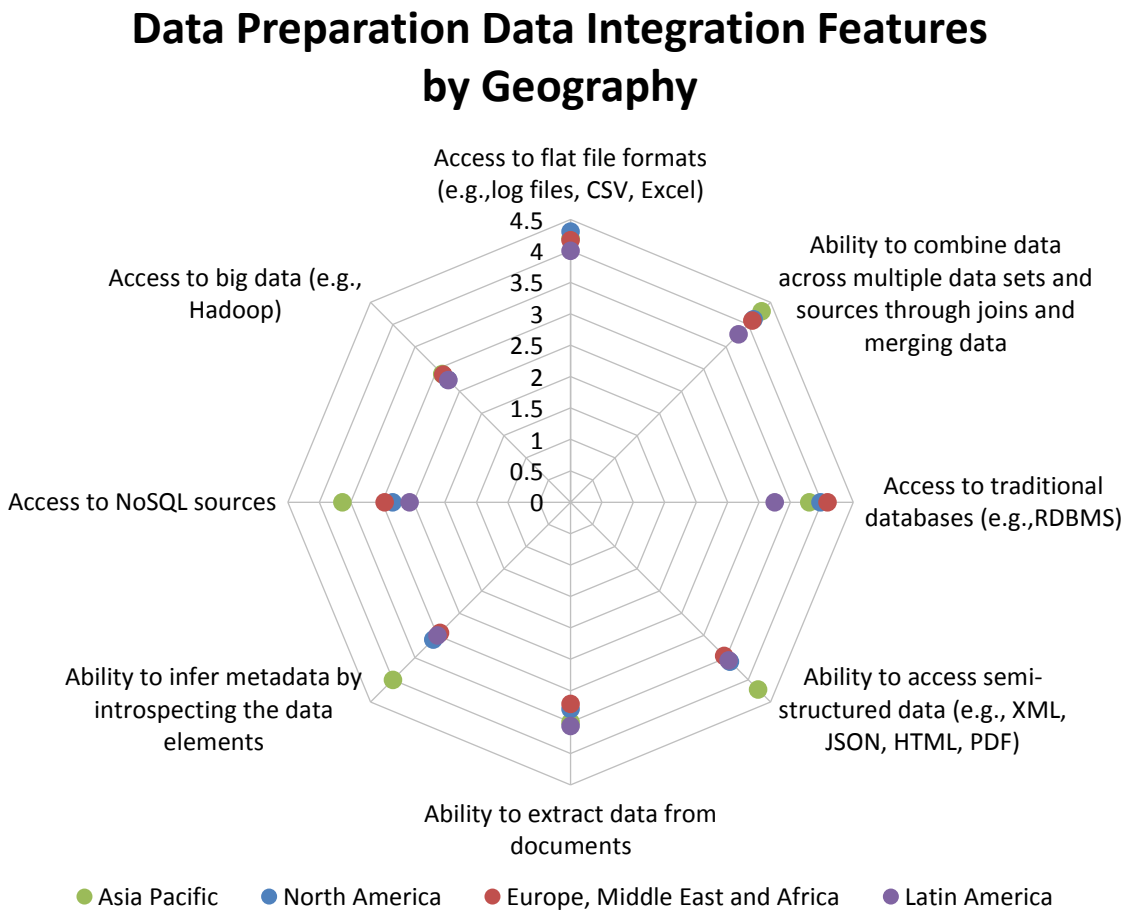


Figure 40 – Data preparation data integration features by geography

Interest in data preparation integration features often tightly clusters across organizations of different size and generally increases, though not always, with organization size (fig. 41). Very large organizations lead interest in varying degree toward all but one integration feature (access to traditional databases, led by large organization respondents). Interest in big data is lowest at small organizations and most clearly grows with global headcount.

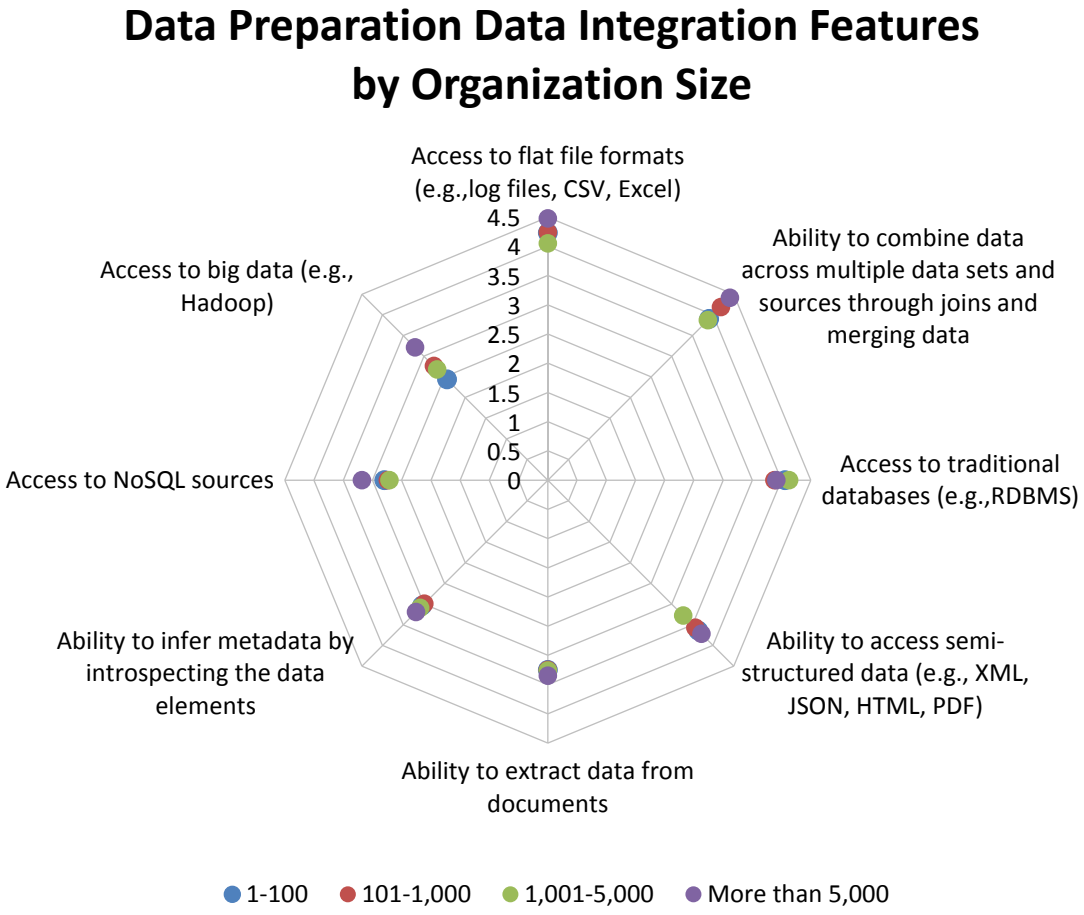


Figure 41 – Data preparation data integration features by organization size

Interest in data preparation integration features varies across industries in 2019 (fig. 42). Almost every industry we sample reports its own unique top feature interest. Perhaps most notably, Telecommunications respondents report the highest levels of interest in newer NoSQL and big data approaches to integration. Education, Retail/Wholesale, and Technology comparably have the least interest in these areas.

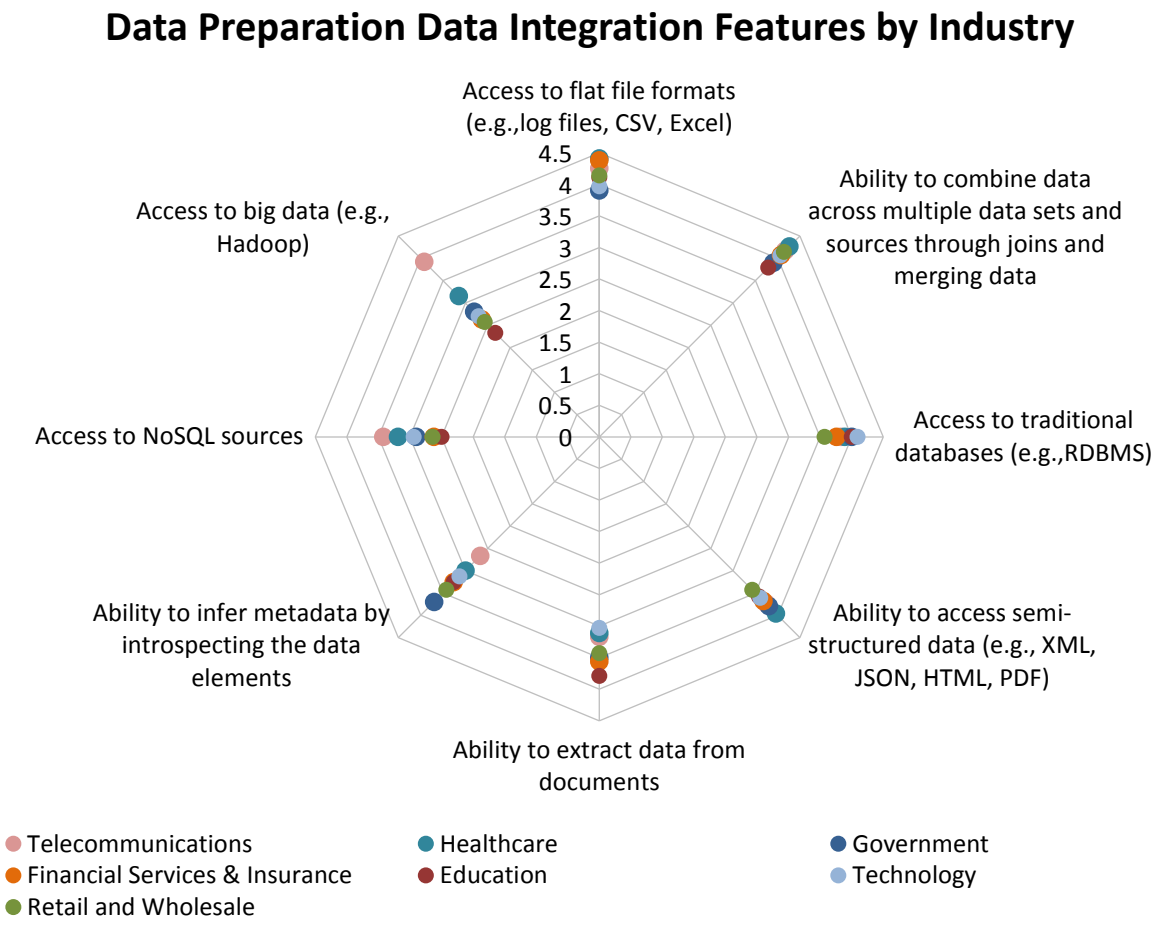


Figure 42 – Data preparation data integration features by industry

Data Preparation Manipulation Features

We asked organizations to score their interest in specific data-manipulation features and once again find a very high and broad level of interest (fig. 43). All but the two lowest-ranked features ("session-ize log or event data" and "window and time-series functions") are at least "very important" to 50 percent or far more respondents. The top two features, "ability to aggregate and group data" and "ability to pivot data," are most often "critical" to respondents and reflect classic spreadsheet operations. Many other features, including "ability to normalize, standardize and enrich data" and "ability to derive new data features from existing data," are likewise attuned to improving common data manipulations.

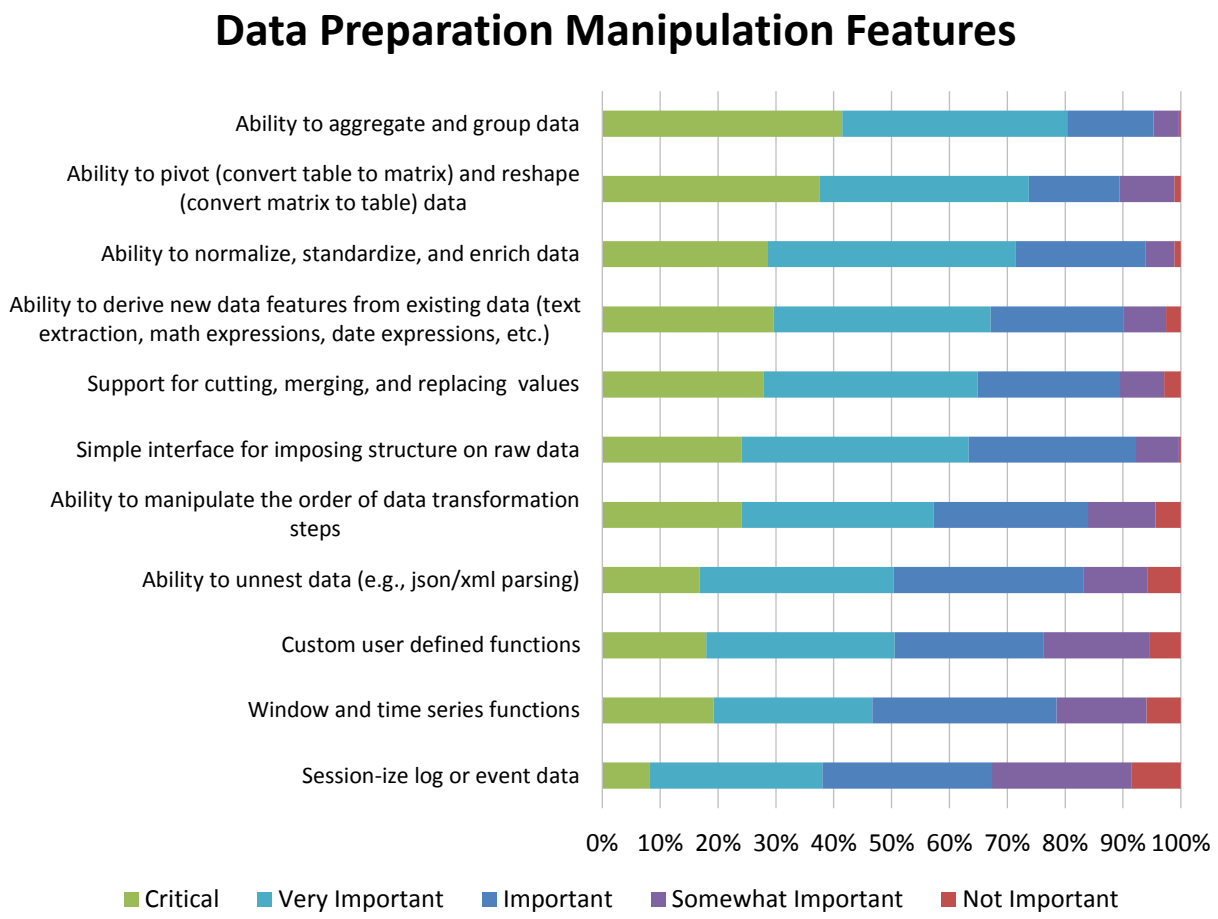


Figure 43 – Data preparation manipulation features

Year-over-year interest in data preparation manipulation features falls at least slightly across all measures except two ("ability to pivot" and "ability to unnest data") in 2019 (fig. 44). Priority rankings over the last 12 months are nearly unchanged. We characterize demand for manipulation features as generally steady across the five years of our report.

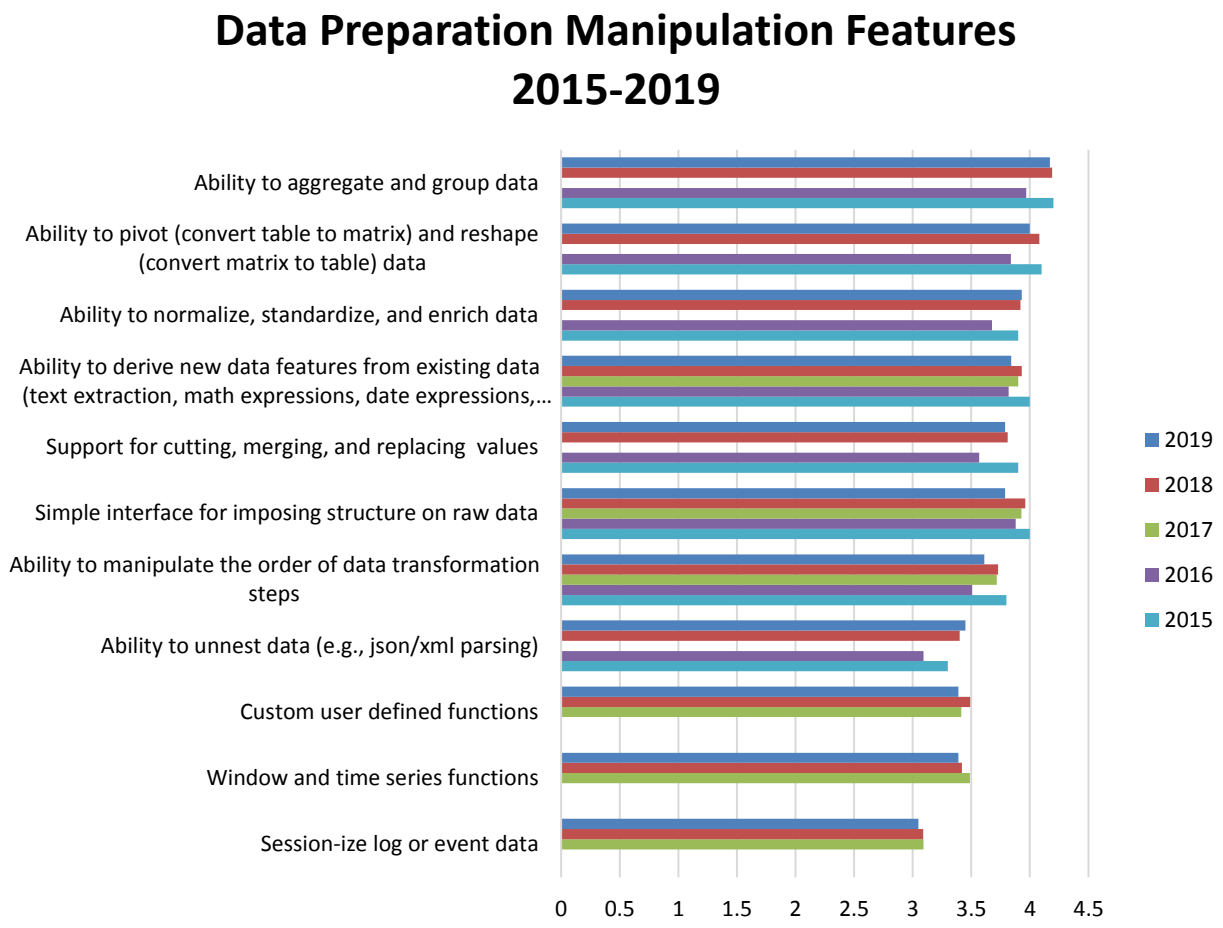


Figure 44 – Data preparation manipulation features 2015-2019

Interest in data manipulation features for data preparation varies rather broadly by function (fig. 45). Marketing/Sales stands out as being most interested in six of 11 functions we sample, including the top three. Operations and BICC respondents have the most interest in "ability to derive new data features from existing data sets." Finance reports the most interest in "simple interface for imposing structure on raw data."

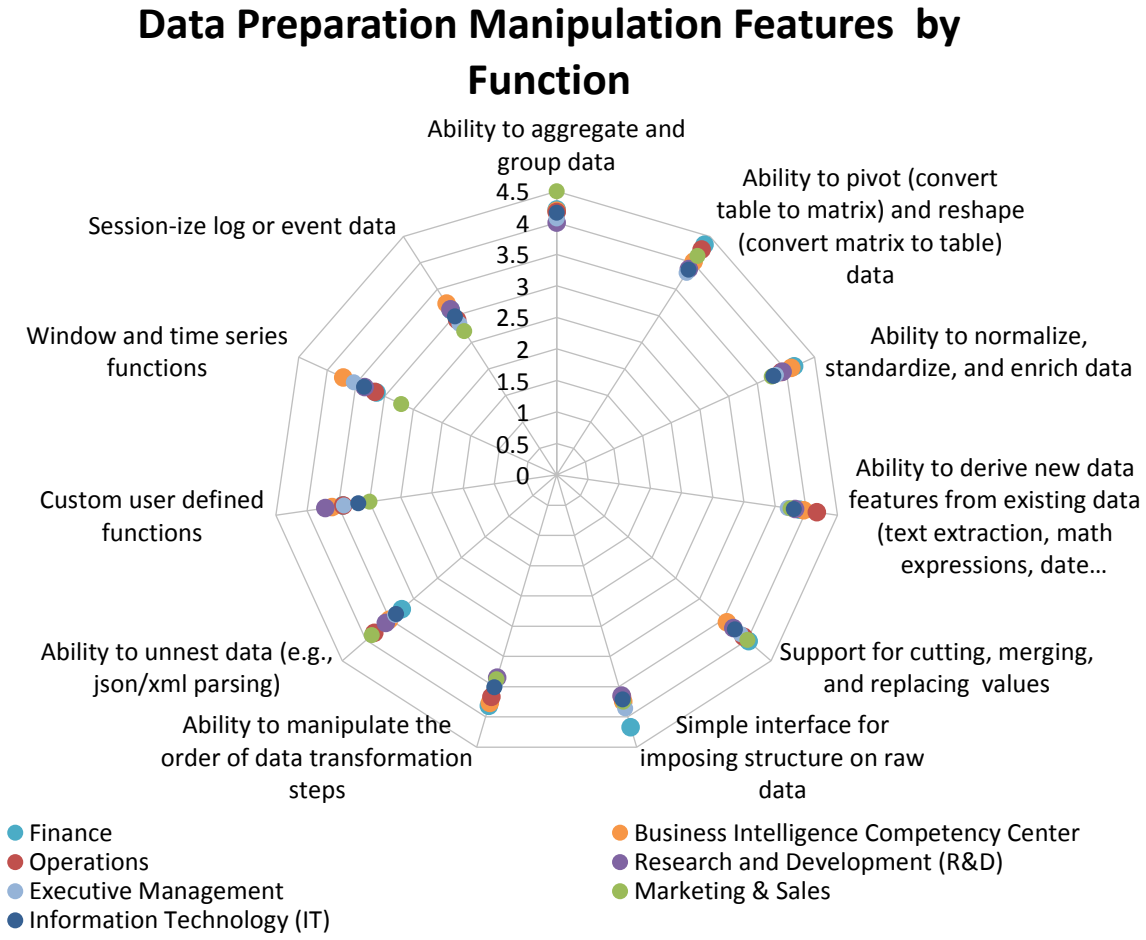


Figure 45 – Data preparation manipulation features by function

In nearly every case, Asia-Pacific respondents lead interest in data preparation manipulation features (fig. 46). In many cases, Asia-Pacific scores approach levels of 4.5, or greater than "very important." North American respondents are generally the second most interested in manipulation features by geographic region, often closely trailed by EMEA respondents.

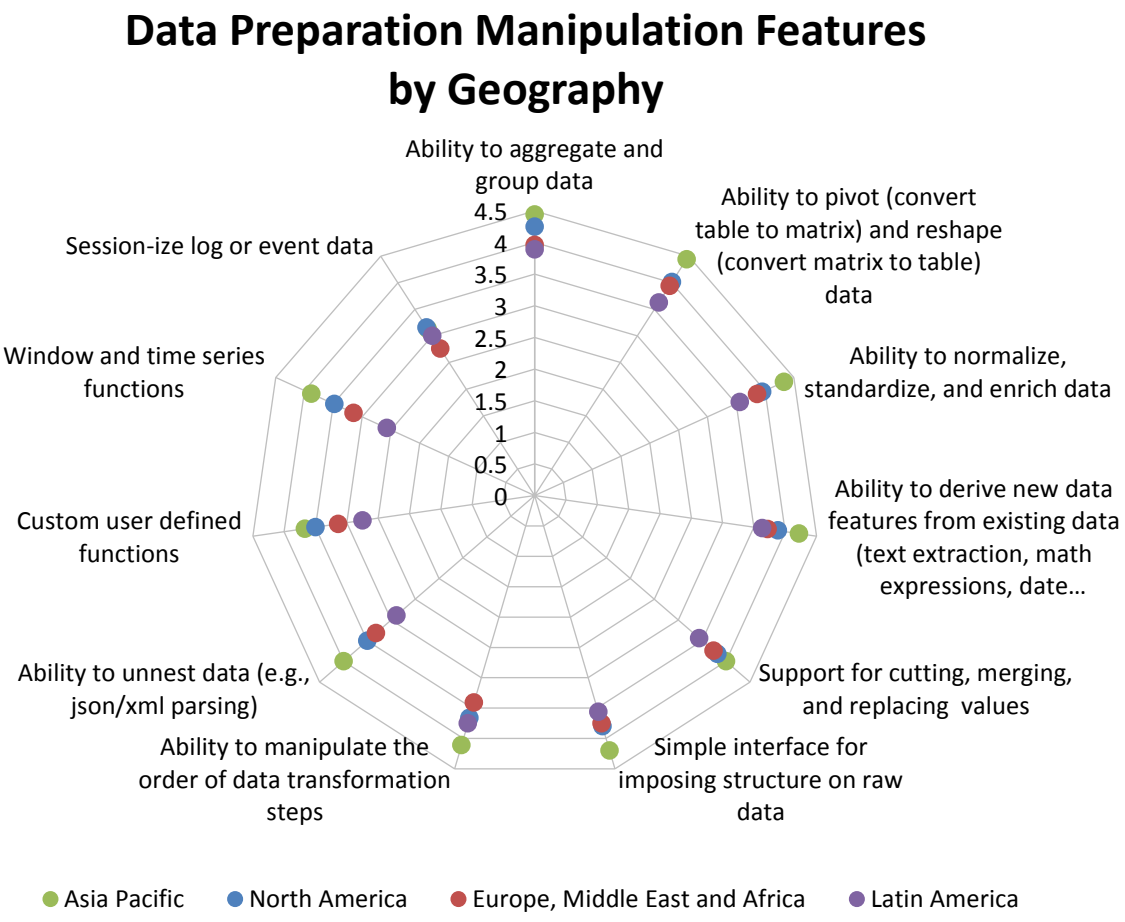


Figure 46 – Data preparation manipulation features by geography

Amid results that often cluster, very large organizations (> 5,000 employees) lead overall interest in data preparation manipulation features (fig. 47). Generally, interest in these features declines with organization size, though interest in large organizations (1,001-5,000 employees) often falls below mean values, especially for lower-ranked manipulation features.

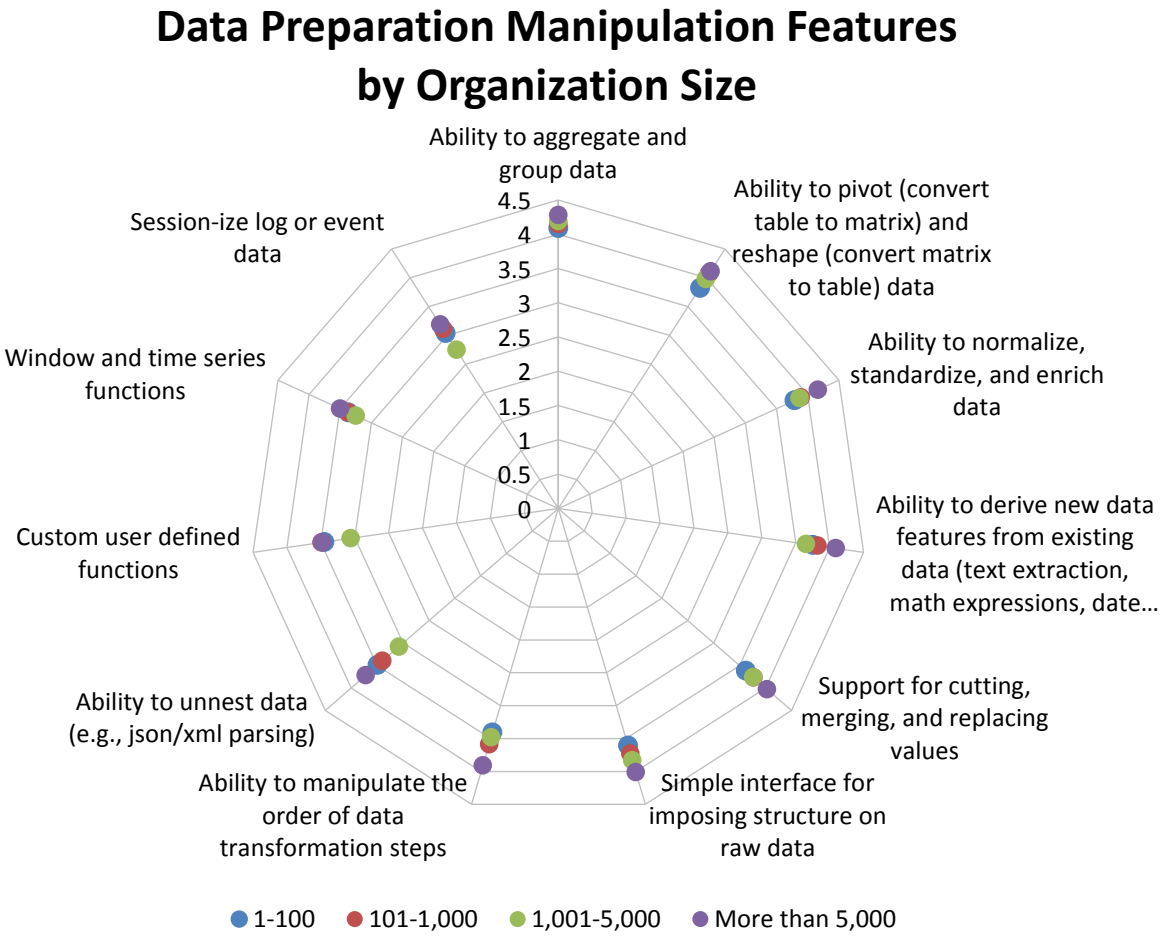


Figure 47 – Data preparation manipulation features by organization size

Interest in data preparation manipulation features varies by industry, though many feature scores cluster (fig. 48). In 2019, the top feature, "ability to aggregate and group data," shares similar high interest in all industries except Government. Retail/Wholesale reports very high interest in "ability to pivot" and gives the high score for "simple interface for imposing structure on raw data." Financial Services and Technology respondents share the most interest in "ability to derive new data from existing data."

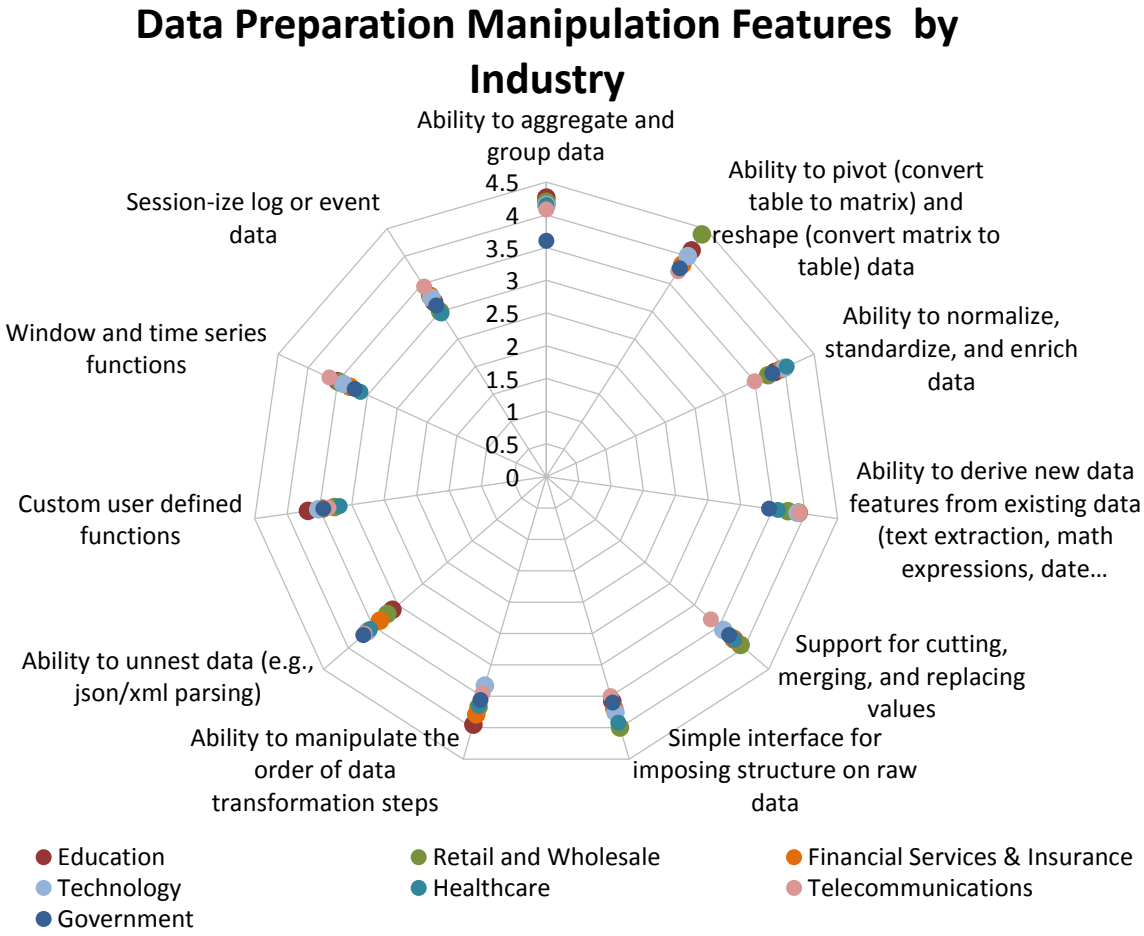


Figure 48 – Data preparation manipulation features by industry

Data Preparation Supported Outputs

Respondents say the most important data prep output is to Excel and CSV, followed by traditional relational databases (71 percent) (fig. 49). Somewhat distantly behind these long-held preferences, JSON (44 percent) and popular BI tools are the next most required outputs for support. Various popular big data and service-based outputs are by comparison relatively unimportant compared to flat files in 2019. For example, users are more than three times likely to seek flat file outputs than outputs for Azure or Hadoop. (The predominance of Excel, shown over time in the following chart, provides another useful perspective.)

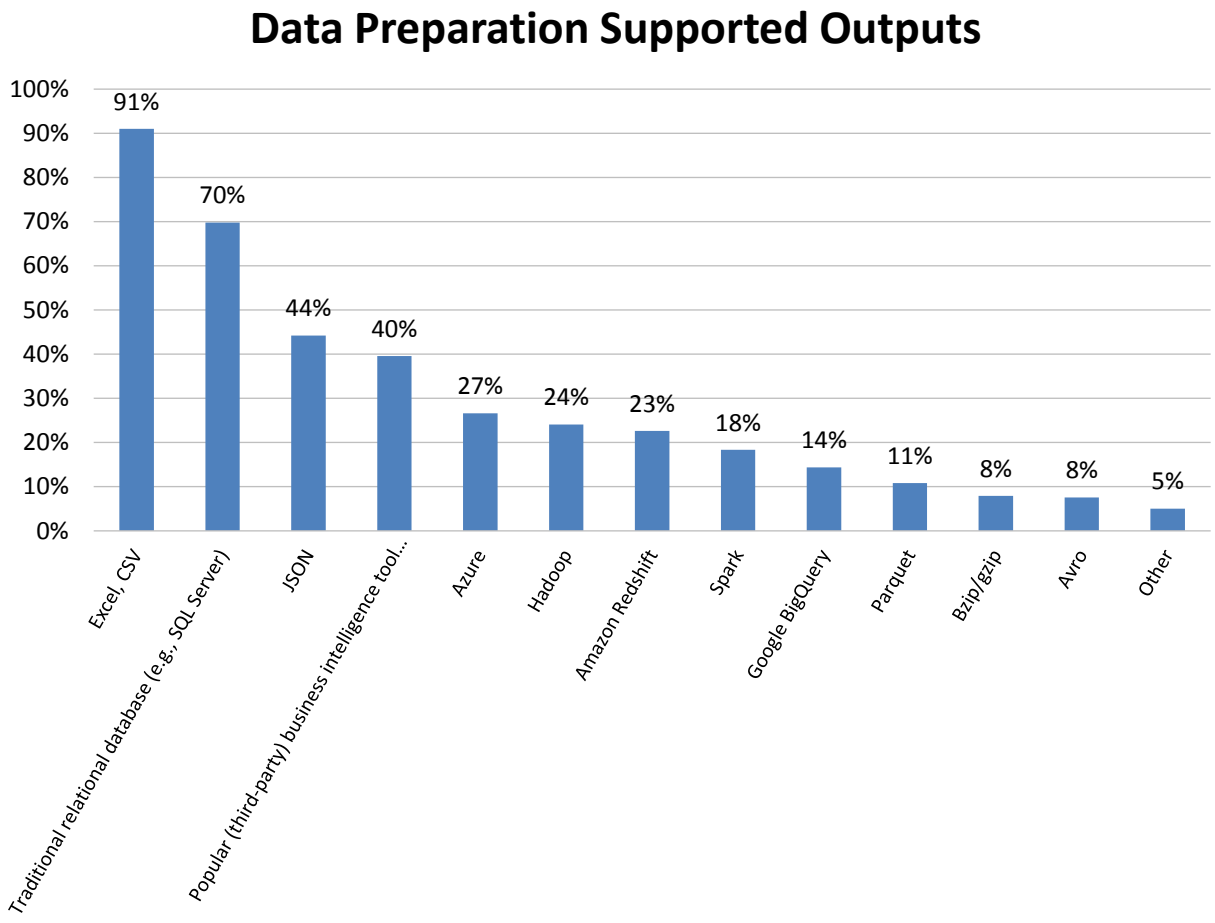


Figure 49 – Data preparation supported outputs

The dominance of Excel, traditional database and other leading output formats (previous chart) continues in 2019 but gathers context in the growth of some competing technologies (fig. 50). Across three years of data, traditional output formats have declined marginally in importance while Microsoft Azure, Amazon Redshift, and Google BigQuery notably gather traction. Demand for Azure in particular nearly double in importance between 2018 and 2019, albeit to a bit less than 30 percent among respondents. The biggest year-over-year decline in importance is reported for Hadoop, which fell from 32 to 24 percent.

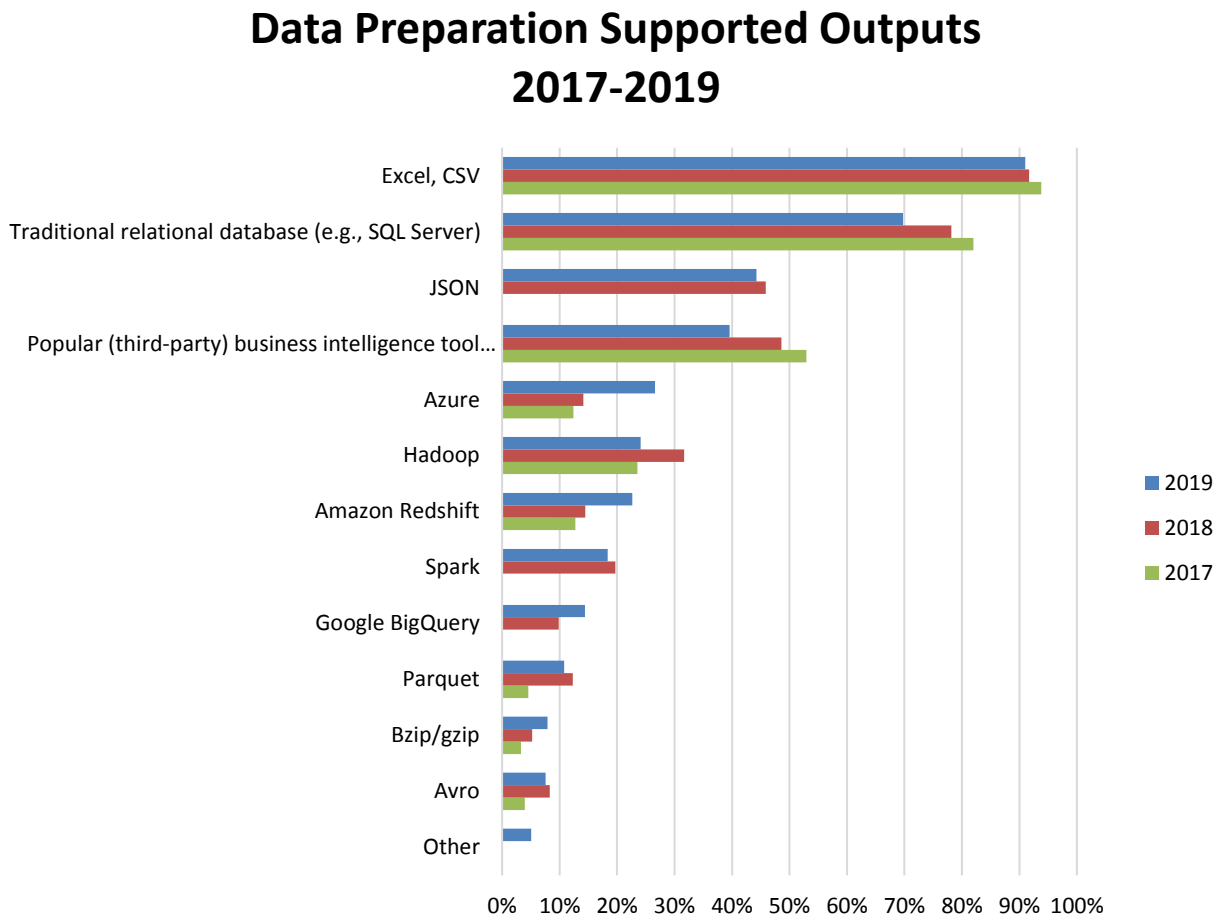


Figure 50 – Data preparation supported outputs 2017-2019

Viewed by function, R&D respondents are most likely to specify the top three supported data preparation output formats (relational database, JSON. and popular BI tool formats) in 2019 (fig. 51). R&D is also most likely to specify developer-oriented Hadoop and Spark outputs. Marketing/Sales respondents most often require data preparation output to third-party services Azure and Google BigQuery.

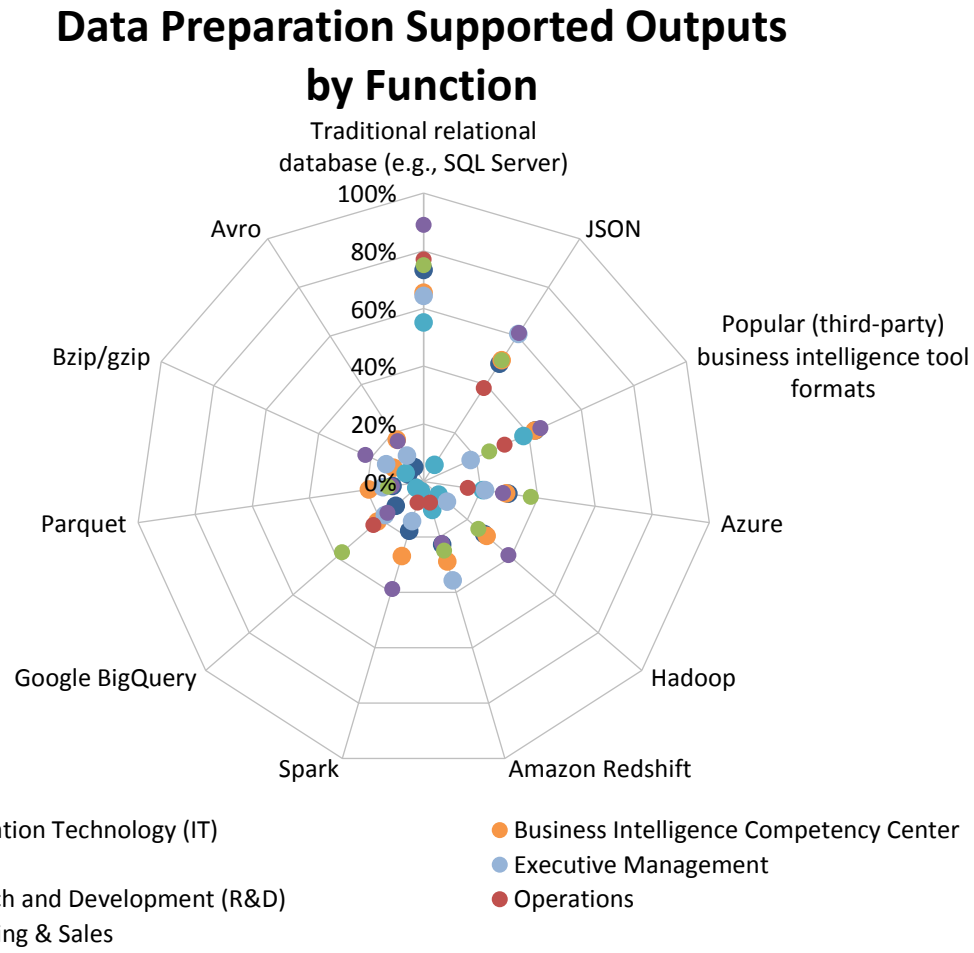


Figure 51 – Data preparation supported outputs by function

Preference for flat file Excel and CSV outputs for data prep is uniformly highest across all geographic regions and highest overall in Latin America, North America, and EMEA (fig. 52). EMEA and North American respondents are most likely by region to require traditional relational database, JSON, and popular third-party business intelligence tools outputs. Interest in Azure Hadoop and Amazon Redshift is above the global average in Latin America and Asia Pacific.

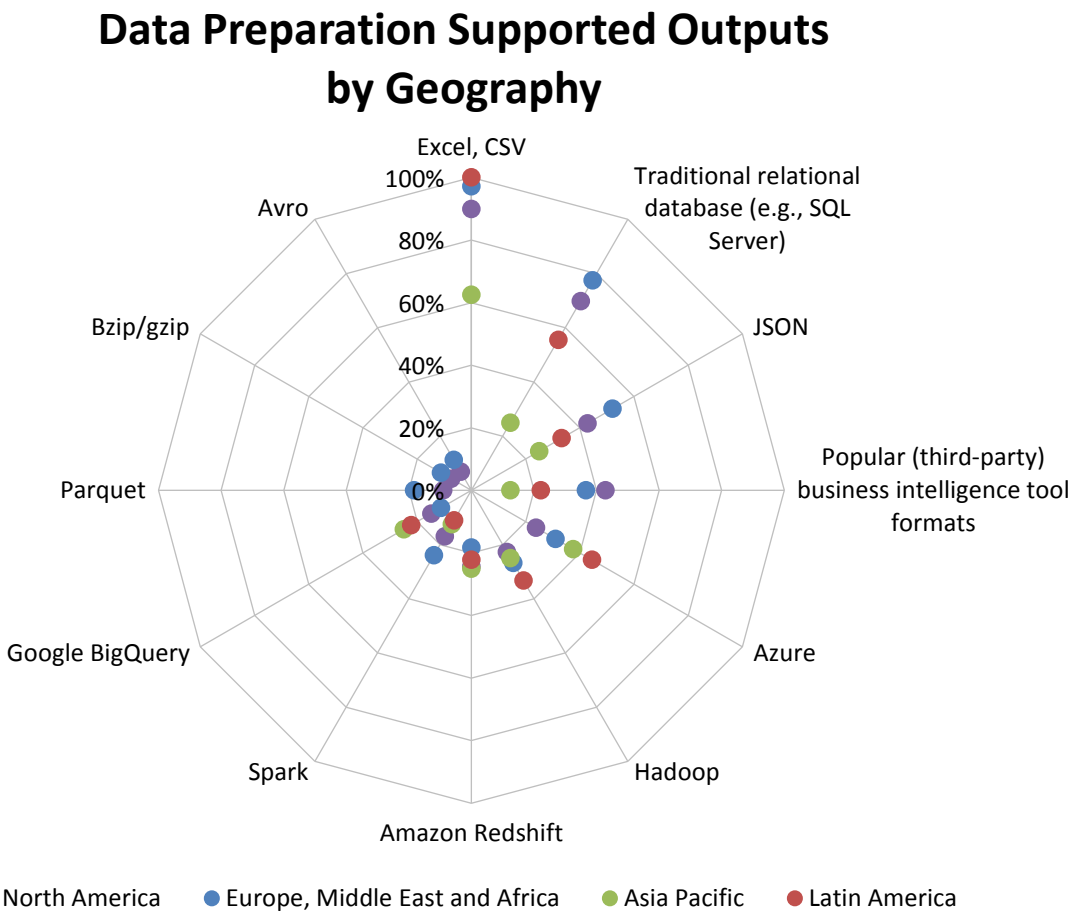


Figure 52 – Data preparation supported outputs by geography

A great majority of organizations of any size (led by large and very large organizations) share the highest preference for flat file output support of data preparation (fig. 53). Again, traditional relational database output support is the strong second choice, led by respondents in large and very large organizations. With the exception of JSON and Amazon Redshift, where small organizations (1-100 employees) show the most interest, very large organizations lead demand for all other formats for data preparation output support.

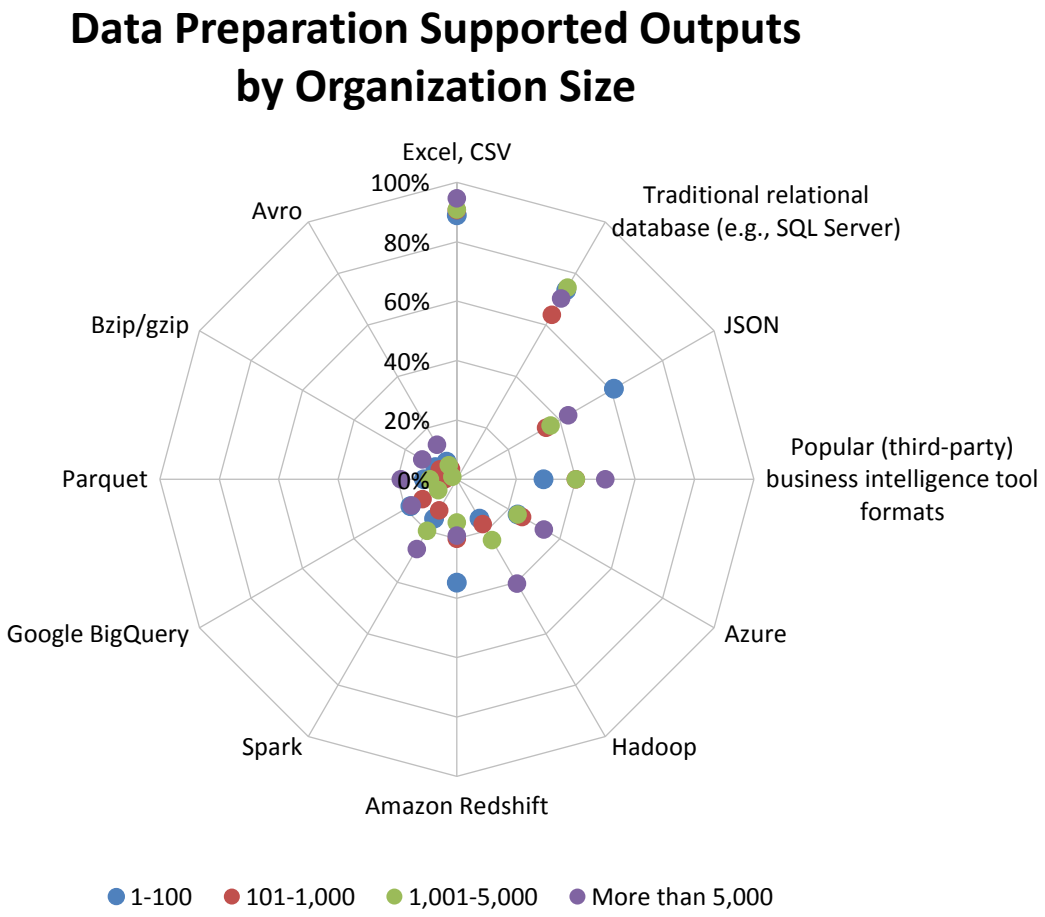


Figure 53 – Data preparation supported outputs by organization size

Most but not all industries we sampled in 2019 share the greatest preference for Excel/CSV output with a second choice of traditional relational database output support (fig. 54). Telecommunications respondents, however, are more likely to specify relational database versus flat file output. Interest in relational databases versus Excel by industry inverts for Retail/Wholesale and Education respondents. Telecommunications and Financial Services slightly more often require JSON output support. Telecommunications respondents report by far the highest interest in Hadoop and Spark, while Technology respondents widely lead interest in Amazon Redshift.

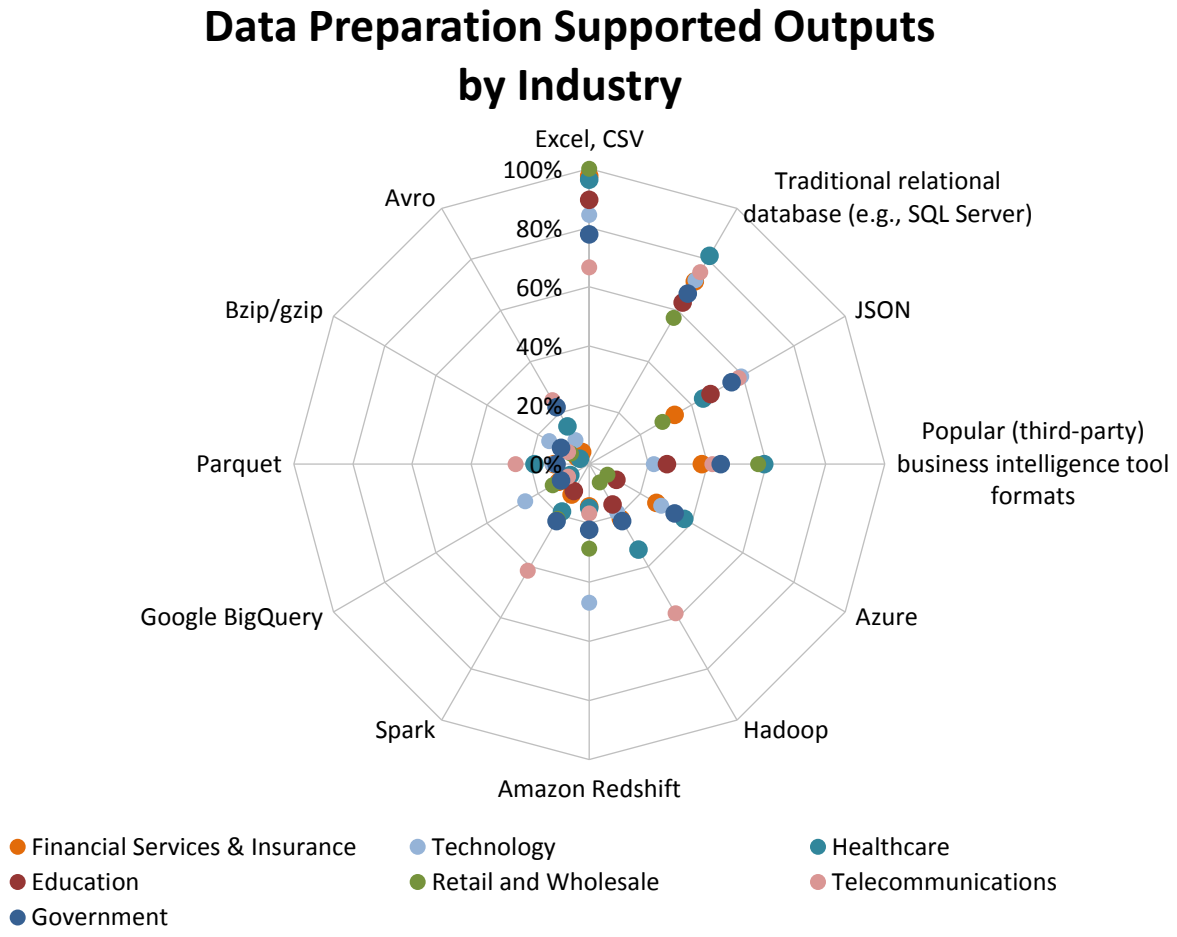


Figure 54 – Data preparation supported outputs by industry

Data preparation Deployment Features

We asked respondents about their preferences for scheduling, monitoring, and testing aspects that make data preparation part of a more formal ongoing process (fig. 55). While such features resonate less with respondents compared to other data preparation capabilities, the two most popular features ("schedule a process to run on a time-based or trigger-based event" and "ability to schedule execution/replay of data transformation processing ") are either "critical" or "very important" to 60 percent or more of respondents. To a similar degree, “ability to alert on anomalies or changes in the structure” is, at minimum, “very important” to 58 percent of respondents.

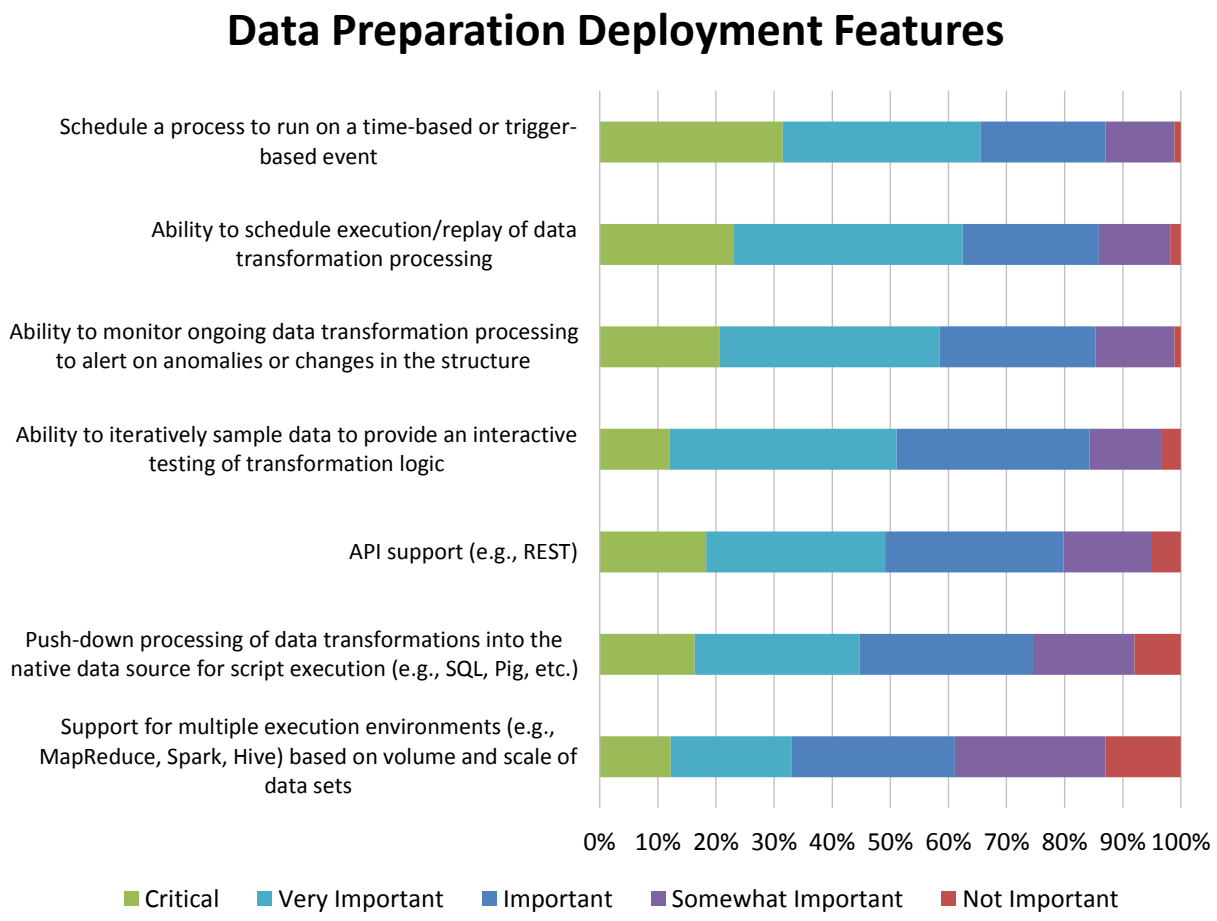


Figure 55 – Data preparation deployment features

In the most recent two to three years of our study, interest in data preparation deployment features holds steady with only minor changes, while importance rankings remain unchanged (fig. 56). (The top feature, “schedule a process” is newly added for 2019.) The importance of the second and third feature choices (“ability to schedule execution/replay” and “ability to monitor ongoing data transformation”) is unchanged year over year, with weighted mean importance above 3.5 or between “important” and “very important.”

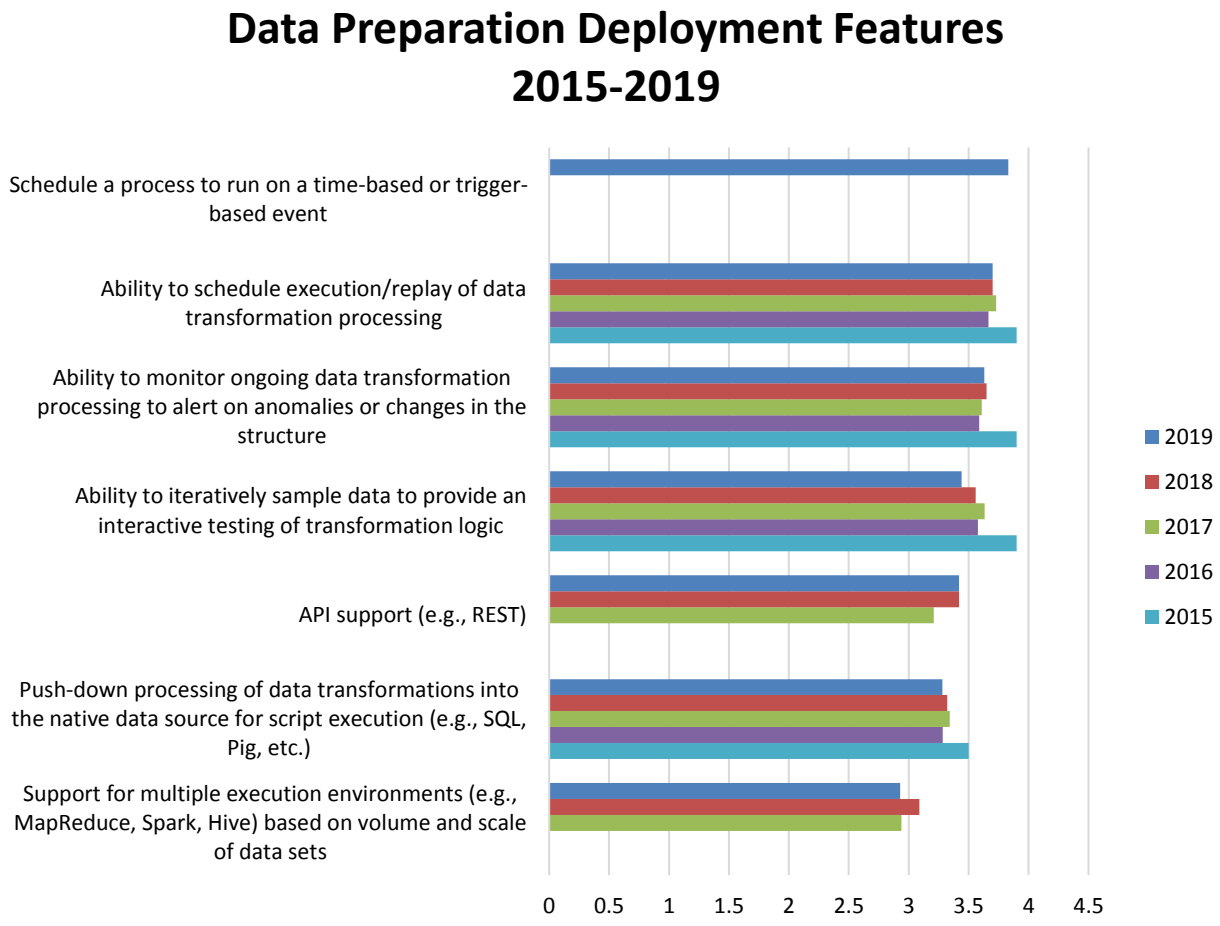


Figure 56 – Data preparation deployment features 2015-2019

Sentiment toward the top three data preparation deployment features is high across functions with mean interest mostly ranging from well above "important" toward "very important" (fig. 57). Interest in the top feature, “schedule a process,” is highest in Operations, which also leads interest in API support. Interest in the next three top choices (“ability to schedule,” “ability to monitor,” and “ability to iteratively sample”) is highest among BICC respondents, as is “push-down processing.” Marketing/Sales and BICC respondents are most likely to seek “support for multiple execution environments.”

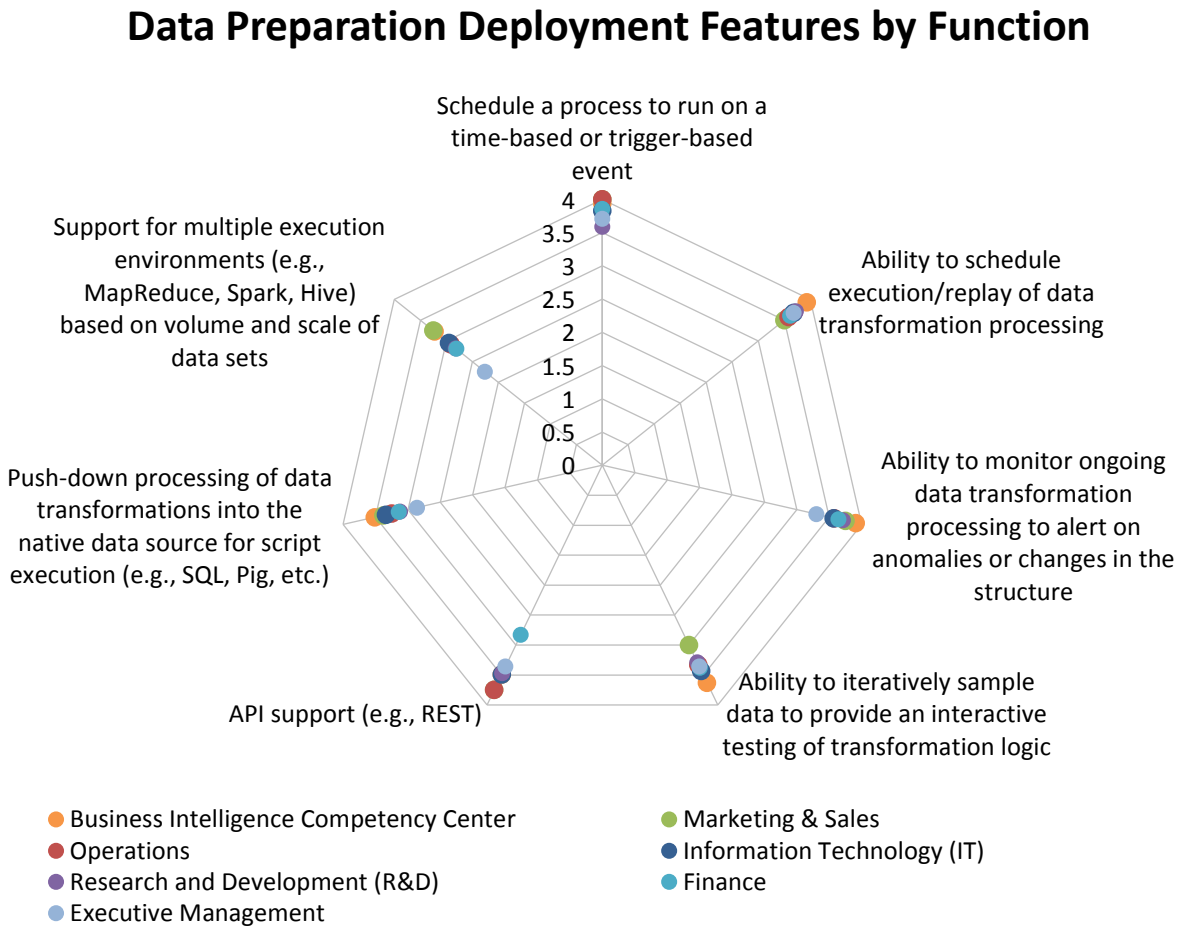


Figure 57 – Data preparation deployment features by function

Interest in data preparation scheduling, monitoring, and testing features varies somewhat by geography (fig. 58). In 2019, North American respondents have the greatest interest in newly added “schedule a process.” Latin American respondents narrowly lead tightly clustered respondent sentiment toward “ability to schedule” and “ability to monitor.” Asia-Pacific respondents reserve their highest score (above “very important”) for "ability to iteratively sample data." API support interest is highest in North America and EMEA.

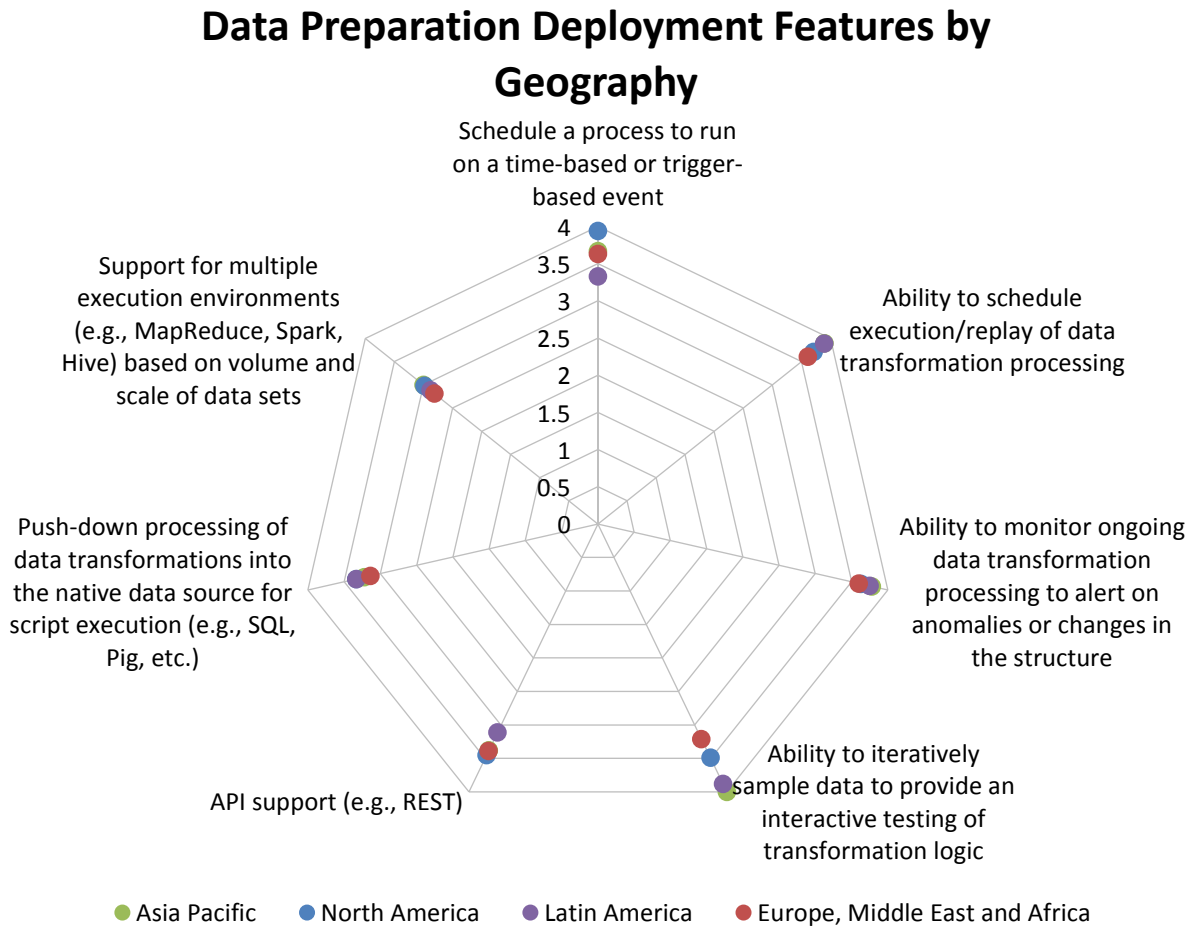


Figure 58 – Data preparation deployment features by geography

Interest in data preparation deployment features for the most part increases with organization global headcount (fig. 59). An exception in 2019 is for "schedule a process to run," where mid-sized organizations (101-1,000 employees) report the highest demand. (Mid-sized organizations also report above-mean interest in some trailing deployment feature requirements.) Outside this exception, very large organizations (>5,000 employees) give the highest scores to all other deployment features, and small organizations (1-100 employees) report the lowest or near-lowest scores.

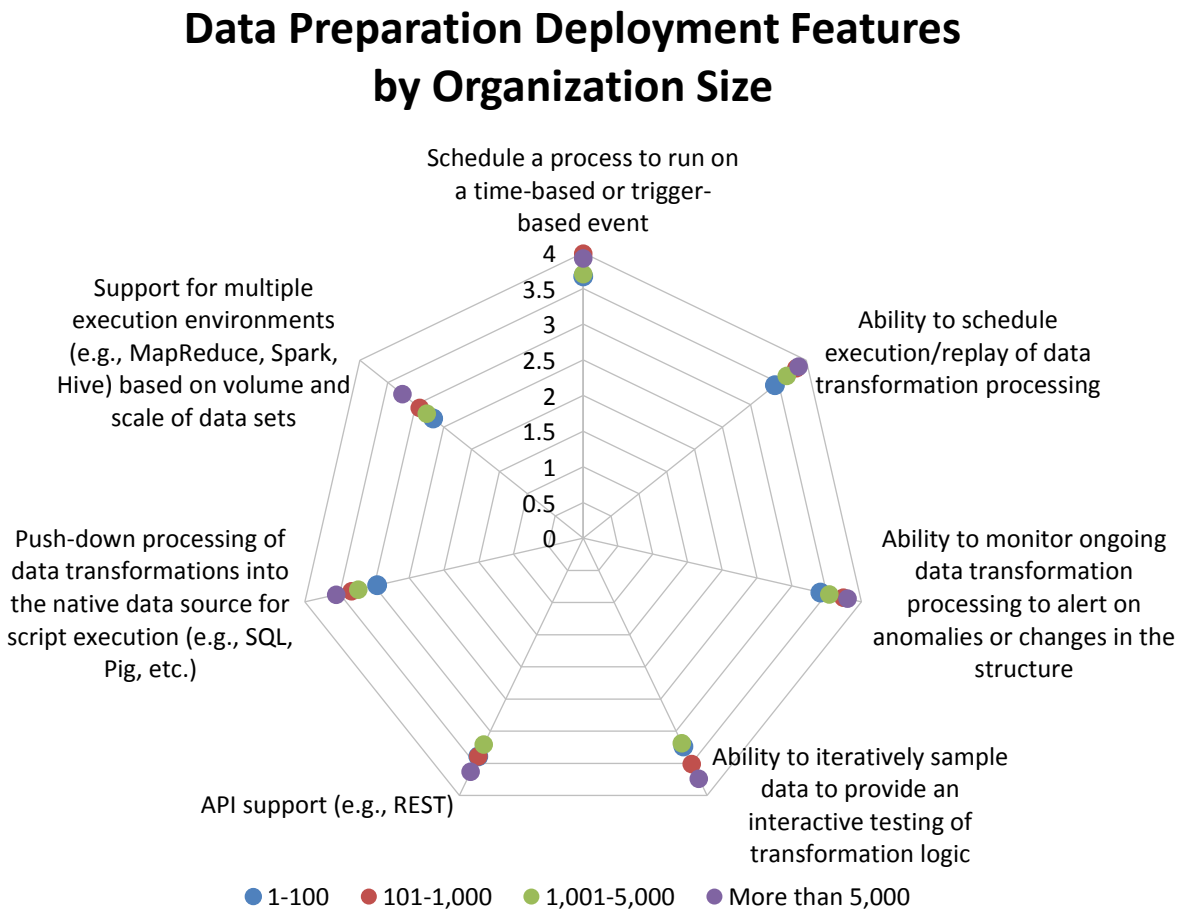


Figure 59 – Data preparation deployment features by organization size

Interest in data preparation deployment features varies broadly by industry (fig. 60). In 2019, Telecommunications respondents post the highest importance for "schedule a process to run" and "support for multiple execution environments." Education respondents have the greatest interest in "ability to monitor" and "ability to iteratively sample," while Government and Telecom respondents most often require "API support." "Push-down processing" gets top scores from Healthcare respondents.

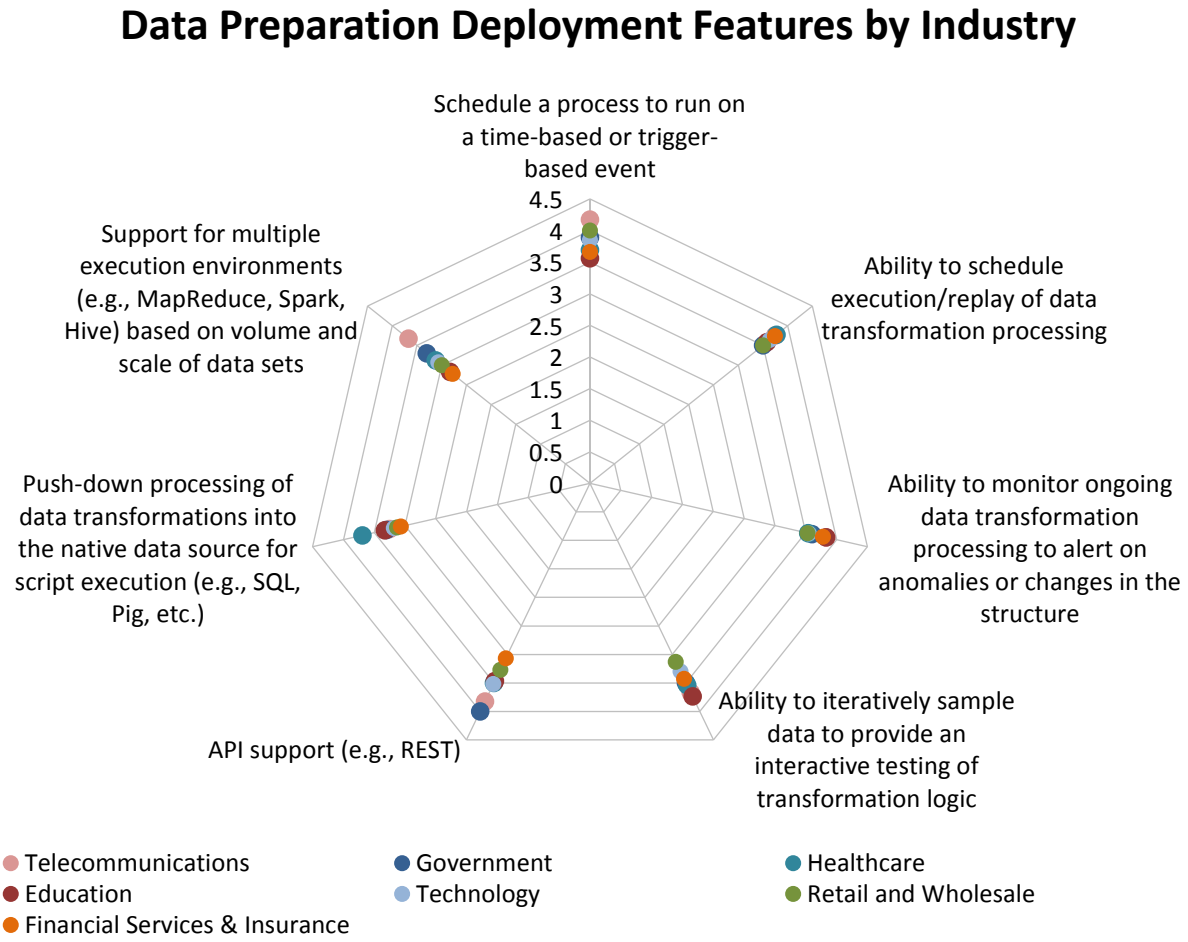


Figure 60 – Data preparation deployment features by industry

Location of Data preparation Capabilities

We gave respondents three choices to describe their preferred deployment location scenario for data preparation capabilities (fig. 61). In 2019, respondents prefer on-premises deployment (which might include desktop, LAN, or other configurations inside the firewall). Compared to on-premises deployments, which are "critical" or "very important" to almost 60 percent of respondents, private cloud deployments are "critical" or "very important" to a little less than half our sample. Public cloud sentiment is the least-regarded option, considered "critical" or "very important" to a little more than one-third of the sample.

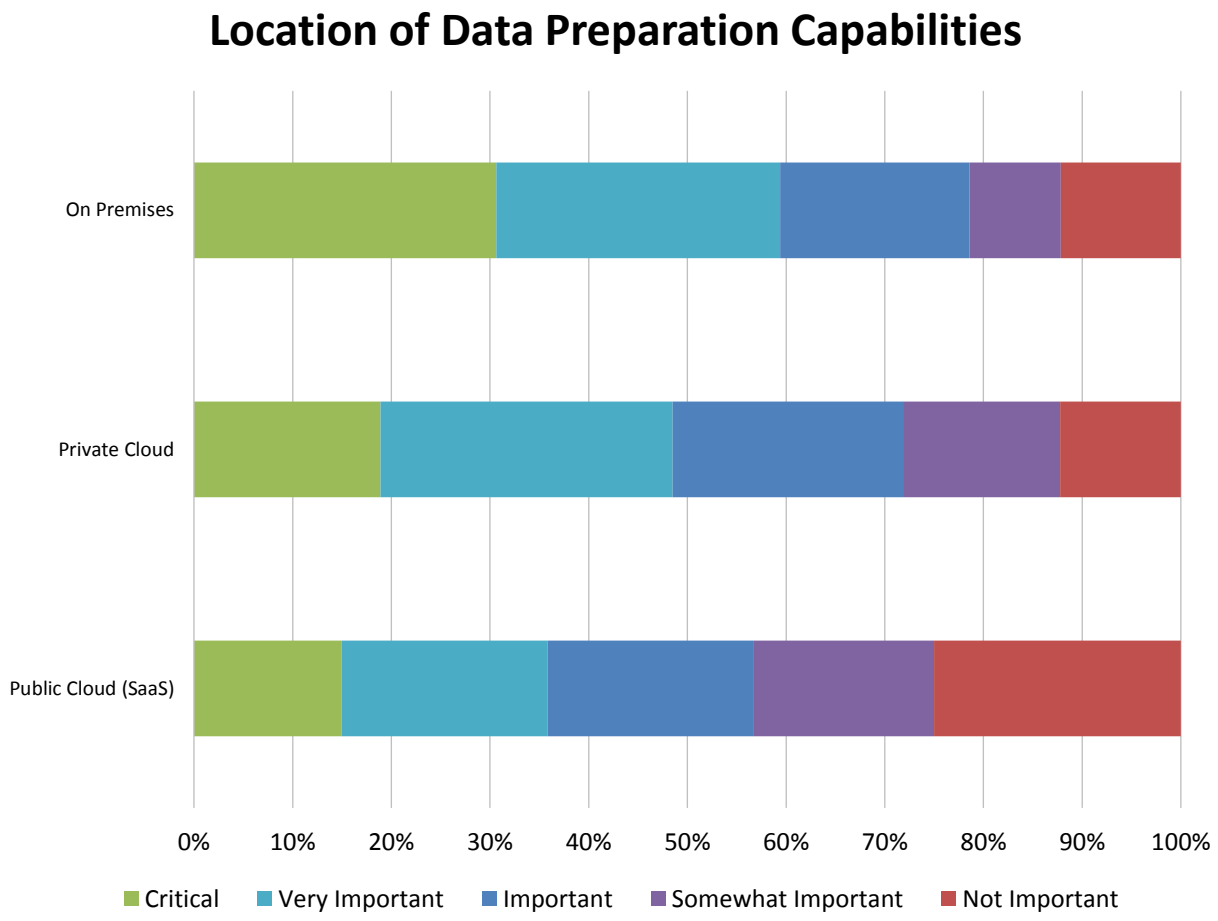


Figure 61 – Location of data preparation capabilities

Not every function most prefers on-premises deployment in 2019 (fig. 62). In our latest sample, Marketing/Sales, Operations, BICC, and IT most prefer on-premises deployments to other models. Finance is more evenly divided between on-premises and private cloud, while R&D is more likely to choose private cloud over on-premises deployment. All functions prefer on-premises or private cloud choices except Executive Management, which most often selects public cloud.

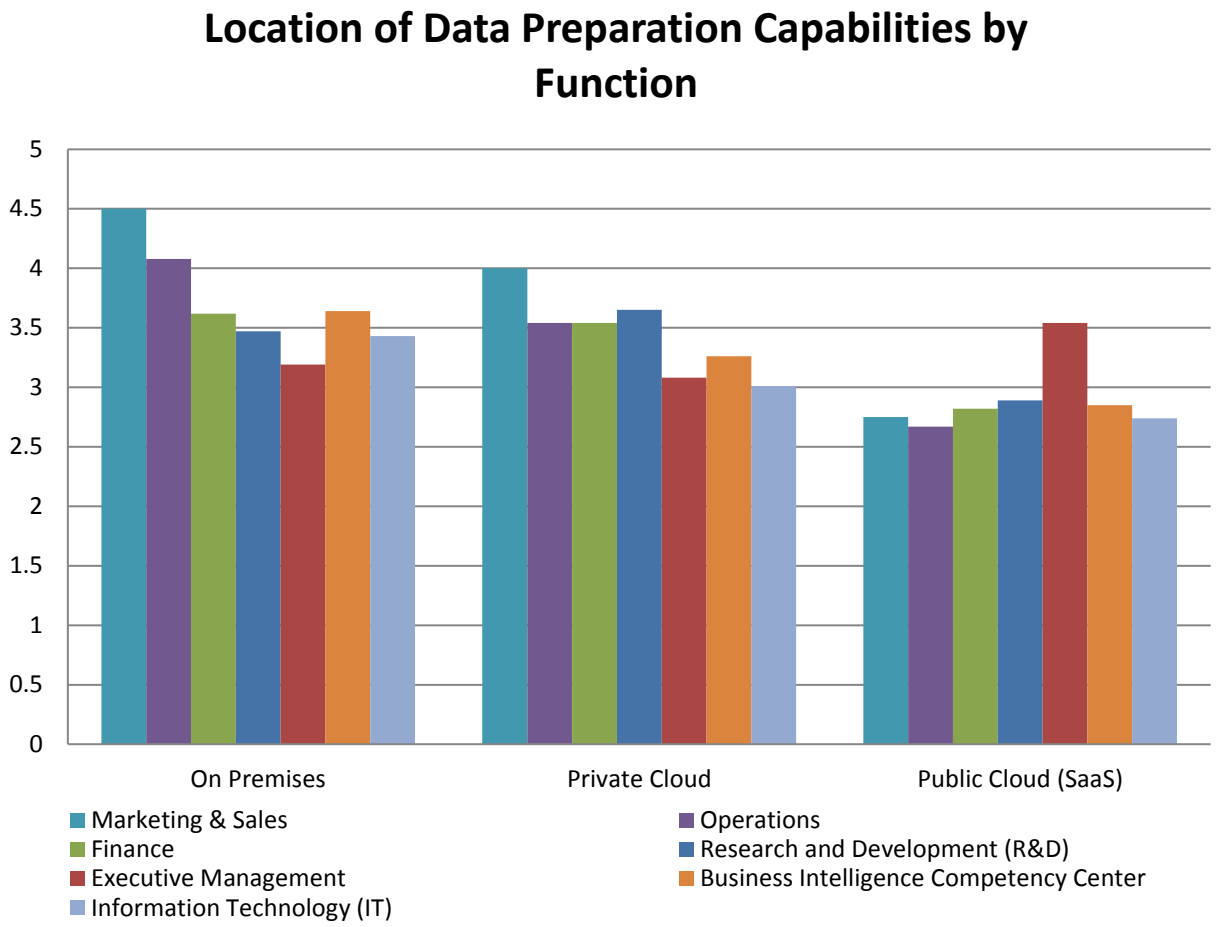


Figure 62 – Location of data preparation capabilities by function

The preference for on-premises capabilities for data preparation extends to most geographic regions in 2019 (fig. 63). Asia-Pacific, EMEA and Latin American respondents most prefer on-premises to private cloud deployment. North American respondents are more narrowly divided between on-premises and private cloud location of data preparation capabilities. Asia-Pacific respondents are the most likely users of public cloud for hosted data preparation, to a degree about equal to that for private cloud.

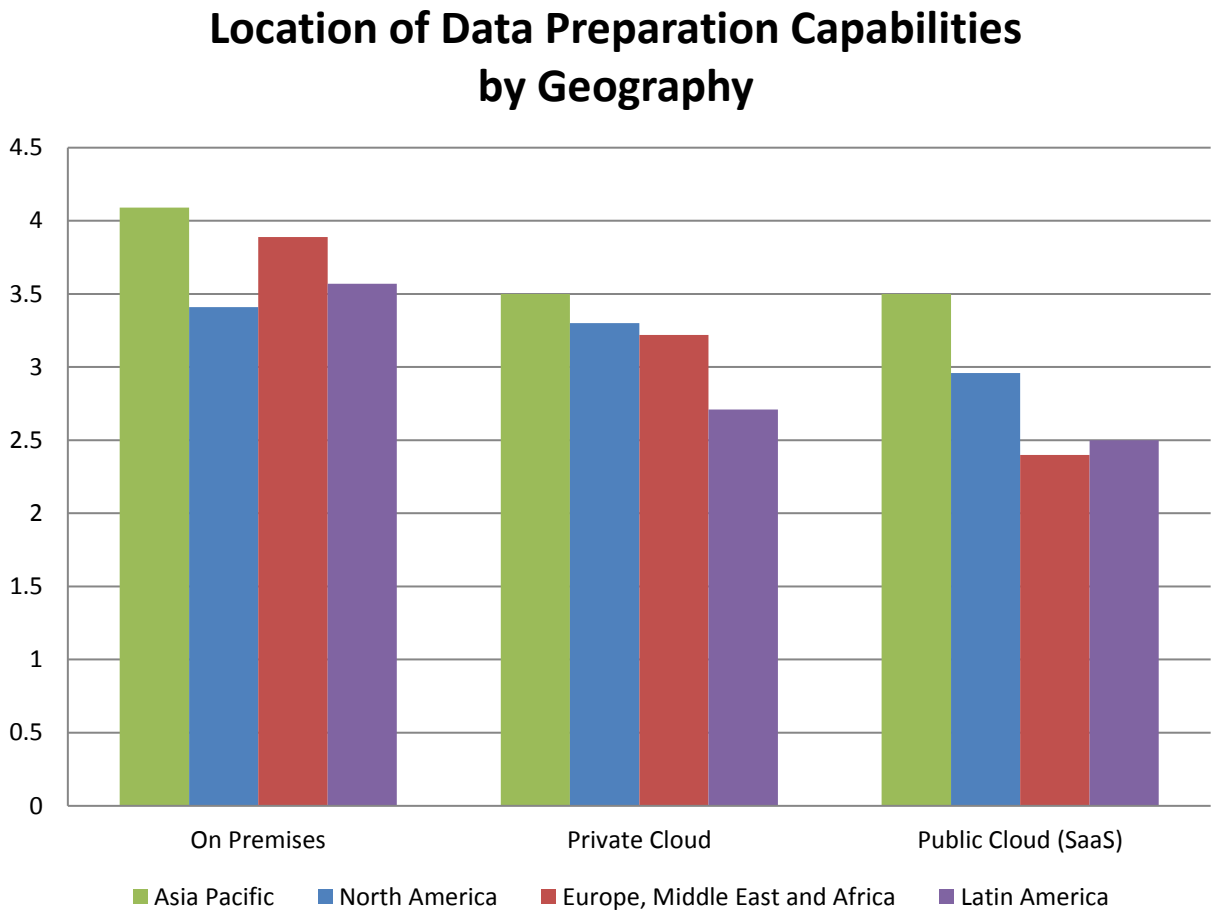


Figure 63 – Location of data preparation capabilities by geography

The preference for on-premises location of data preparation capabilities is predictably highest at large and very large organizations that are historically likely to maintain in-house infrastructure (fig. 64). Small organizations (1-100 employees) are more likely to have private cloud and most likely to use public cloud compared to on-premises capabilities. In 2019, large organizations (1,001-5,000 employees) are least likely to use private or public cloud compared to on-premises capabilities.

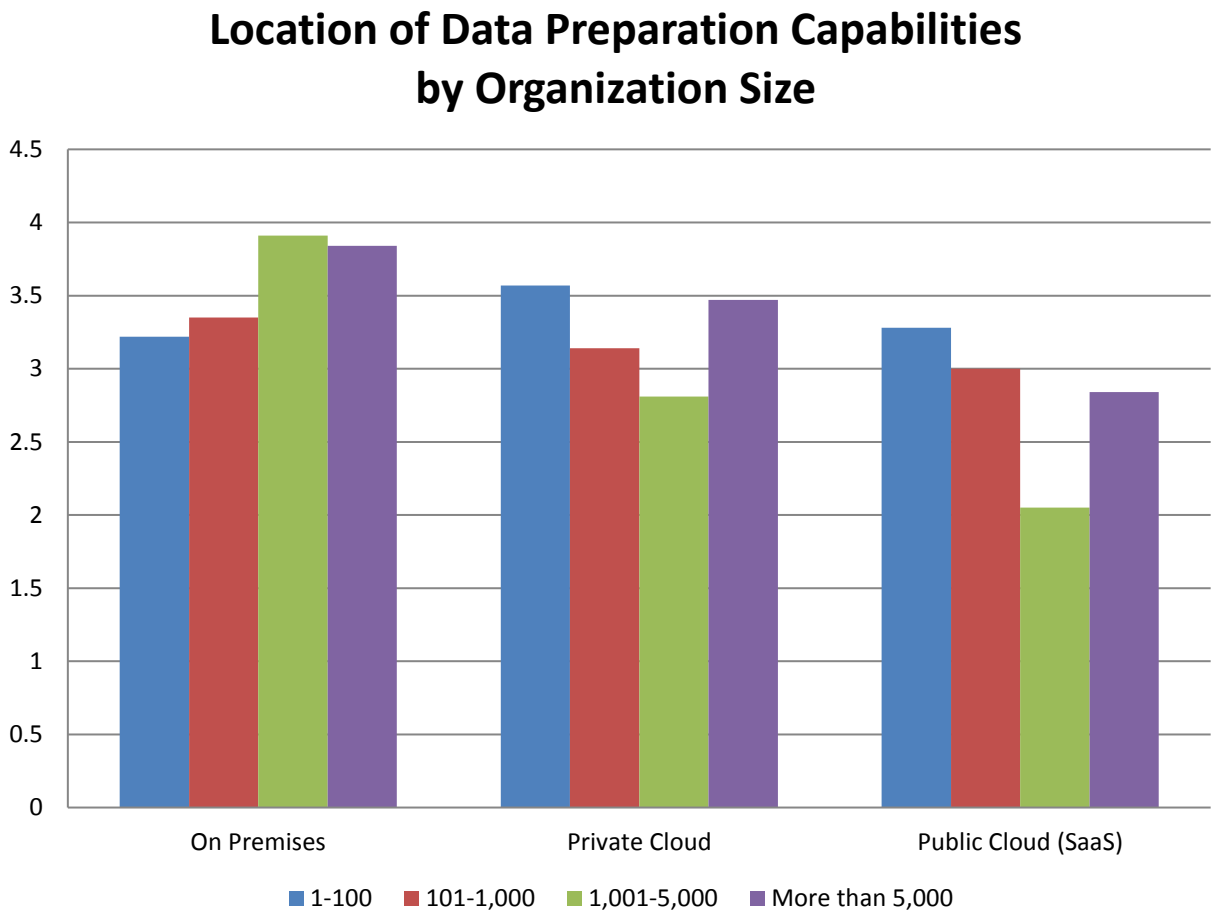


Figure 64 – Location of data preparation capabilities by organization size

Across vertical industries, the strongest preference for on-premises data preparation location is among Government, Retail/Wholesale, Financial Services, and Healthcare respondents (fig. 65). Conversely, Technology and Education organizations are more likely to use private cloud than on-premises or public cloud location, and Telecommunications respondents are quite notably more likely to locate on public cloud than on private cloud or deploy on premises.

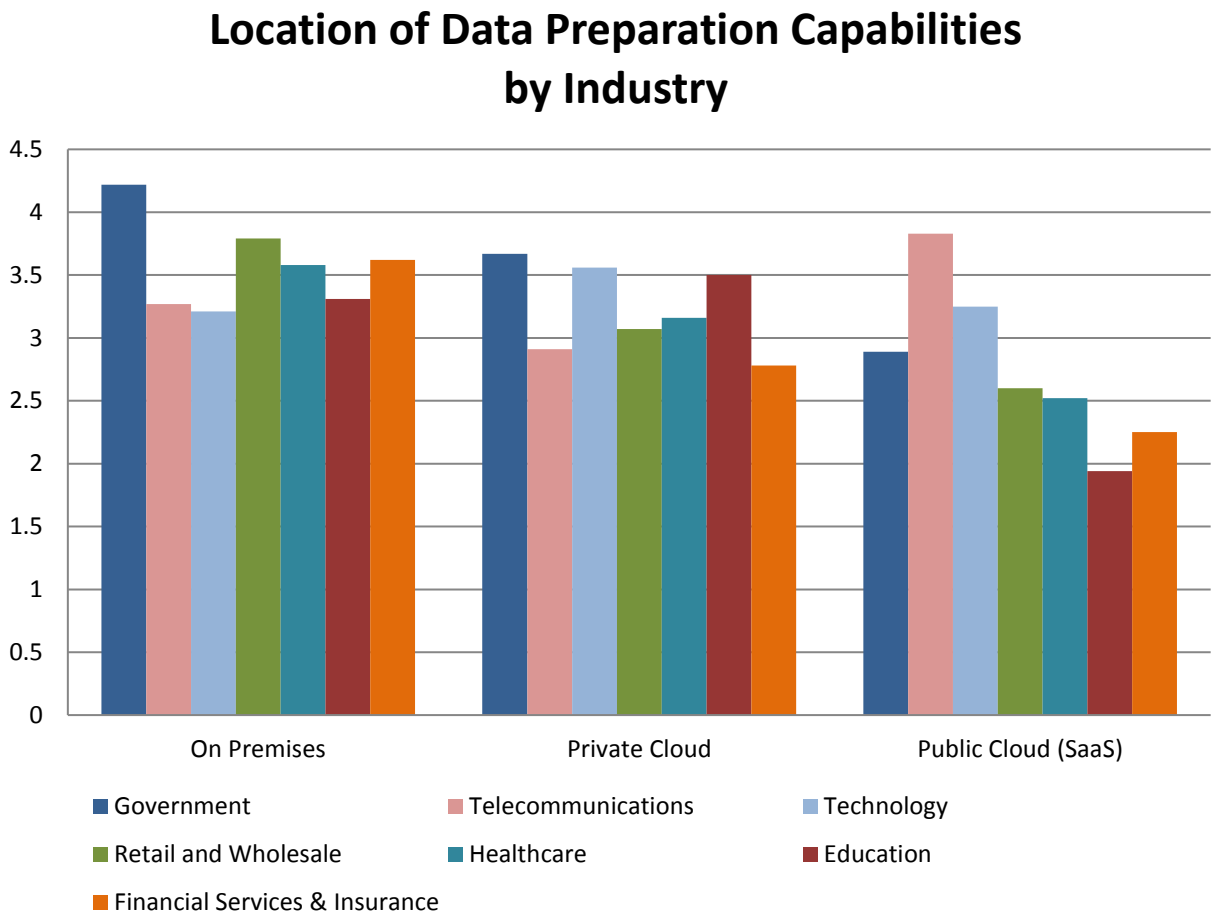


Figure 65 – Location of data preparation capabilities by industry

Industry Support for Data reparation

Like the end-user respondent community, the provider software and services industry attaches high importance to data preparation (fig. 66). That said, it appears that parts of the industry may feel data prep "crossed the chasm" since criticality and mean levels of importance peaked around 2016 and somewhat flattened since then. This year, sentiment ticks up slightly, with similar "critical" scores and more "very important" and "somewhat important" scores than were recorded in 2018. Current mean industry importance of about 3.5 is slightly below but in the range of user importance of about 3.7 (fig. 7, p. 20). While other BI imperatives may come more to the fore, we are confident that data preparation will be a common and transparent component or feature of BI tools going forward.

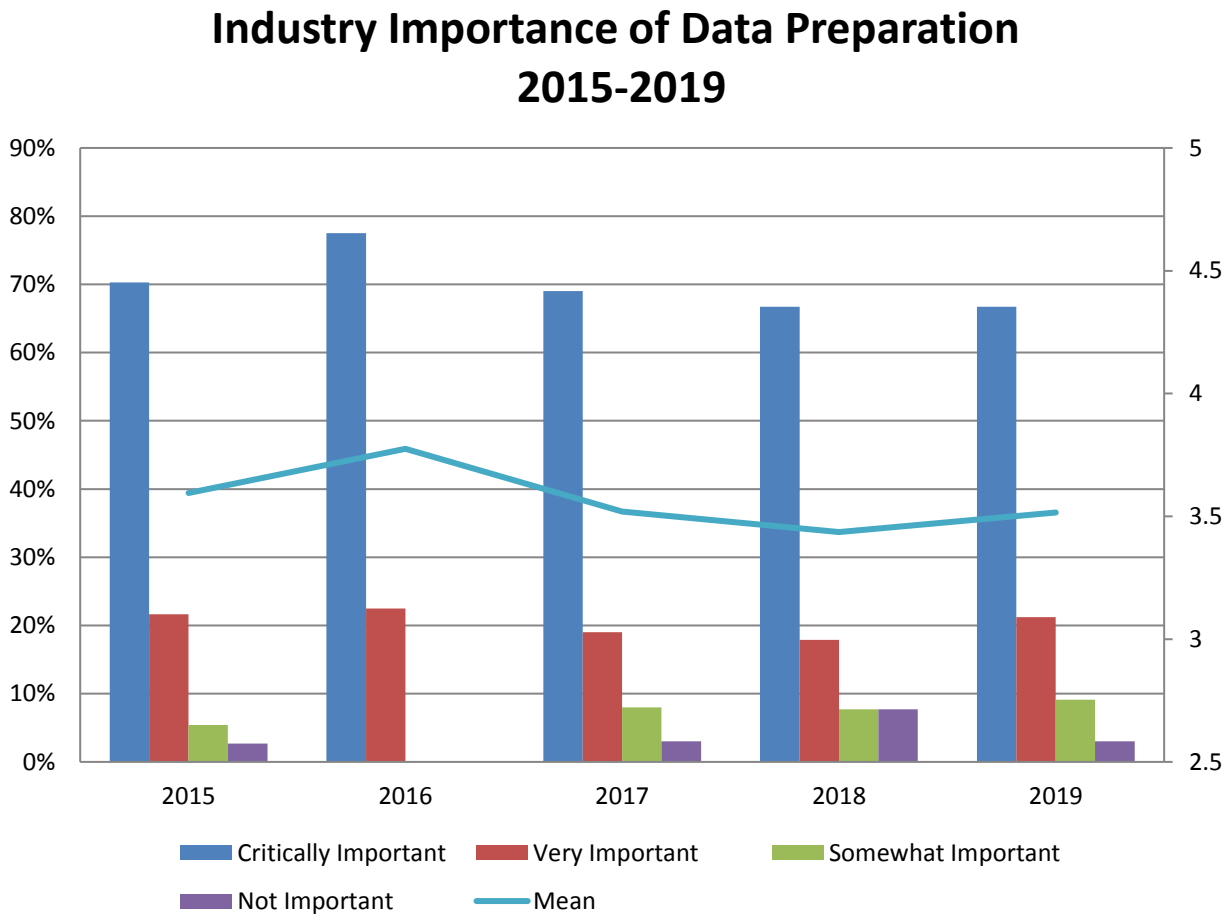


Figure 66 – Industry importance of data preparation 2015-2019

Industry Support for Data preparation Usability

We asked vendors to describe their current and future support for 13 usability features associated with data preparation (fig. 67). Generally, we can report that the industry provides high levels of support for user demand. The most and fully supported feature today, "support for entire data transformation process," along with other features are not neatly aligned with user priorities but are well ahead of current user needs (fig. 31, p. 44) in measures of support versus user criticality. The least-supported feature, machine learning, is far ahead of user requirements in 2019.

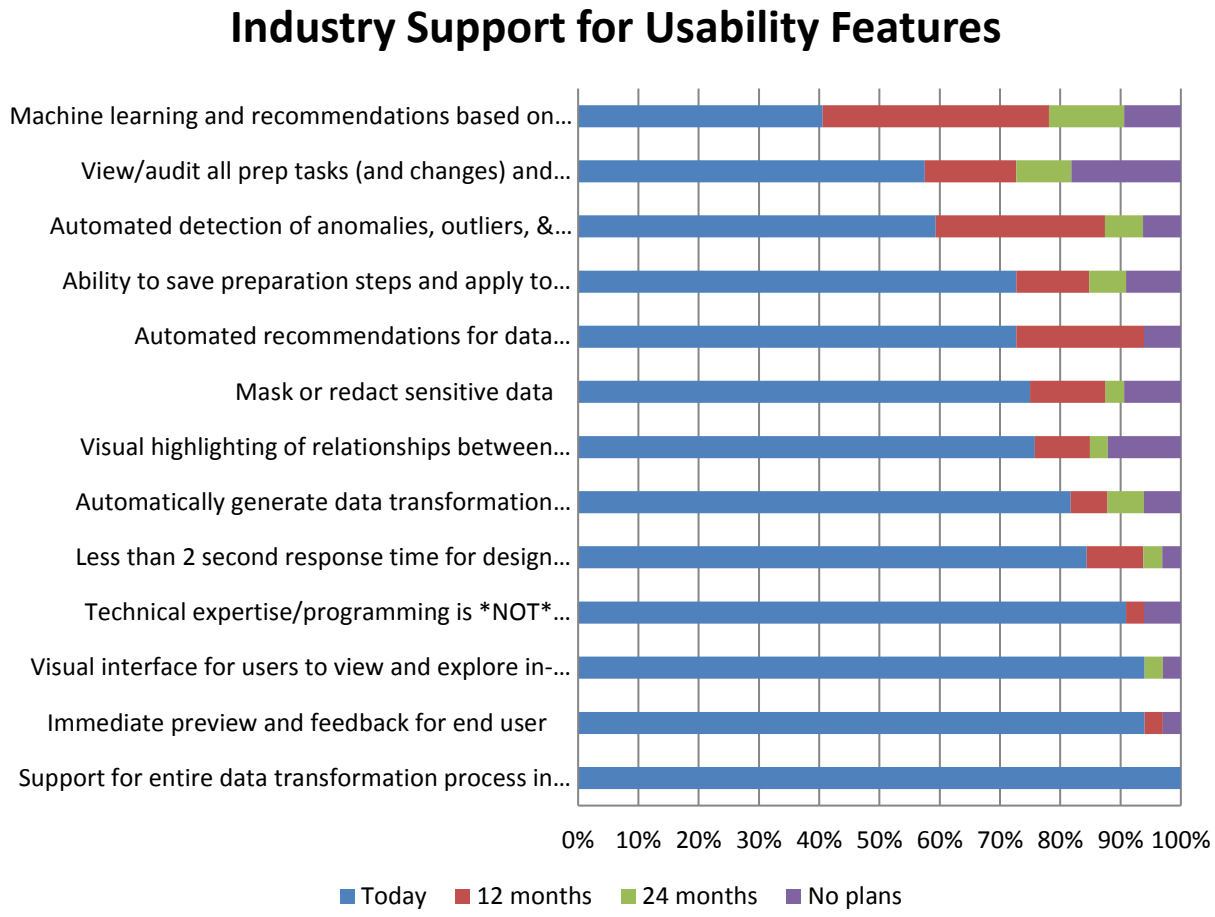


Figure 67 – Industry support for usability features

Industry Support for Data preparation Integration

Industry investment and support for data preparation integration features is also very strong and mature, with levels of current support above 80 percent for all but one function under our study in 2019 (fig. 68). All industry participants already support “ability to combine data.” There is near universal support for "access to traditional databases," "access to flat file formats," and other top user choices. Vendors expect all integration features except “ability to extract data from documents” to have greater than 90 percent support in the coming 12 months. Such robust support certainly answers all current user expectations for integration features (fig. 37, p. 50).

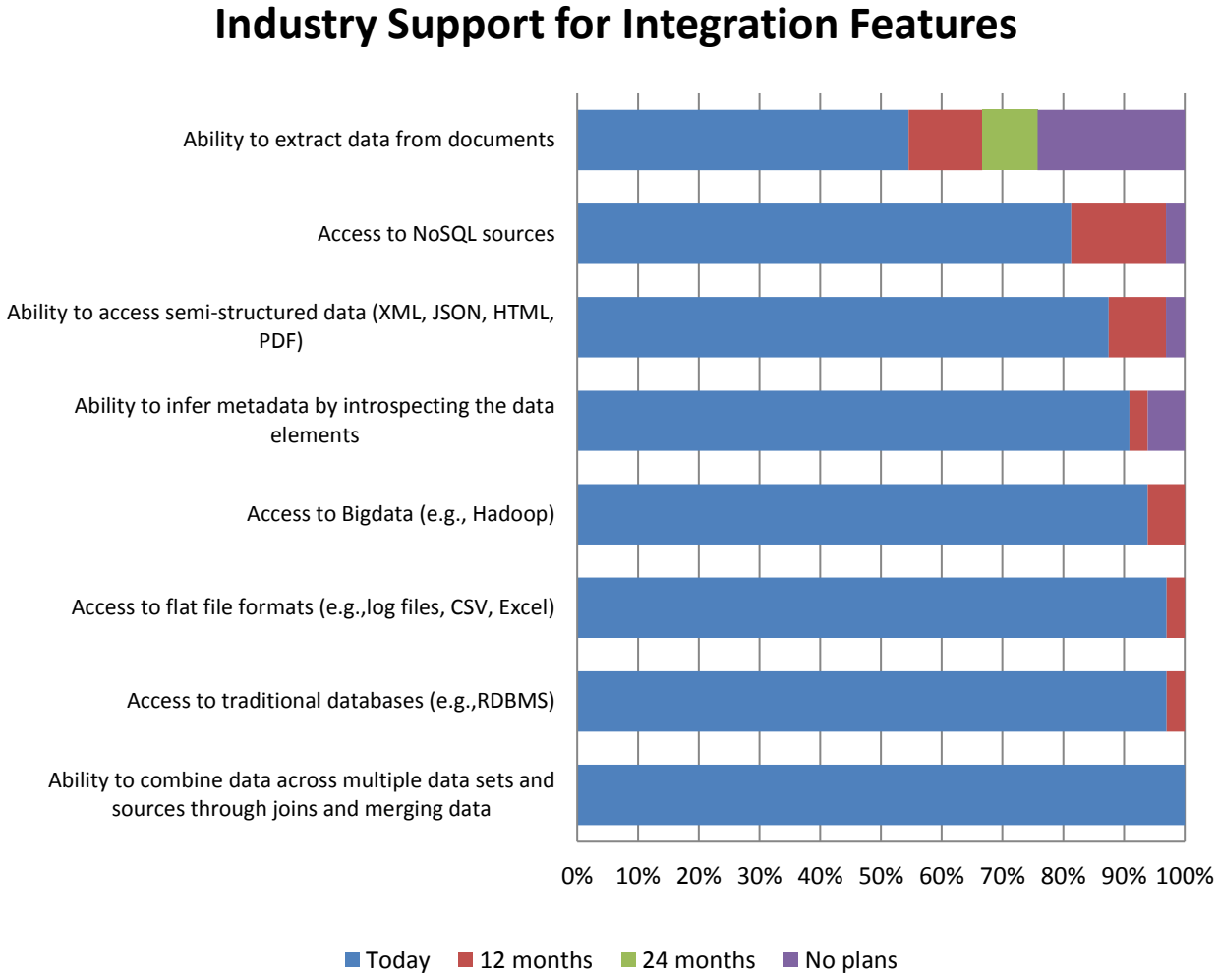


Figure 68 – Industry support for integration features

Industry Support for Data preparation Output Options

Industry support for output options is less robust than for integration and usability features but appears to meet the "80-20 rule" in addressing market demand (fig. 69). The four most-supported outputs match the top four user requirements (fig. 49, p. 62) in identical order. Industry current support also appears to address the growing user uptake of Azure and other outputs (fig. 50, p. 63).

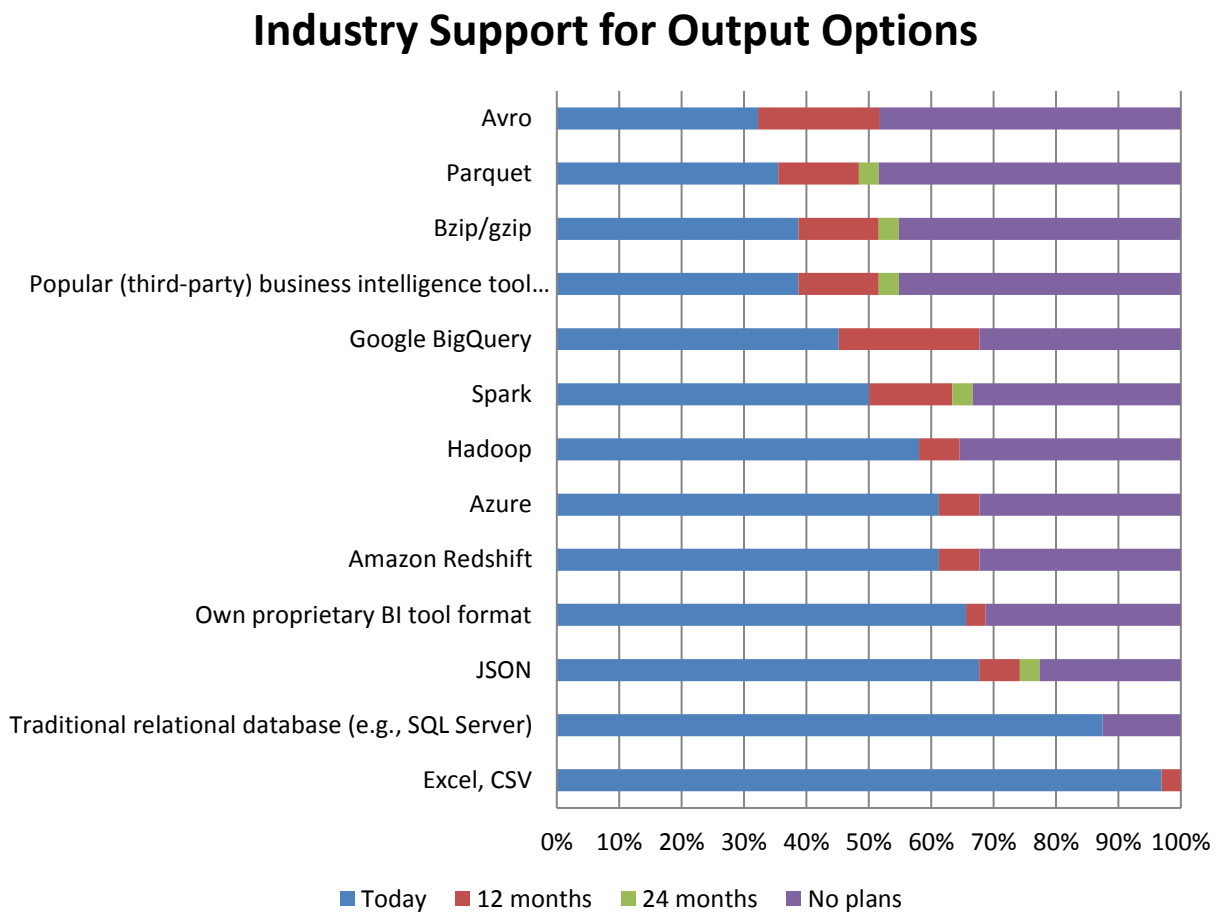


Figure 69 – Industry support for output options

Industry Support for Data preparation Data Manipulation Features

Industry support for data manipulation data preparation features is strong "across the board" in 2019 (fig. 70). The top seven features all currently enjoy 80 percent or greater support, and vendors expect all but two data manipulation features we sampled to approach 80 percent support within 12 months. The top industry-supported manipulation features mostly match and easily accommodate the top user priorities (fig. 43, p. 56).

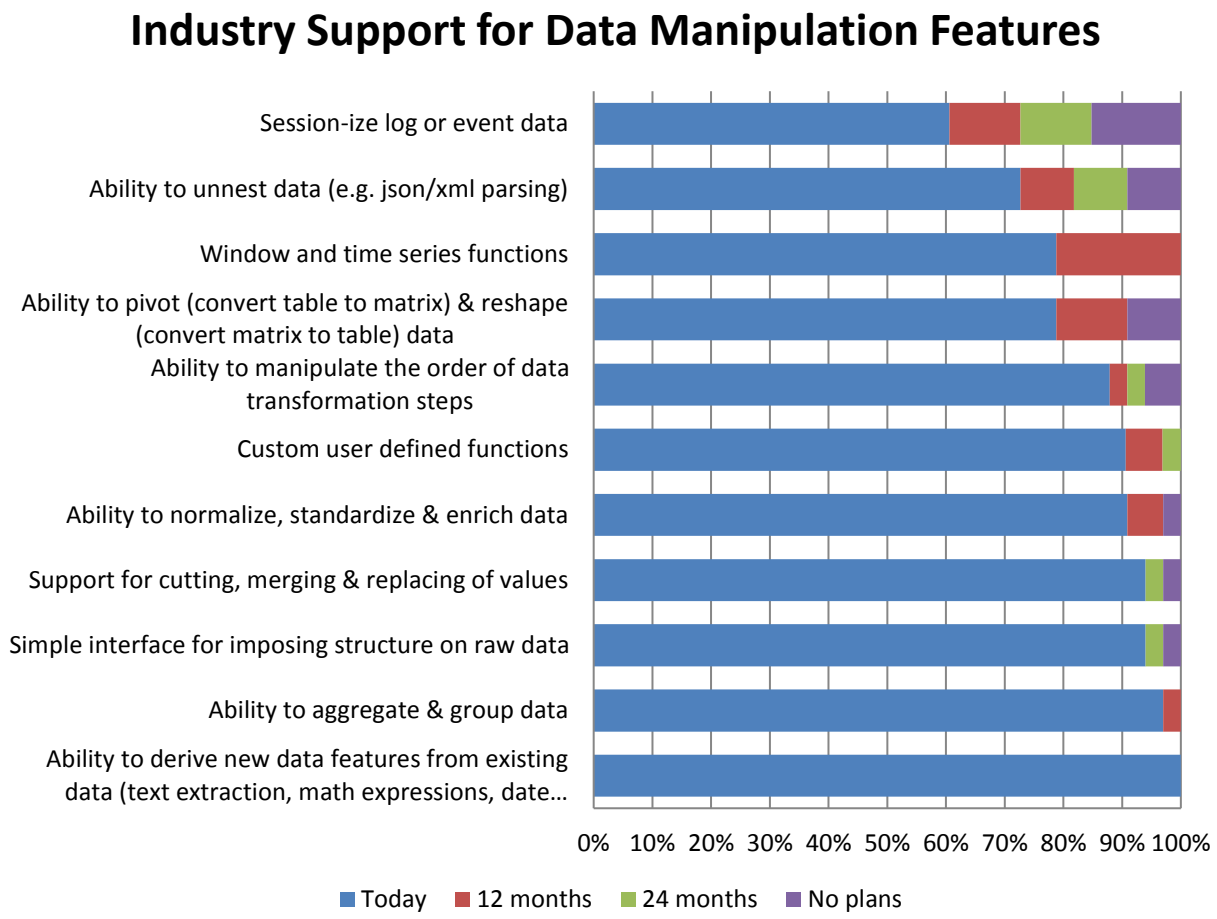


Figure 70 – Industry support for data manipulation features

Industry Support for Data preparation Deployment Features

Industry support of data preparation deployment features is robust, and vendors expect to extend investment in capabilities in 2019 (fig. 71). Top user priorities (fig. 55, p. 68), are somewhat aligned with current industry support, and current support appears to easily surpass user requirements. The third top user requirement, “ability to monitor,” is among the less-supported features today, but vendors expect future investment to keep availability well ahead of user requirements.

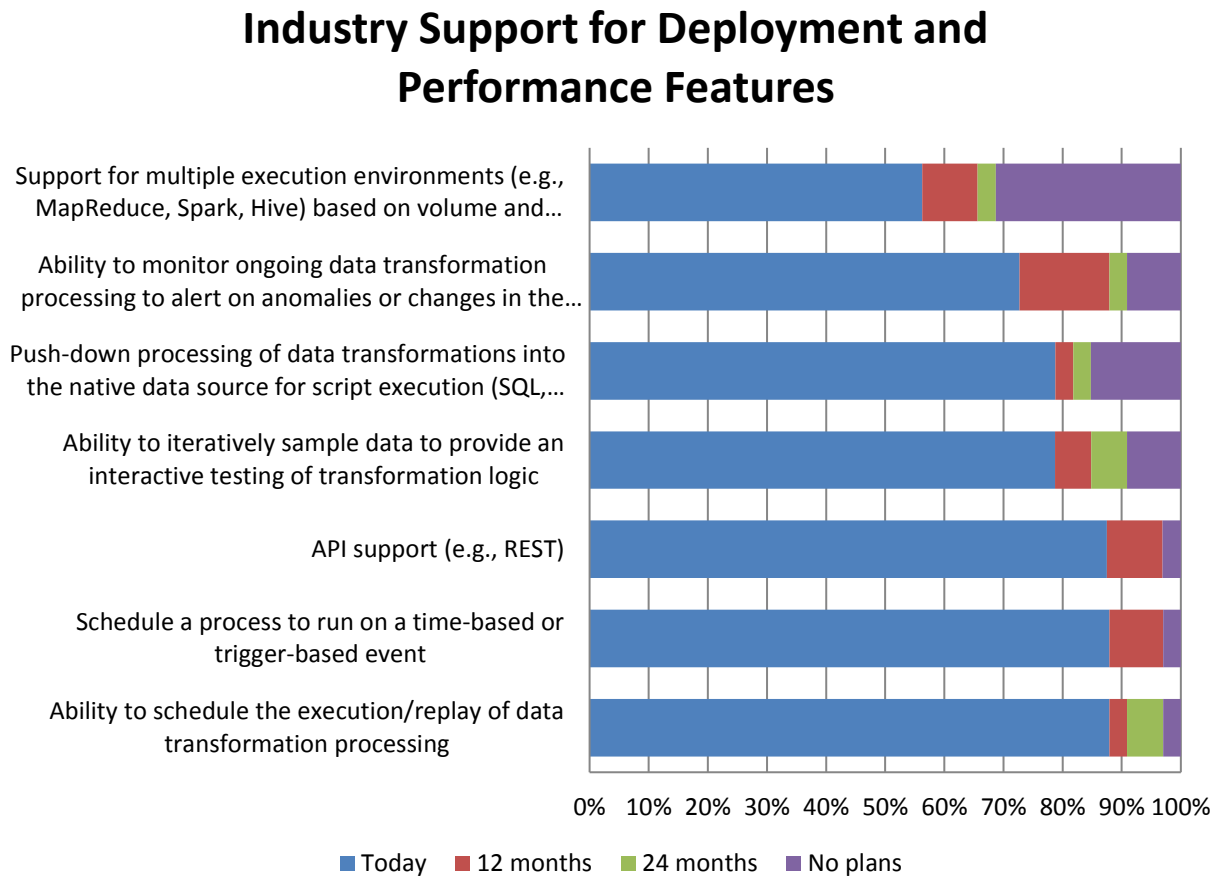


Figure 71 – Industry support deployment and performance features

Industry Support for Data preparation—Cloud versus On-Premises

Industry support for data preparation industry deployment options is near maturity in 2019 with little future growth of penetration expected (fig. 72). Currently, more than 80 percent of vendors support on-premises deployment; even more, about 91 percent currently support cloud deployment. We note for the first time in our study that more industry products are available for cloud versus on-premises deployment of data preparation. As noted earlier (fig. 61, p. 74), user demand still leans toward on-premises versus public or private cloud deployment. Assuming users shift towards greater cloud deployment, industry support will already be in place.

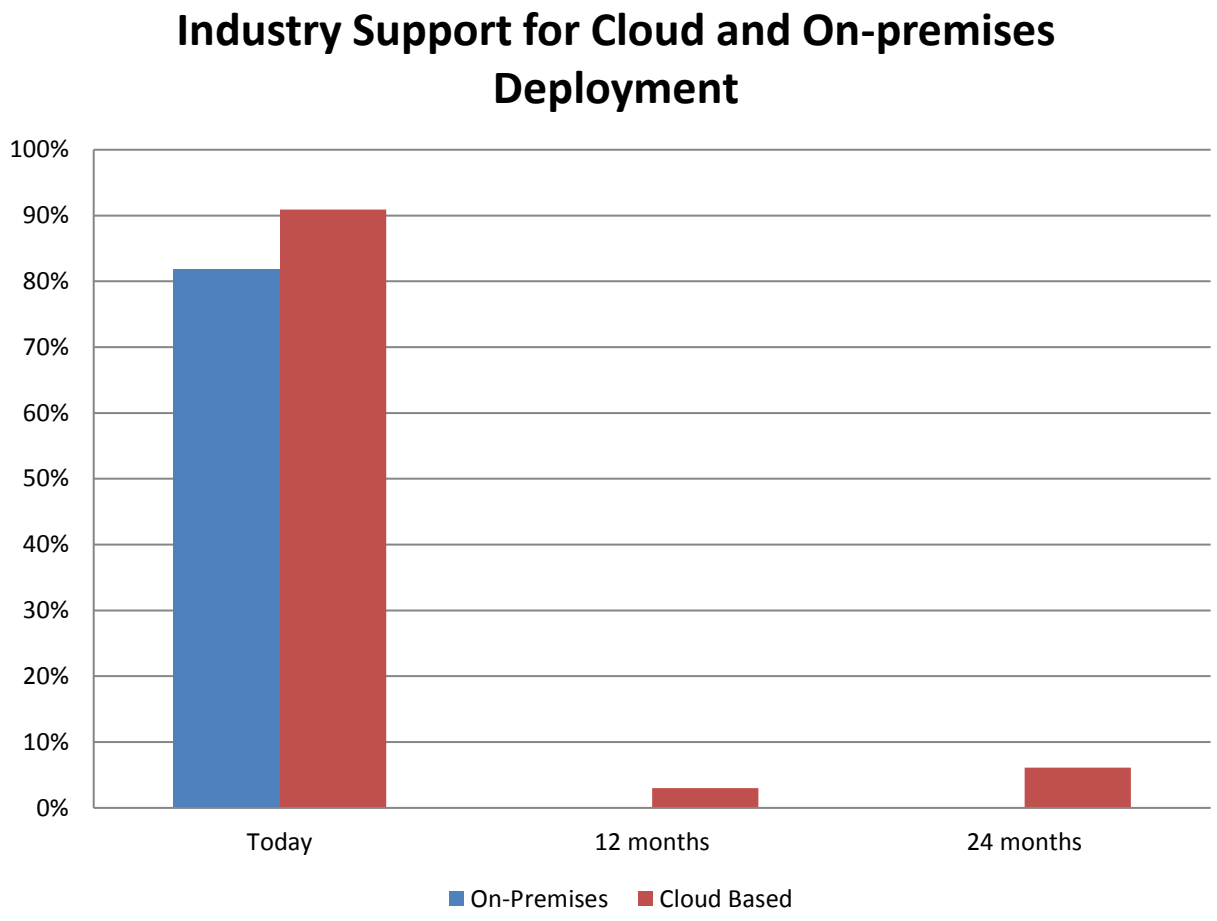


Figure 72 – Industry support for cloud and on-premises deployment

Fig. 73 shows another instructive view of industry support growth for cloud-based data preparation deployment over time. Our 2019 study shows not only a 7 percent increase in the number of cloud-based offerings but a similar decline in on-premises data product and service offerings compared to 2018. Indeed, on-premises capabilities appear to diminish by about 10 percent from a 2017 high, while the five-year trajectory of cloud-based data preparation has been growing.

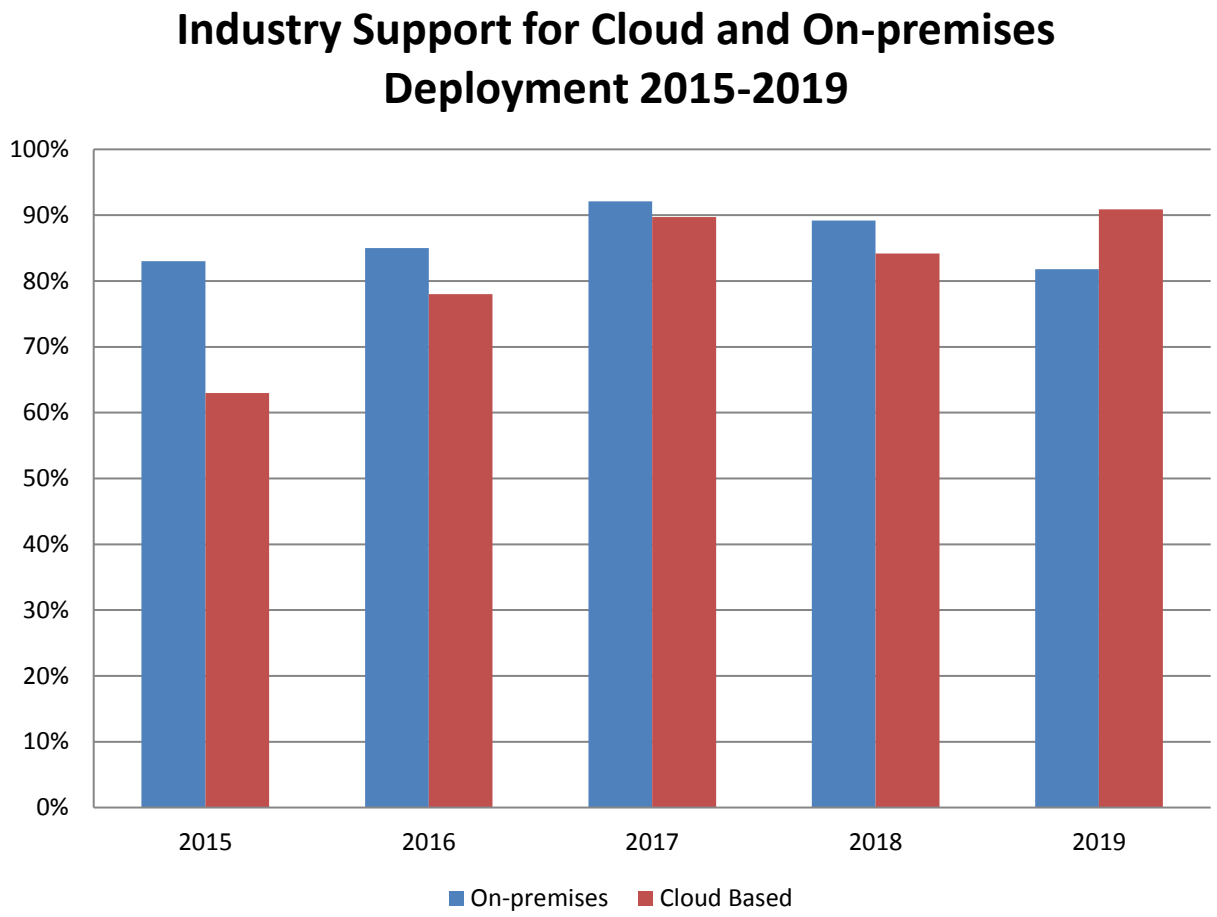


Figure 73 – Industry support for cloud and on-premises deployment 2015-2019

Data Preparation Vendor Ratings

We include 24 vendors in our data preparation ratings (fig. 74). For each vendor, we consider usability, integration, output, data manipulation, and deployment features. Only vendors that score 50 percent or greater are included in this report.

Top-rated vendors include Trifacta (1st), ClearStory Data (2nd), Information Builders (3rd), Datameer (4th) and Altair/Datawatch (5th).

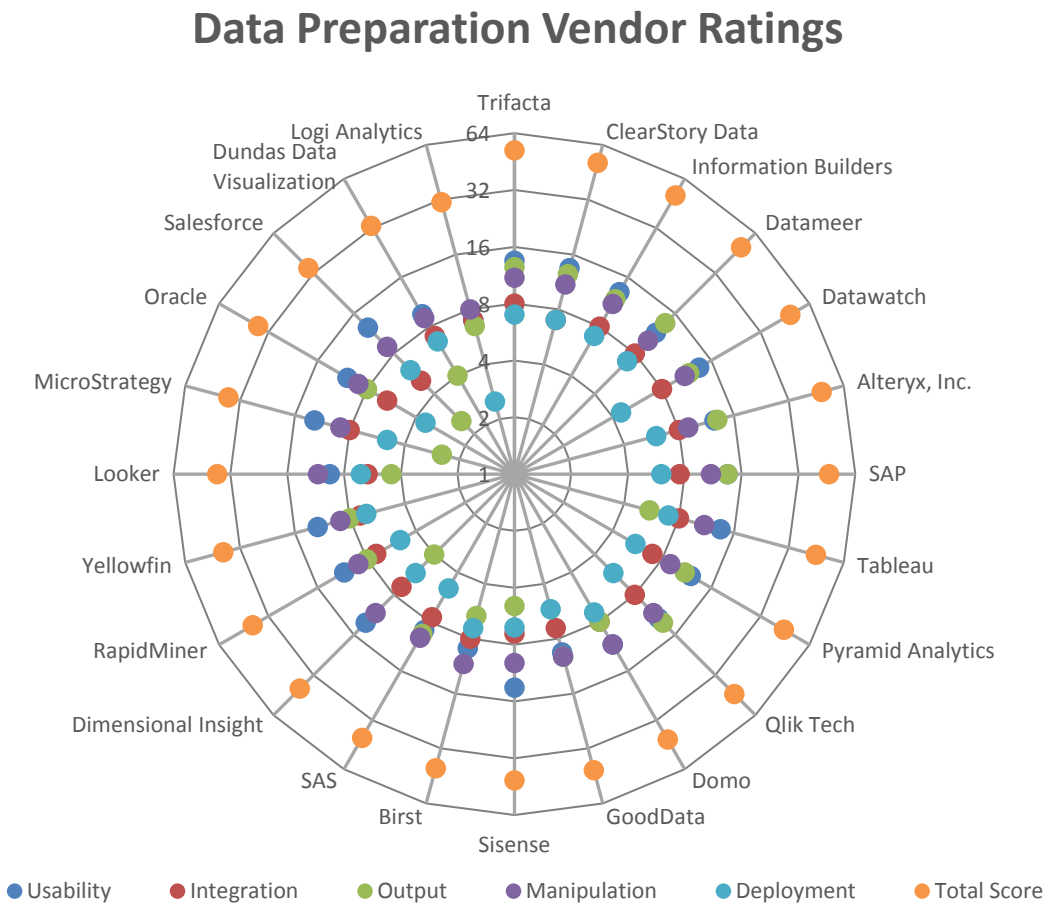


Figure 74 – Data preparation vendor ratings

Other Dresner Advisory Services Research Reports

- "Flagship" Wisdom of Crowds® Analytical Data Infrastructure Market Study
- "Flagship" Wisdom of Crowds® Business Intelligence Market Study
- "Flagship" Wisdom of Crowds® Enterprise Planning market Study
- Advanced and Predictive Analytics
- Big Data Analytics
- Business Intelligence Competency Center
- Cloud Computing and Business Intelligence
- Collective Insights®
- Data Catalog
- Embedded Business Intelligence
- Enterprise Planning
- IoT Intelligence®
- IT Analytics
- Location Intelligence
- Small and Mid-Sized Enterprise Business Intelligence
- Small and Mid-Sized Planning

Appendix: Data preparation Survey Instrument

Name*: _____

Company Name: _____

Address 1: _____

Address 2: _____

City: _____

State: _____

Zip: _____

Country: _____

Email Address*: _____

Phone Number: _____

Major Geography

( ) Asia/Pacific

( ) Europe, Middle East and Africa

( ) Latin America

( ) North America

What is your current title?

_____

What function are you a part of?

( ) Business intelligence competency center

( ) Executive management

( ) Finance

( ) Information Technology (IT)

( ) Manufacturing

( ) Marketing

( ) Project/program management office

( ) Sales

( ) Research and development (R&D)

( ) Other - Write In: _____

Please select an industry

( ) Advertising

( ) Aerospace

( ) Agriculture

( ) Apparel and accessories

( ) Automotive

( ) Aviation

( ) Biotechnology

( ) Broadcasting

( ) Business services

( ) Chemical

( ) Construction

( ) Consulting

( ) Consumer products

( ) Defense

( ) Distribution & logistics

( ) Education

( ) Energy

( ) Entertainment and leisure

( ) Executive search

( ) Federal government

( ) Financial services

( ) Food, beverage and tobacco

( ) Healthcare

( ) Hospitality

( ) Gaming

( ) 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 - 1000

( ) 1001 - 5000

( ) More than 5000

How important is it for users to be able to prepare data (e.g., combine, clean, shape datasets) prior to analysis?*

( ) Critical

( ) Very important

( ) Important

( ) Somewhat important

( ) Not important

What tool(s) do users currently use to prepare data for analysis?

_____

_____

_____

_____

How effective is the current approach to Data preparation for Business Intelligence/user analysis today?

( ) Highly effective

( ) Somewhat effective

( ) Somewhat ineffective

( ) Totally ineffective

How often do users have to prepare data (e.g., combine, clean and shape datasets) to get it in a format that can be used for analysis?

( ) Constantly

( ) Frequently

( ) Occasionally

( ) Rarely

( ) Never

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

( ) Constantly

( ) Frequently

( ) Occasionally

( ) Rarely

( ) Never

Should Data preparation be a standalone capability or part of another tool?

( ) Standalone

( ) Part of business intelligence tools

( ) Part of existing data quality/data integration tools

Please indicate the importance of the following usability features for Data preparation software:

	Critical	**Very important**	**Important**	**Somewhat important**	**Not important**
Technical expertise/programming is *NOT* required to build/execute data transformation scripts	( )	( )	( )	( )	( )
Immediate preview and feedback for end user	( )	( )	( )	( )	( )
Automated recommendations for data relationships & keys for combining data across multiple data sets and sources	( )	( )	( )	( )	( )
Visual interface for users to view and explore in-process data sets, interactively profile and refine data transformations prior to execution	( )	( )	( )	( )	( )
Visual highlighting of relationships between columns, attributes & datasets	( )	( )	( )	( )	( )
Automated detection of anomalies, outliers, & duplicates	( )	( )	( )	( )	( )

Automatically generate data transformation code/scripts for execution	( )	( )	( )	( )	( )
Support for entire data transformation process in a single application/user interface	( )	( )	( )	( )	( )
Machine learning and recommendations based on usage data gathered across users, groups, or organizations	( )	( )	( )	( )	( )

Please indicate the importance of the following data integration features for Data preparation software:

	Critical	**Very important**	**Important**	**Somewhat important**	**Not important**
Access to traditional databases (e.g., RDBMS)	( )	( )	( )	( )	( )
Access to big data (e.g., Hadoop)	( )	( )	( )	( )	( )
Access to NoSQL sources	( )	( )	( )	( )	( )
Access to	( )	( )	( )	( )	( )

file formats (e.g., log files, CSV, Excel)					
Ability to infer metadata by introspecting the data elements	( )	( )	( )	( )	( )
Ability to combine data across multiple data sets and sources through joins and merging data	( )	( )	( )	( )	( )

What output formats should an Data preparation solution support?

[ ] Traditional relational database (e.g., SQL Server)

[ ] Excel, CSV

[ ] Popular (third-party) business intelligence tool formats

[ ] Hadoop

[ ] Redshift

[ ] Azure

[ ] Avro

[ ] Parquet

[ ] Bizp/gizp

[ ] Other - Write In: _____

Please indicate the importance of the following data manipulation features for Data preparation software:

	Critical	**Very important**	**Important**	**Somewhat important**	**Not important**
Simple interface for imposing structure on raw data	( )	( )	( )	( )	( )
Ability to un-nest 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	( )	( )	( )	( )	( )

Please indicate the importance of the following deployment features for Data preparation software:

	Critical	**Very important**	**Important**	**Somewhat important**	**Not important**
Ability to iteratively sample data to provide an interactive testing of transformation	( )	( )	( )	( )	( )

logic					
Push-down processing of data transformations into the native data source for script execution (SQL, Pig, etc.)	( )	( )	( )	( )	( )
Ability to schedule the execution/replay of data transformation processing	( )	( )	( )	( )	( )
Ability to monitor ongoing data transformation processing to alert on anomalies or changes in the structure	( )	( )	( )	( )	( )
Support for multiple execution environments (e.g., MapReduce, Spark, Hive) based on volume and scale of data sets	( )	( )	( )	( )	( )
API support (e.g., REST)	( )	( )	( )	( )	( )

Where should Data preparation functionality reside?

	Critical	**Very important**	**Important**	**Somewhat important**	**Not important**
On-premises	( )	( )	( )	( )	( )
Private cloud	( )	( )	( )	( )	( )
Public cloud (SaaS)	( )	( )	( )	( )	( )