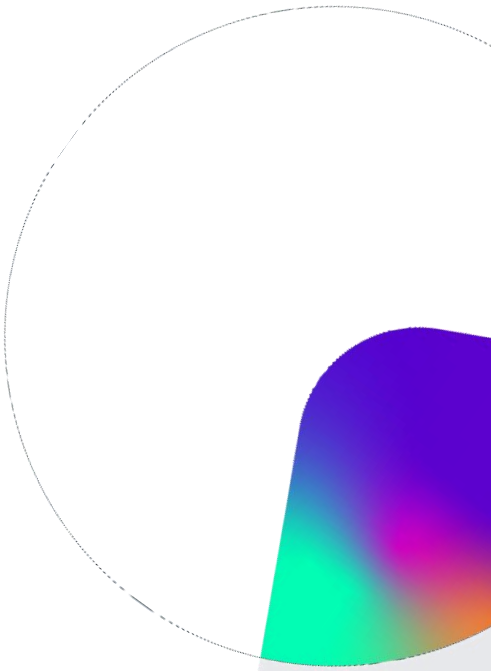
2024 State of Big Data Analytics: Constant Compromising Is Leading to Suboptimal Results

Survey Report, June 2024

Table of Contents

Introduction and Key Findings



Introduction & Methodology

The rapid growth in generative AI over the past couple of years has resulted in the need to analyze massive (and growing) volumes of data. As a company that provides enterprises with critical business intelligence from massive data stores, we were keen to explore how this rapid growth in GenAI has impacted big data analytics in 2024.

The aim of this survey was therefore to better understand the connection between the *cost* of analytics and the *results* of analytics (cost-performance), to identify the primary pain points and challenges currently facing those who work in the analytics space, and to discover the methods organizations are using to address these challenges.

The report is aimed at team leaders, CIOs, heads of data and other data-oriented professionals in large to very large organizations (500+ employees).

Methodology

To get more insight into current trends shaping big data analytics, we commissioned a survey of 300 senior professionals and decision makers in data management and FinOps roles (financial managers who engage cross-functional teams in a collaborative effort to control cloud computing infrastructure and costs) to shed light on their most pressing challenges and priorities.

This report was administered online by Global Surveyz Research, an independent global research firm. The survey is based on responses from data leaders, including CIOs, CDOs, Heads of Data and Heads of Analytics (69%), and FinOps executives (31%).

Respondents hailed from US companies with at least $5M+ annual spend on cloud, and using either AWS, GCP (Google) or Azure (Microsoft) for their cloud infrastructure. 46% of the companies surveyed manage over 1PB data+, 41% with 100TB-1PB, the rest under 100TB. Ten industries were represented by the participating companies, including Banking and Financial Services, Health and Pharma, Information Technology, Insurance, Manufacturing, Media, Retail and eCommerce, Software Development, Technology (excluding software development) and Telecom.

The respondents were recruited through a global B2B research panel and invited via email to complete the survey, with all responses collected during March-April 2024. The average amount of time spent on the survey was 5 minutes and 49 seconds. The answers to most of the non-numerical questions were randomized to prevent order bias in the answers.

Key Findings

1 Most organizations experience analytics “bill shock” often, with more than two-thirds experiencing it at least quarterly, or more frequently

When it comes to data analytics processes, the larger the volumes of data a company works with, the more compute power it needs, and the higher the associated costs. Although billing cycles vary from company to company, when asked how often they experience bill shock, 71% of respondents (2 out of 3 companies) reported they are surprised by the high costs of their cloud analytics bill fairly frequently (Figure 3), with 5% experiencing bill shock monthly, 25% bimonthly and 41% quarterly – which is pretty significant. Almost a third of the companies surveyed (29%) say they experience bill shock either annually or semiannually. None of the companies surveyed indicated that they never experience bill shock at all, demonstrating the prevalence of this problem throughout the entire industry.

2 41% of companies report high costs as the leading challenge associated with ML/data analytics today

As with data analytics, the cost-performance of ML projects is key to successful business predictions. But given that in ML, the more experimentation a company conducts, the better the final result – it is no surprise that 41% of companies consider the high costs involved in ML experimentation to be the primary challenge associated with ML/data analytics today (Figure 5). This suggests that companies would be able to achieve better ROI from their ML projects by using a platform that can turbo-boost the experimentation process and enable more iterations at a reasonable cost.

3 98% of companies experienced ML project failures in 2023, with poor data cleansing and lackluster cost-performance the primary causes

The top contributing factor to ML project failures in 2023 was insufficient budget (29%), which is consistent with many other findings throughout the report (Figure 6). But aside from the cost concerns, the other top contributing factors to project failures were poor data preparation (19%) and poor data cleansing (19%) – both of which are crucial to the success of ML projects, because they have a direct impact on the number of successful ML iterations that can be achieved within the available project budget. The more inefficient the data preparation and cleansing, the less opportunity there is to maximize the volume of iterations required to achieve optimal results.

4 The top factor impacting data analytics and AI/ML goals in 2024 is “adding more CPU”, but “adding GPU instances to the stack” is rapidly gaining ground

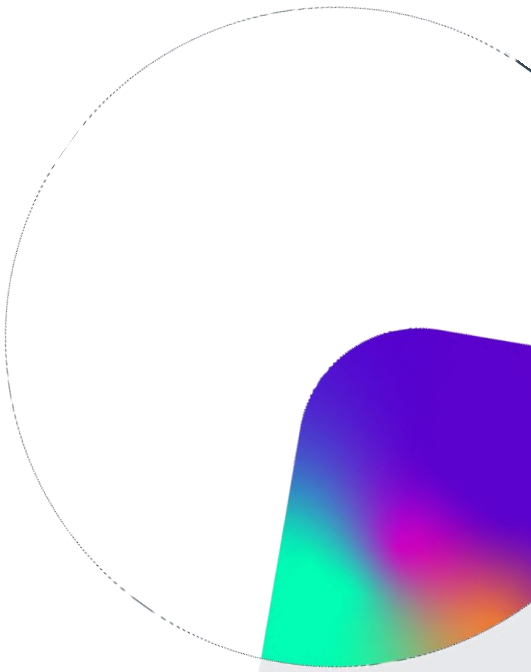
Although 78% of companies say that adding more general compute power (CPU) will have the most impact on their data analytics and AI/ML goals in 2024 (Figure 7), the second most impactful factor is adding GPU instances to their analytics stack (75%). This suggests that a growing proportion of organizations are realizing that CPU-based MPP systems can’t scale infinitely, and that once certain levels of complexity or dataset size are reached, performance will not improve. In other words, three-quarters of the companies surveyed understand that relying on “more compute power to yield better results” indefinitely – is an untenable approach, and that the contribution of GPUs to enterprise AI and analytics is crucial.

5 92% of companies are actively working to "rightsize" cloud spend on analytics, but almost half are doing it by compromising on query complexity, project volume and dataset size

The ability to query all of a company’s data – Including complex questions that can potentially make a significant impact – is at the heart of data-driven businesses. And yet, about half of the respondents admitted they compromise on the complexity of queries (48%) and on the volume of projects (46%) in an effort to manage and control analytics costs – especially in relation to cloud resources and compute loads (Figure 8). In other words, they are presenting stale dashboards and leveraging only semi-reliable data because they are using less data and asking only “simple” questions.

This finding is therefore a wakeup call for companies to find other cost management methods, because compromising on query complexity, project volume and dataset size is clearly far from ideal.

Challenges in Data Management



Number of Data Related Tools, Platforms or Solutions in Use by Task

We asked respondents how many tools, platforms or solutions are currently in use by their organizations for a variety of tasks. For data management, most organizations use two tools (46%), for machine learning/data science most use 3-4 tools (65%), for data preparation/ ETL most use 3-4 tools (42%) and for business intelligence most use 3-4 tools (46%).

Companies typically use multiple tools because it makes sense to use specific tools for certain file formats or workflows, such as batches versus streams. But **using many different tools by many users in parallel can be problematic**, because it is more difficult to find bottlenecks, and therefore more complicated to analyze and accelerate processes. Using too many tools also means that there is no single source of truth, making it more challenging to manage governance and orientation and to ensure that the ecosystem runs smoothly.

When it comes to data preparation, a third of the companies surveyed (33%) are using 5-10 solutions/platforms, which makes this task extremely complicated. It is therefore no wonder that the solution the industry is leaning towards is based on unified lakehouse management capabilities directly on the cloud storage, which allows stakeholders throughout an organization to access the same information more efficiently.

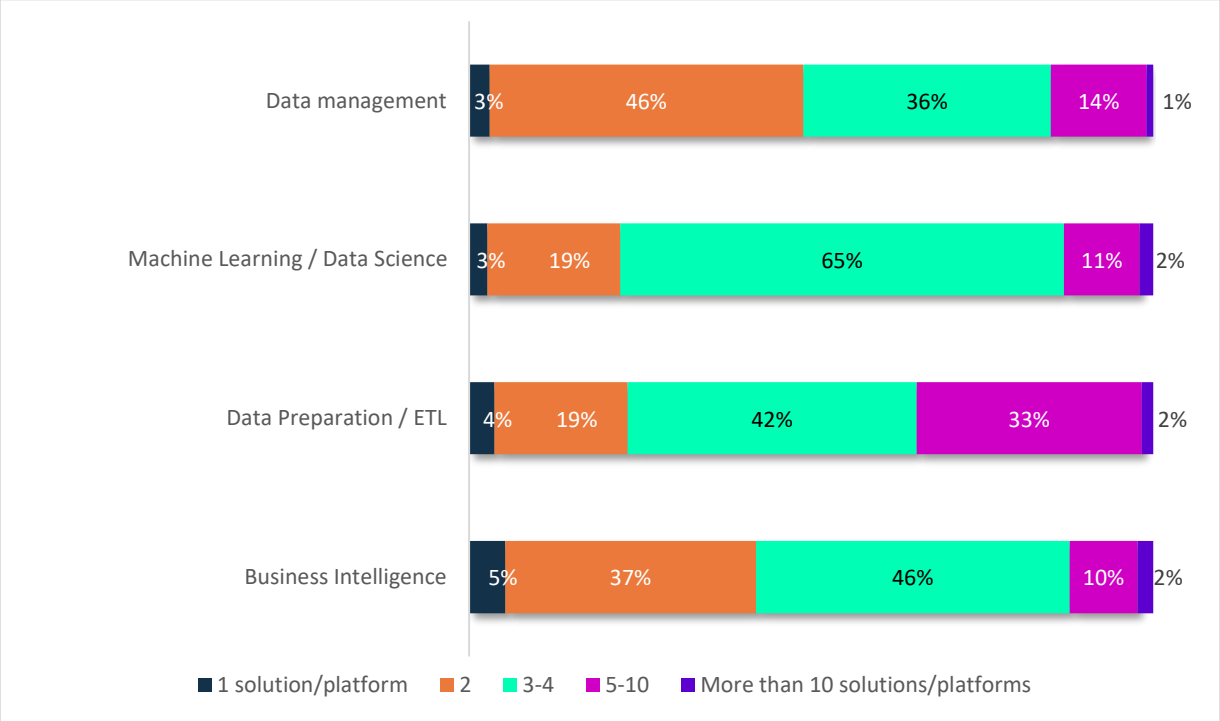


Figure 1: Number of Data Related Tools, Platforms or Solutions in Use by Task

Top Challenges in Handling and Analyzing Data at Scale

In the era of cloud analytics, most challenges can be measured as a cost challenge, in that most issues can be solved by simply paying more. The problem is that the cost of solving these issues can often become exorbitant. It is therefore no surprise that **the top challenge companies experience in handling and analyzing data at scale is the budget/total cost of analytics (27%)**, followed by the time of analytics performance (23%) and scale, i.e. the ability to analyze ALL the data needed (14%).

At the moment, most companies are accustomed to solutions that can provide either great analytics performance at a very high cost, or less-than-great analytics performance at a more reasonable cost. This finding shows, however, that the most important metric for comparing data analytics platforms is cost-performance, where the best results can be achieved in terms of both cost and performance, so that companies do not need to lose out on either aspect.

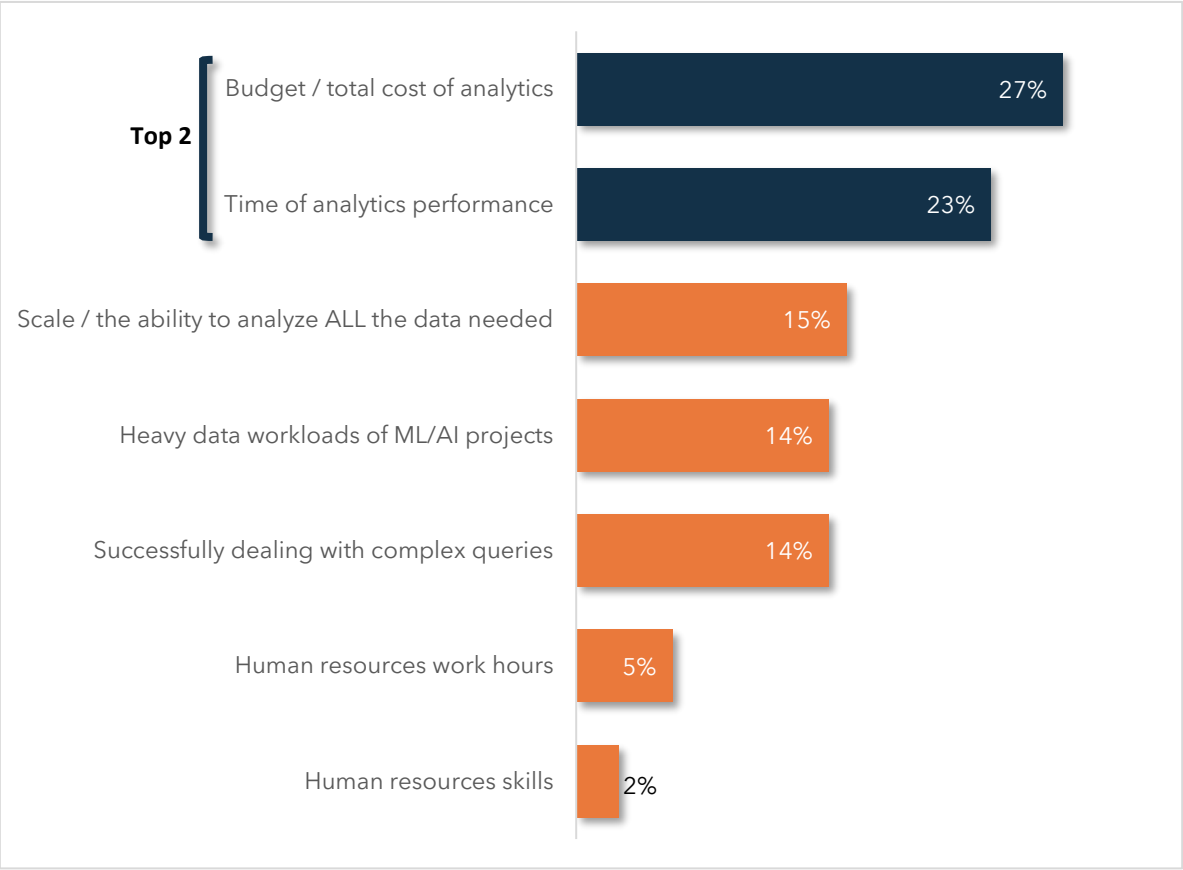


Figure 2: Top Challenges in Handling and Analyzing Data at Scale. The chart illustrates the distribution of these challenges as a percentage of the total, summing up to 100%.

Frequency of Data Analytics "Bill Shocks" of Jobs, Queries and Workloads

“Bill shocks” (relating to data analytics processes) occur when data workflows are either too complex or too big for the existing query engine, because the more compute power is needed, the higher the associated costs. We asked respondents how often they experience bill shock, or in other words, how often they are surprised by the high costs of cloud-based data analytics.

71% of respondents (2 out of 3 companies) reported they are surprised by their cloud analytics bill fairly frequently, with 5% experiencing bill shock monthly, 25% bimonthly and 41% quarterly. That’s pretty significant, considering that most of them – including those who have contracts with the likes of Google, Amazon or other cloud vendors – are not likely to be paying the highest on-demand prices anyway, and using cost visibility tools in an attempt to reduce costs. Almost a third of the companies surveyed (29%) say they experience bill shock either annually or semiannually. None of the companies surveyed indicated that they never experience bill shock at all, demonstrating the prevalence of this problem throughout the entire industry.

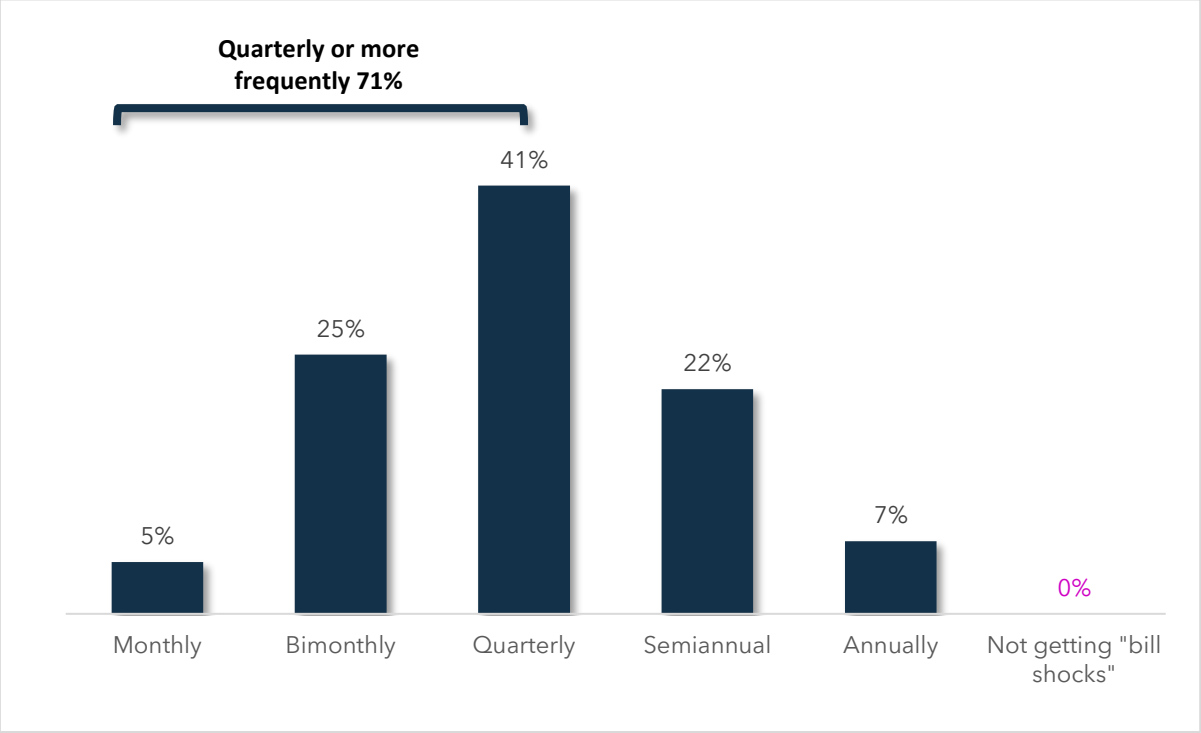


Figure 3: Frequency of Data Analytics "Bill Shocks" of Jobs, Queries and Workloads

Top Areas of Dissatisfaction with Current Data Stack

When asked to rank their level of satisfaction with a variety of aspects relating to their data stack, **respondents reported that their top area of dissatisfaction is the total cost of analytics (27%)**, which also aligns with their top challenge in handling and analyzing data at scale, as seen in Figure 2.

That said, respondents are dissatisfied more or less evenly with other aspects of their data stack such as speed of analytics (25%), the volume of ML projects that make it to production (23%), the ability to analyze all data needed (23%) and to solve complex data queries (20%). The fact that there is a similar level of dissatisfaction with most aspects of their data stack suggests that on the whole, companies would benefit from wide-ranging improvements to boost their overall level of satisfaction.

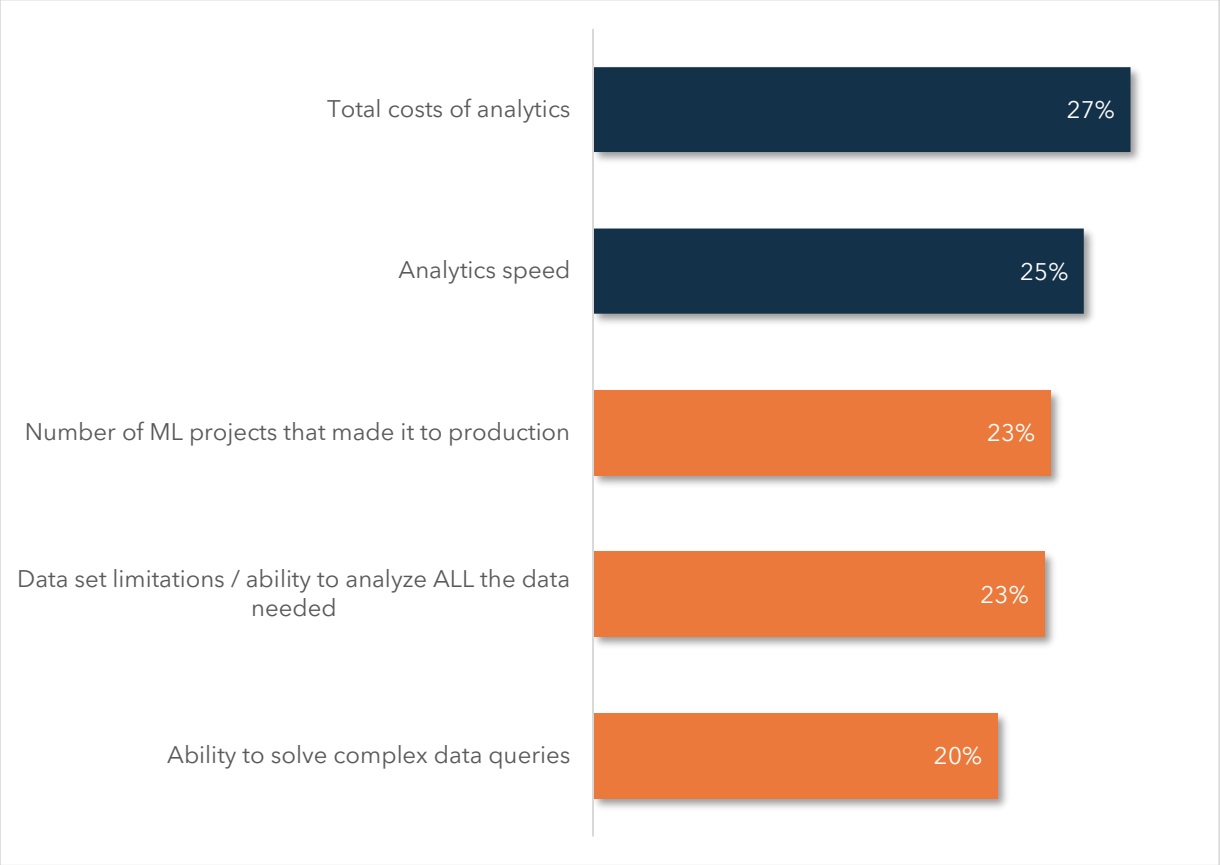


Figure 4: Top Areas of Dissatisfaction with Current Data Stack

*Question allowed more than one answer and as a result, percentages will add up to more than 100%

Top Challenges Pertaining to ML/Data Analytics

ML and data analytics are both crucial for organizations to sustain business progress and optimization of their products. In ML, the general rule of thumb is that the more experimentation a company conducts, the better the final result. This involves a great deal of fine-tuning, which is why it's so important to test the results of the various fine-tuning iterations as quickly as possible. The more iterations are tested, the better the resulting business predictions tend to be. So, as with data analytics, the cost-performance of ML projects is key to successful business predictions.

It is therefore no surprise that **companies consider the high costs involved in ML experimentation to be the primary disadvantage of ML/data analytics today (41%)**, followed by the unsatisfactory speed of this process (32%), too much time required by teams (14%) and poor data quality (13%).

This key finding reinforces the notion that companies can achieve better ROI from their ML projects by using a platform that can turbo-boost the process at a reasonable cost.

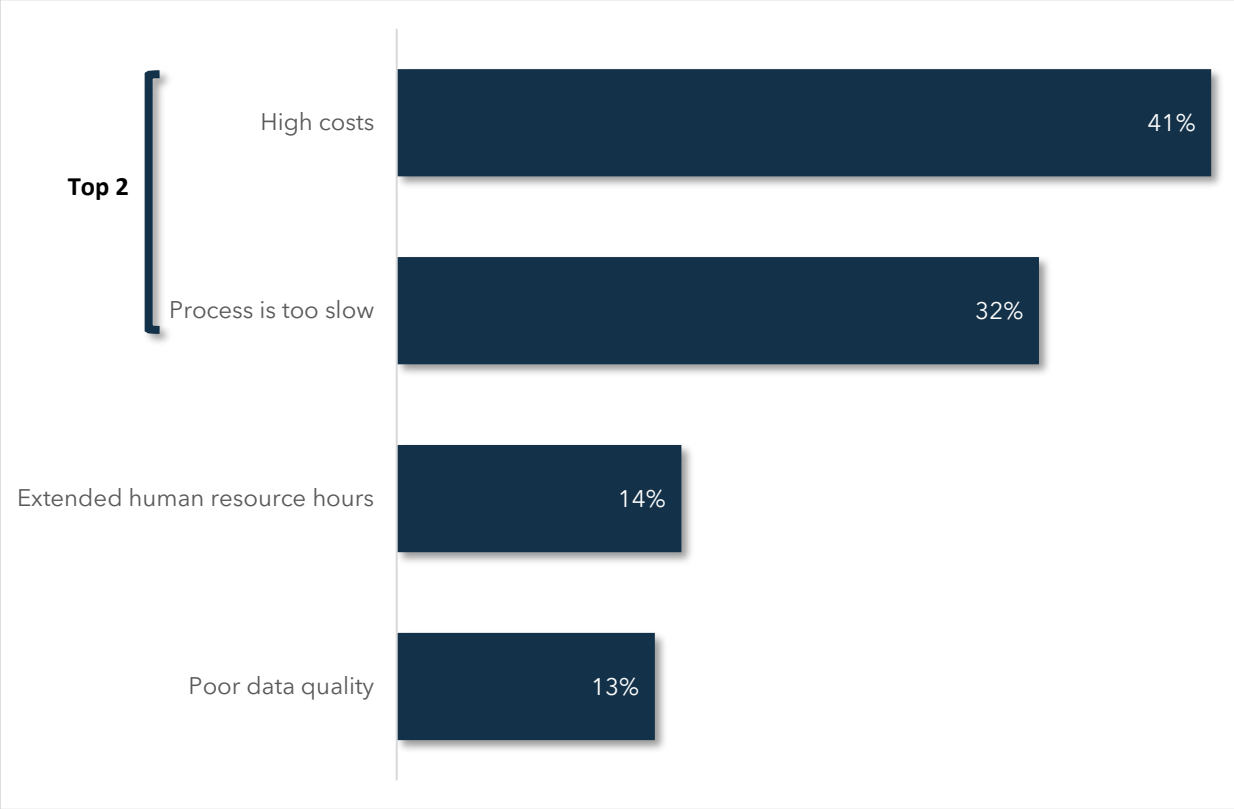


Figure 5: Top Challenges Pertaining to ML/Data Analytics. The chart illustrates the distribution of these challenges as a percentage of the total, summing up to 100%.

Top Factors Contributing to the Failure of ML Projects in 2023

The top contributing factor to ML project failures in 2023 was insufficient budget (29%), which is consistent with previous findings – including the fact that "budget" is the top challenge in handling and analyzing data at scale (Figure 2), that more than two-thirds of companies experience "bill shock" around their data analytics processes at least quarterly if not more frequently (Figure 3), that that the total cost of analytics is the aspect companies are most dissatisfied with when it comes to their data stack (Figure 4), and that companies consider the high costs involved in ML experimentation to be the primary disadvantage of ML/data analytics today (Figure 5).

Interestingly, aside from cost concerns, **the other top contributing factors to project failures were poor data preparation (19%) and poor data cleansing (19%) – both of which are crucial to the success of ML projects.**

Data preparation typically involves data preprocessing (including cleaning, integration, transformation, and reduction) and data wrangling (including filtering, grouping, enhancing features, and enhancing accuracy). Poor data preparation and cleansing can therefore impact the number of successful ML iterations that can be achieved within the available budget, and ultimately contribute to project failure.

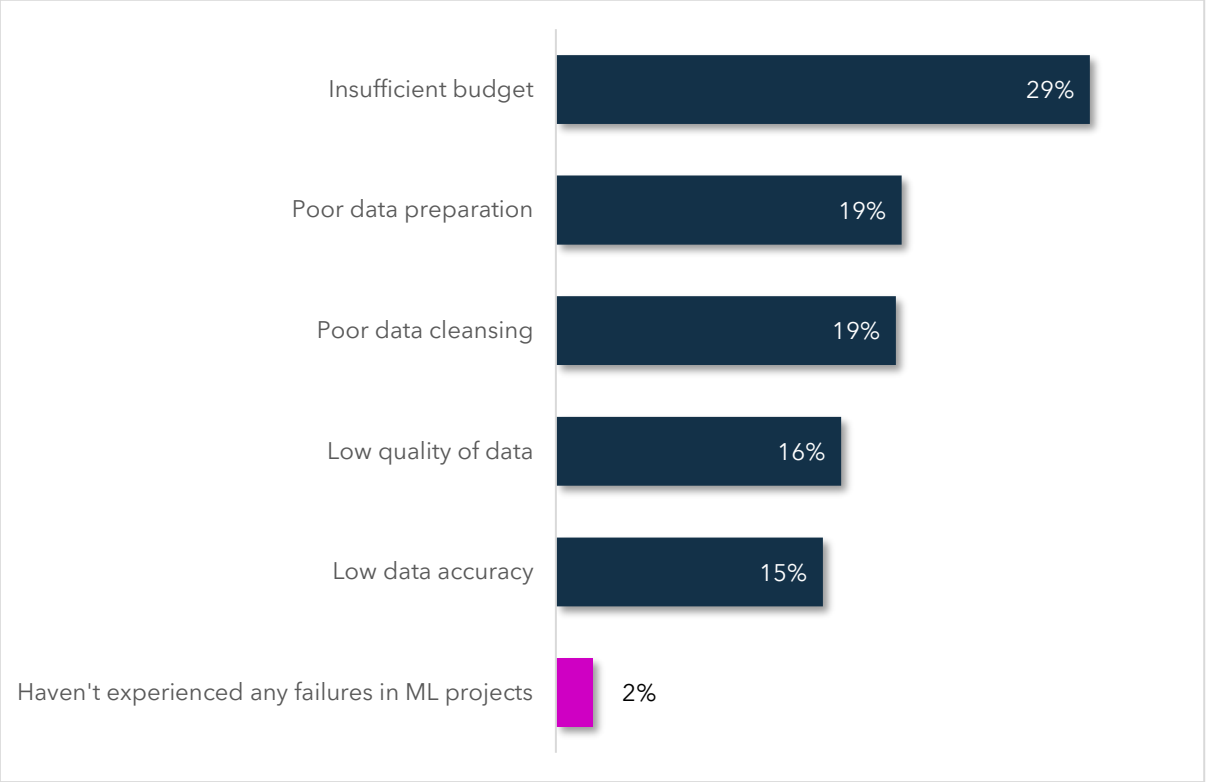


Figure 6: Top Factors Contributing to the Failure of ML Projects in 2023. The chart illustrates the distribution of these challenges as a percentage of the total, summing up to 100%.

Top Factors Impacting Data Analytics and AI/ML Goals in 2024

We asked respondents to weigh in on various factors that would have the most impact ("extremely high" or "high") on their data analytics and AI/ML goals in 2024. 78% say that the top factor is adding more general compute power (CPU), followed closely by adding GPU instances to their company's stack (75%).

Although many companies still seem to rely on the notion that "more compute power will yield better results", this tactic isn't tenable indefinitely. CPU-based MPP systems can't scale infinitely, and once certain levels of complexity or dataset size are reached, performance will not improve.

GPUs would therefore have a much more substantial impact, as they are great not only for AI/ML but can also be used for analytics, and this key finding confirms that organizations are well on their way to realizing this.

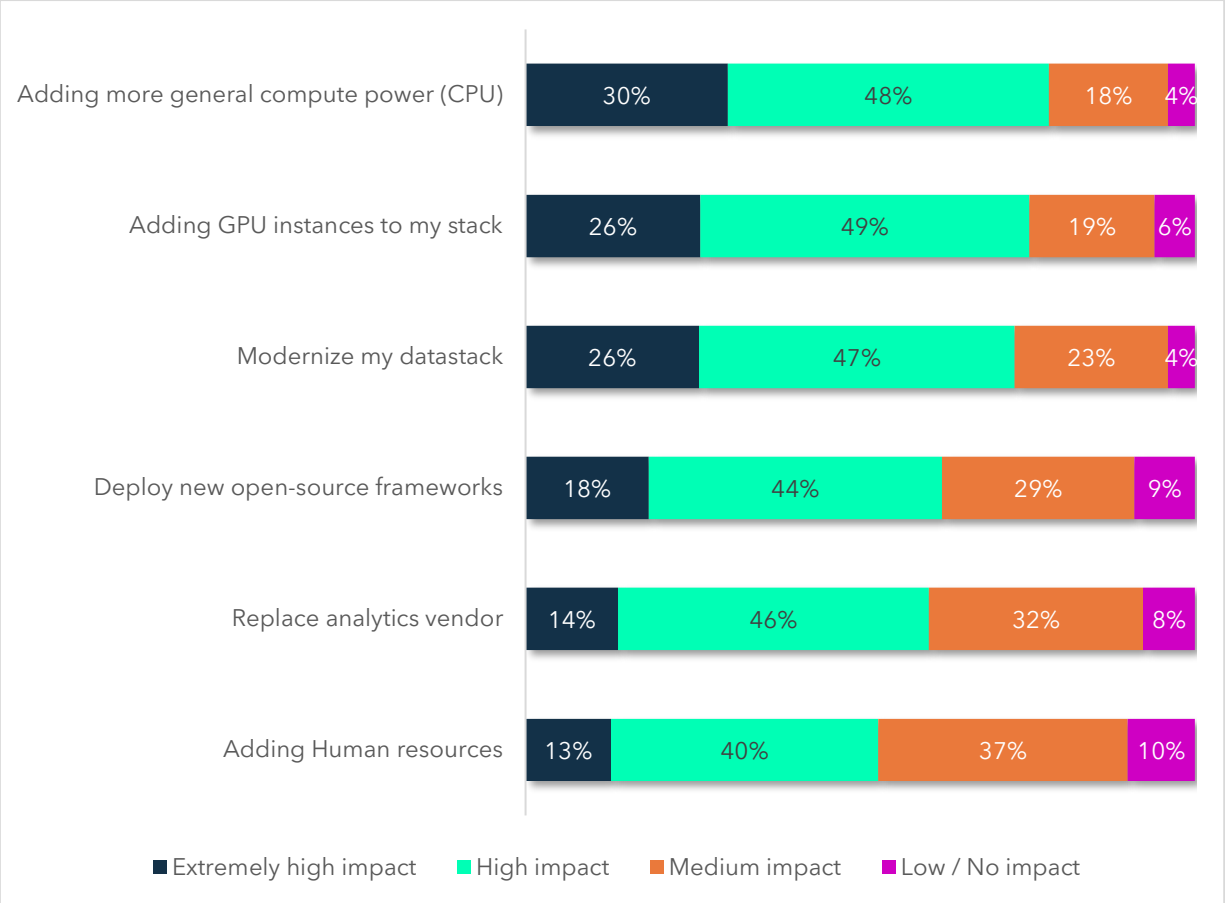
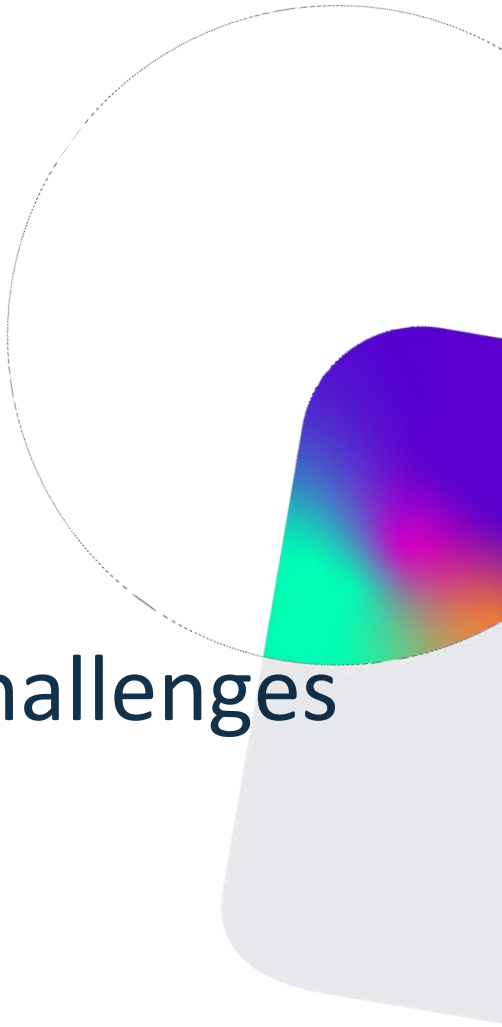


Figure 7: Top Factors Impacting Data Analytics and AI/ML Goals in 2024



Methods to Address Challenges

Top Methods Used to Manage Analytics Costs

We asked the respondents what methods were used in their organization at least once in the past 12 months to manage and control analytics costs, especially in relation to cloud resources and compute loads. Based on their responses, it's clear that companies are used to making compromises in order to manage their costs, with about half admitting they compromise on the complexity of queries (48%) and on the volume of projects (46%).

The ability to query all of a company's data – Including complex questions that can potentially make a significant impact – is at the heart of data-driven businesses. These responses indicate, however, that many companies are compromising on everything that's important (if not critical) in big data analytics, and using data purely as a growth and revenue accelerator. In other words, they are presenting stale dashboards and leveraging only semi-reliable data because they are using less data and asking only "simple" questions.

This key finding is therefore a wakeup call for companies to find other cost management methods, because compromising on query complexity, project volume and dataset size is clearly far from ideal.

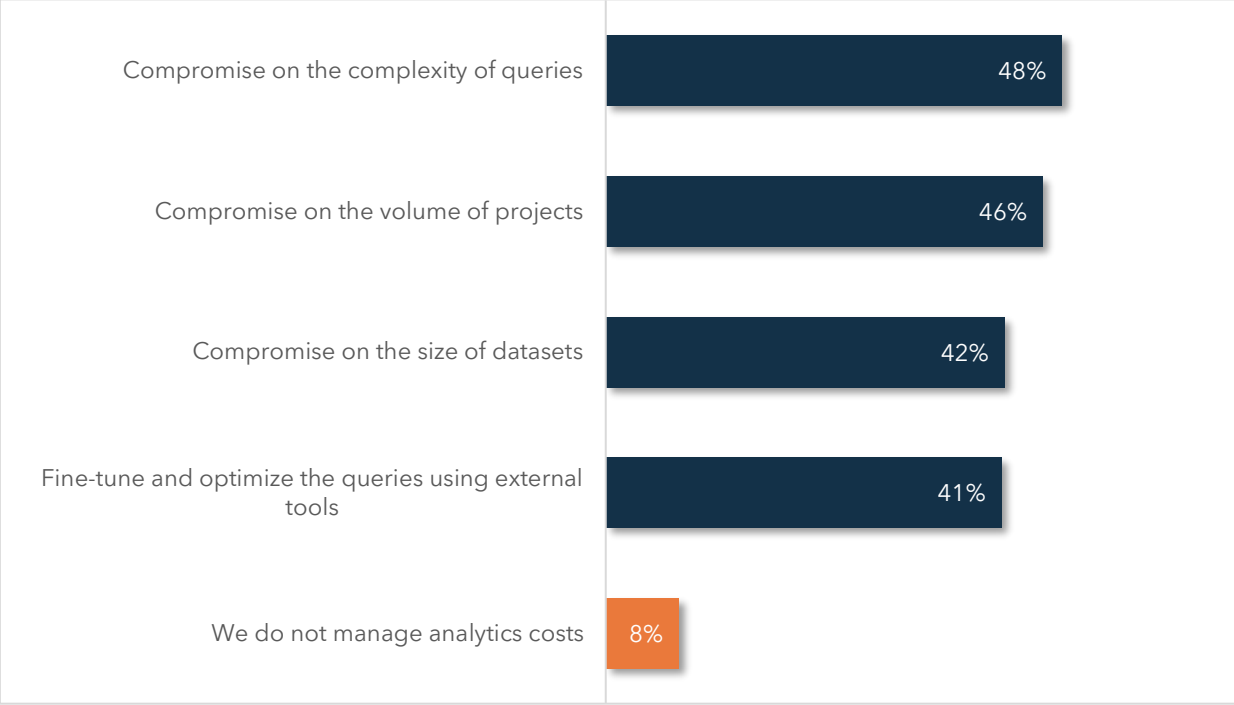


Figure 8: Top Methods Used to Manage Analytics Costs

*Question allowed more than one answer and as a result, percentages will add up to more than 100%

Priority of Optimizing Existing Data Pipelines in Roadmap for 2025

Data is typically inaccessible and not 'workable' unless it goes through a certain level of transformation. In fact, since different departments within an organization have different needs, it is not uncommon for the same data to be prepared in various ways. Data preparation pipelines are therefore the foundation of data analytics and ML, so **we asked respondents to what extent their companies are prioritizing the optimization of existing data pipelines in their roadmap for 2025. Unsurprisingly, around two-thirds (63%) reported that it is either a high or very high priority**, and around a third reported that it is of moderate priority (34%). Only 3% think of optimizing existing data pipelines in their roadmap for 2025 as a low priority.

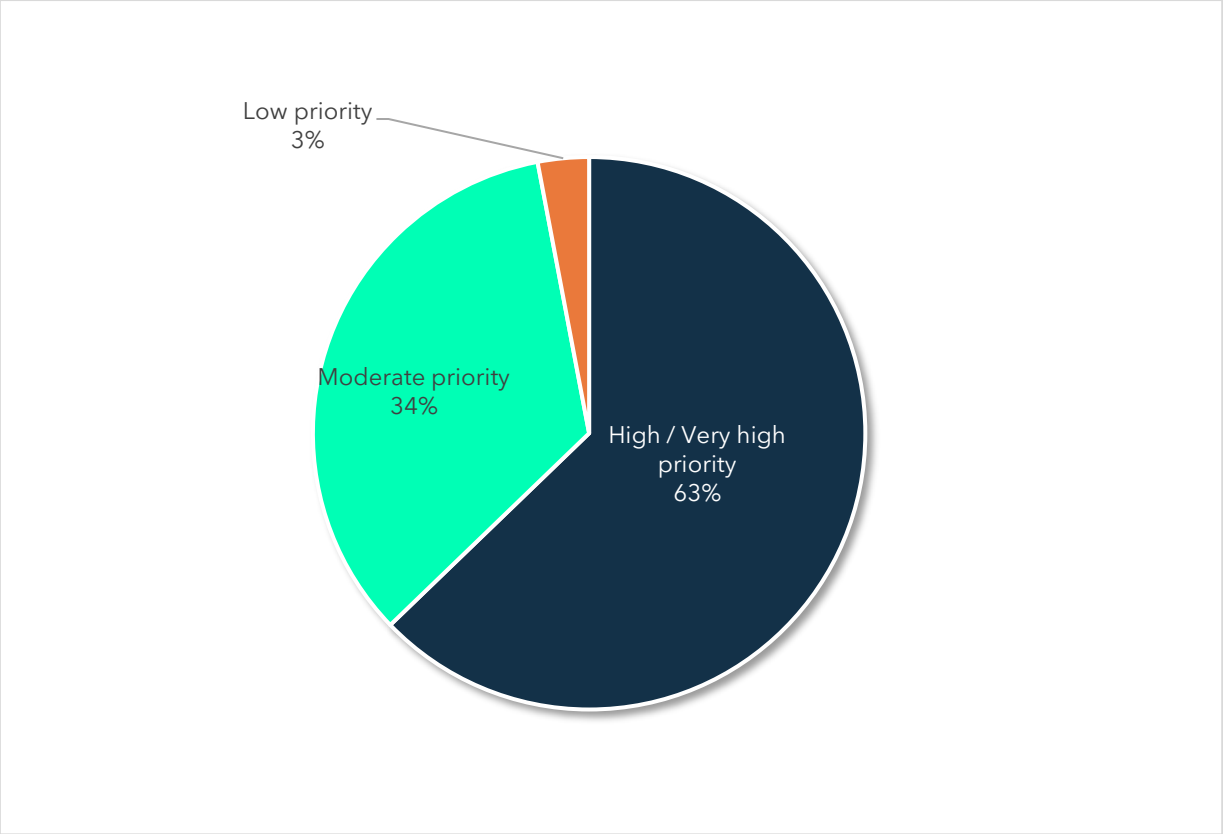


Figure 9: Priority of Optimizing Existing Data Pipelines in Roadmap for 2025

Top Expected Trends in Big Data Analytics, 2025-2028

Given that the survey's participants included visionary leaders in the data analytics industry – who are often all about "the next big thing" – we were keen to investigate their perspective on expected trends in big data analytics over the next few years. **The top expected trends identified by the respondents are: edge computing (48%), quantum computing (46%) and AI/ML learning integration (35%).**

These results indicate that although Generative AI and LLMs are the current "big thing" – having attracted a great deal of hype very rapidly – they are likely to be overtaken by edge computing and quantum computing, which may not have attracted as much hype (yet) but have actually been expected to be of significant importance in big data analytics for some time.

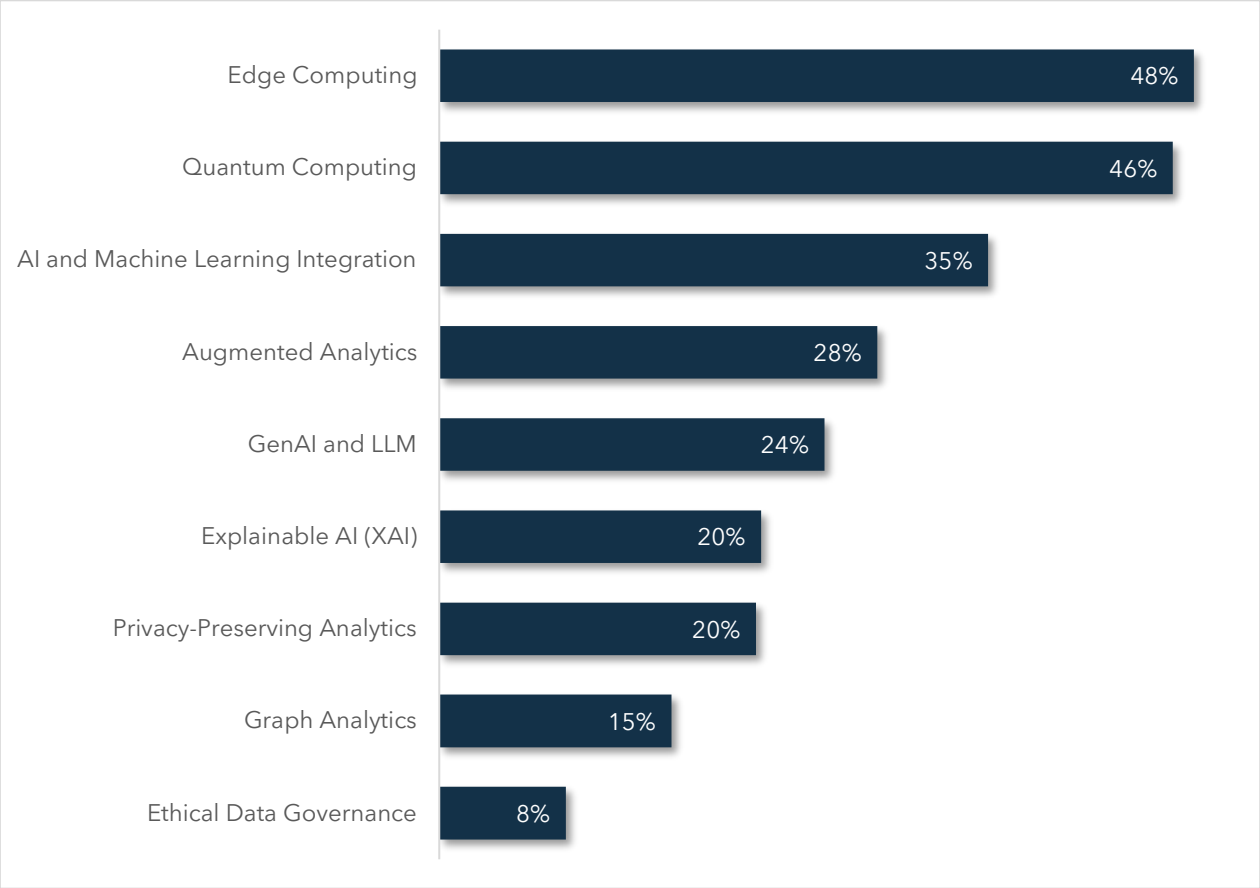
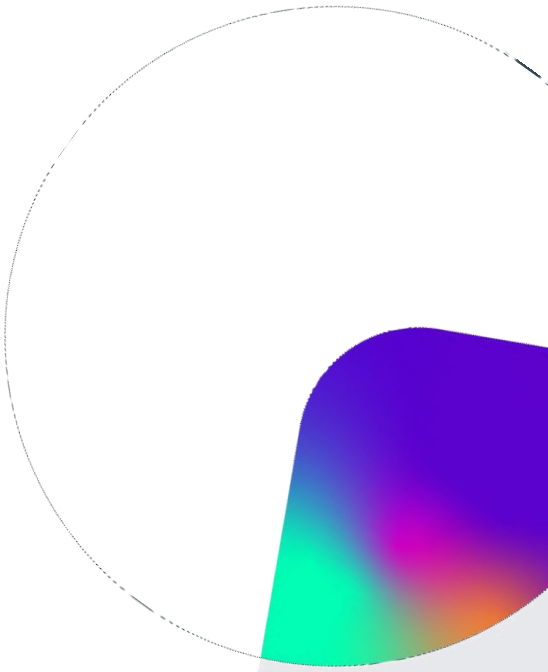


Figure 10: Top Expected Trends in Big Data Analytics, 2025-2028

*Question allowed more than one answer and as a result, percentages will add up to more than 100%

Demographics



Industry, Department, Job Seniority, Role and Size of Data in TB

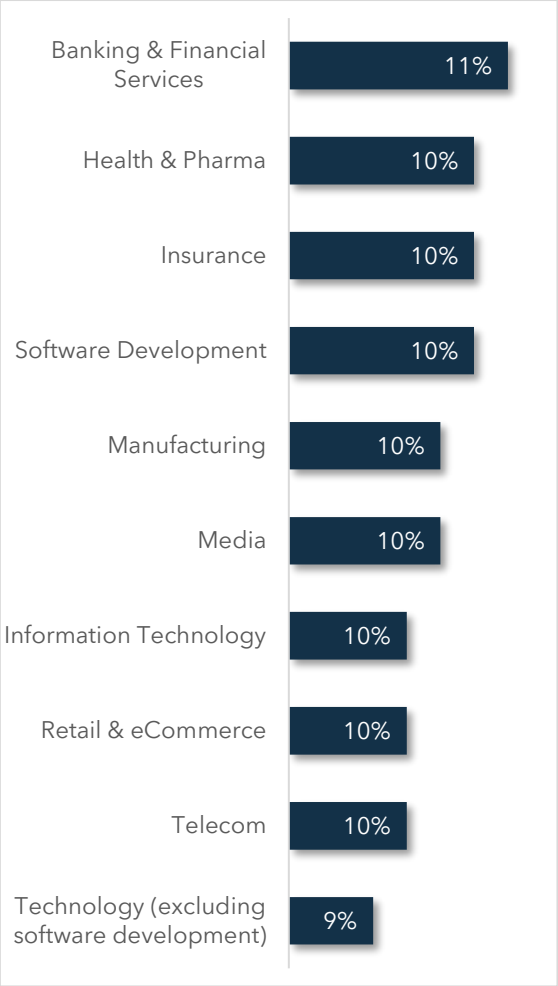


Figure 11: Industry

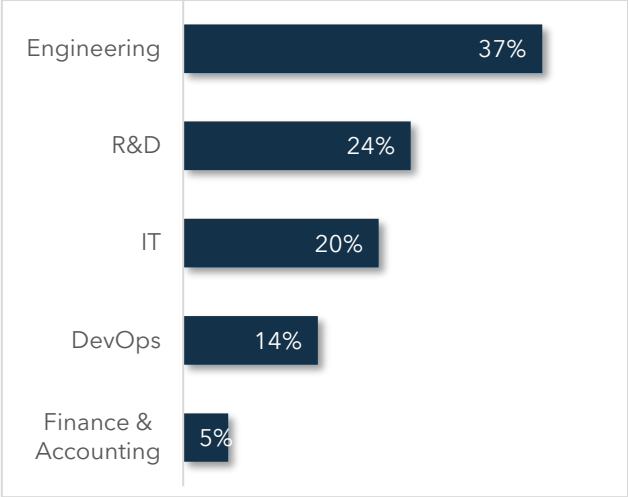


Figure 13: Department

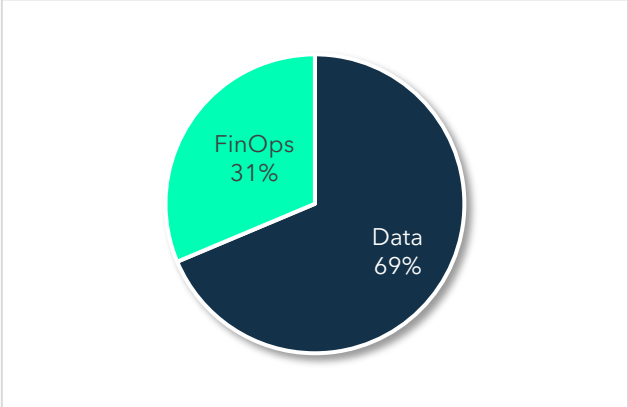


Figure 14: Role

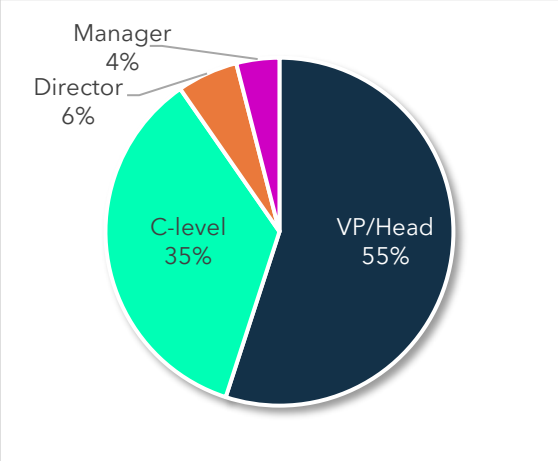


Figure 12: Job Seniority

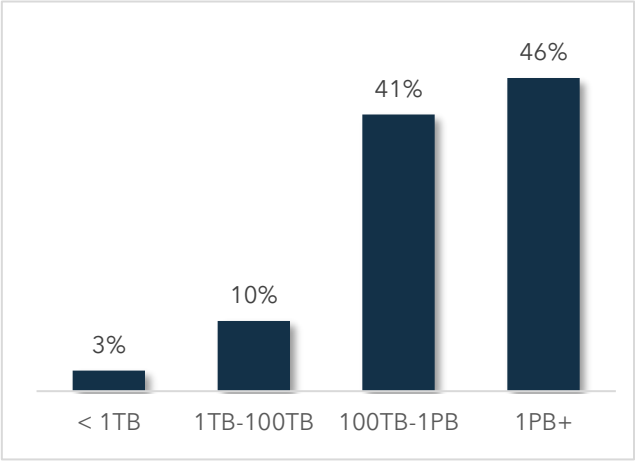


Figure 15: Size of Data in TB

Annual Spend on Cloud Computing / Data Analytics, Main Cloud Computing Providers

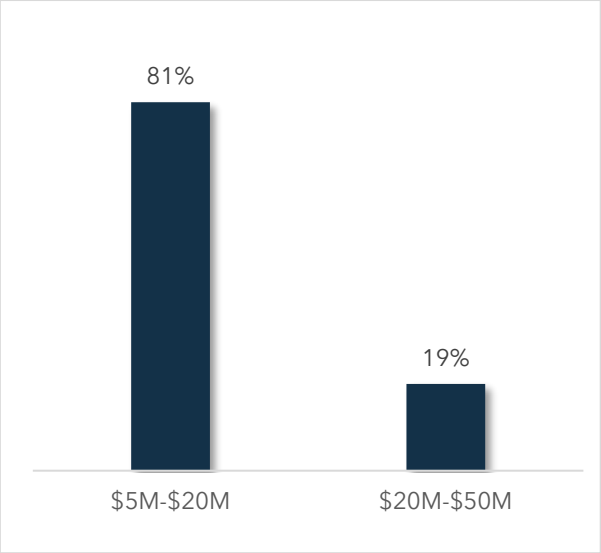


Figure 16: Annual Spend on Cloud Computing

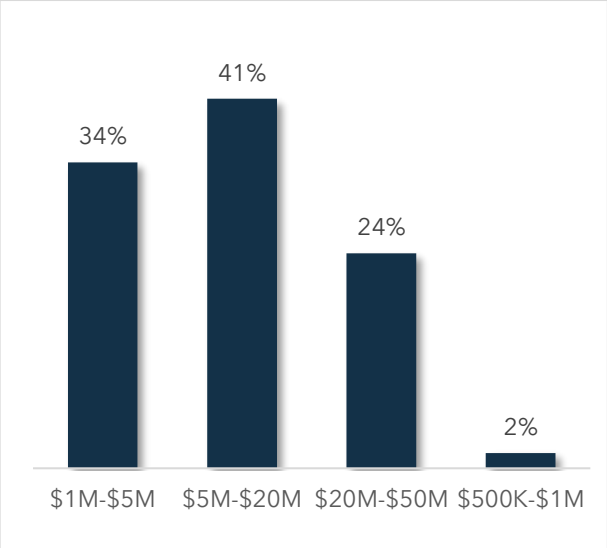


Figure 17: Annual Spend on Data Analytics Services in the Cloud

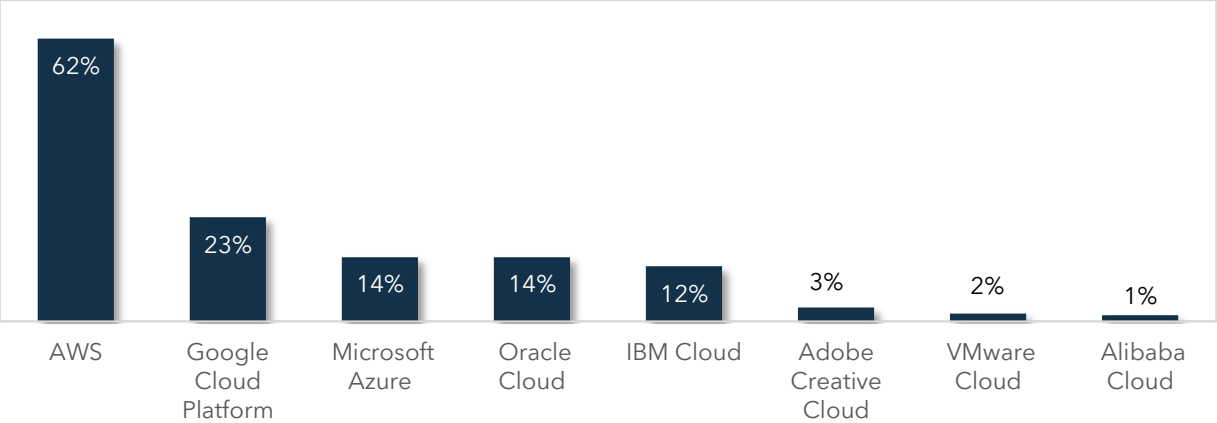


Figure 18: Main Cloud computing Services in Use

About SQream

SQream empowers companies to get value from their data that was unattainable before at an exceptional cost performance. Our data processing and analytics acceleration platform utilizes a GPU- patented SQL engine that accelerates the querying of extremely large and complicated datasets.

By leveraging SQream's advanced supercomputing capabilities for analytics and machine learning, enterprises can stay ahead of their competitors while reducing costs and improving productivity.

Request a Demo

For more information, please visit us:

Global Headquarters (US): +1-877-878-4081 | R&D Center (Israel): +972-3-544-4871