CHAOSSEARCH

2022 CLOUD DATA AND ANALYTICS SURVEY REPORT

Making Sense of Information Overload with Modern Data Lakes

Data lakes built on dated infrastructure are costing businesses critical time-to-value. It's time to ditch them.

INTRODUCTION

A business's secret weapon is the data already at its disposal, but converting data into actionable insights is increasingly complicated.

Especially as the pandemic has shifted much of the working world to an almost entirely digital model, businesses are sitting on vast quantities of data— it's either a goldmine or a sinkhole, depending on how they manage it.

In theory, this increase in data should present new opportunities for organizations. But the reality of current infrastructures makes it challenging to retain and analyze multi-structured datasets at scale. That, coupled with the prevalent data scientist and data engineer shortage, is leaving enterprises hamstrung on their quest for data-driven insights that will impact the business.

To better understand the challenges organizations face when it comes to accessing and analyzing data, ChaosSearch commissioned a survey of 1,020 U.S. IT professionals. The survey asked questions about their organizations' data retention, data usage, and investments in data lake and cloud data platforms. This report summarizes our findings.

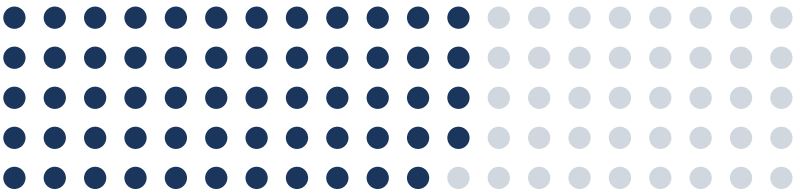
At the highest level, we found that many organizations are making strides in how they govern data management and take action on data-driven insights. However, IT talent is still often wasted moving, migrating, pipelining, and transforming data—a process that can, and should, be automated if the end goal is faster, more insightful analytics.

KEY FINDINGS

IT professionals realize they have a data analytics problem on their hands—and they're looking to technology to remedy it.

When asked to choose between more time, budget, people, or technology to better solve existing analytics challenges, more than half (59% of respondents selected technology—whereas no more than 13% of respondents chose time, budget, or people, respectively.

In particular, investments in data lakes are on the rise. A majority of respondents (69%) indicated that their organizations have implemented a data lake, while 23% of respondents are planning to deploy one but haven't done so yet.



59%

prioritize investments in technology over time, budget, or people.

IT talent is wasted on data prep–even more so in organizations with data lakes.

Respondents spend almost as much time prepping data (6.6 hours per week) as they do analyzing it (7.2 hours per week)—and data lakes aren't helping. Time spent moving, migrating, pipelining, or transforming data increased to 7.1 hours per week for respondents who have a data lake.

Additionally, 30% of respondents indicate that their end-user consumption/visualization tools aren't directly connected to the lake, resulting in data duplication and data movement challenges limiting insights and time to value.

But there are benefits to investing in data lakes.

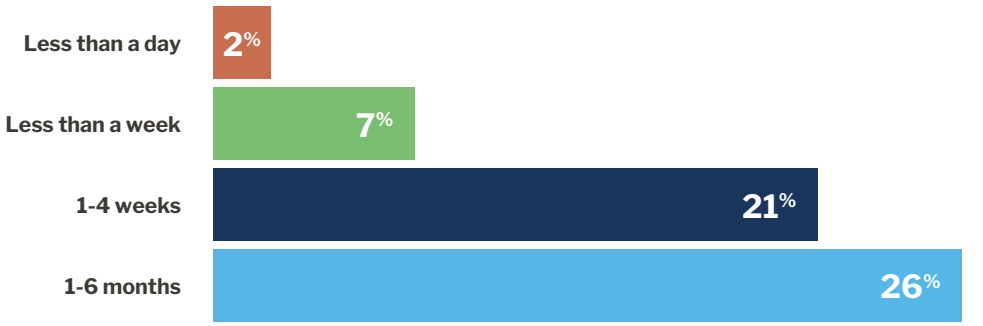
Thirty-eight percent of respondents with data lakes are able to respond to data requests within an hour, compared to 24% without data lakes. Additionally, 87% of respondents using data lakes indicate an improved ability to make organizational decisions.

There's a disconnect on whether organizations are using all the data available to them—and it's leaving them vulnerable.

Eighty-seven percent of respondents agree to some extent that their department is using all the data at its disposal to make informed business decisions, but those who agreed with the statement are retaining less log data (only 1-6 months) than those who disagreed (7-12 months).

That's an issue considering ChaosSearch's Log Management and Analytics 2021 Benchmark Report found that the top two use cases for leveraging log data are security (70%) and IT monitoring (68%).

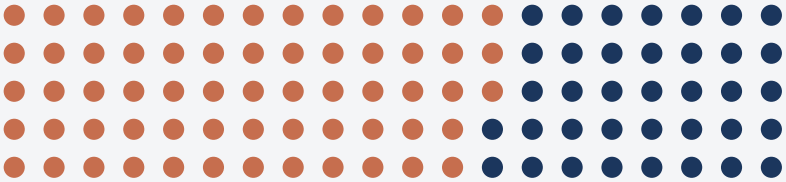
HOW LONG DOES YOUR ORGANIZATION RETAIN LOG DATA FOR, ON AVERAGE?



Saving data for so short a time to make room for new data won't make the data lake useful in analyzing and preventing cyber security breaches, since users won't be able to proactively identify vulnerabilities in their systems that could have been prevented sooner. Likewise, organizations that continually refresh their data stores will never build up a wide breadth of data, making any data analysis shallow in what it reveals about the organization. Only once they acquire unlimited data storage can organizations truly be using all the data available to them. Yet less than a quarter (24%) of respondents are retaining log data indefinitely.

When you have an easily accessible data lake capable of identifying vulnerabilities before they become outright threats, your organization can stay ahead of malicious players. For instance, whereas 47% of respondents who retain less than 7 months of log data have experienced a breach in the last year, only 24% of respondents who retain 7+ months of log data experienced a breach in the last year. Now that's data put to good use.

HAS YOUR ORGANIZATION EXPERIENCED A SECURITY BREACH IN THE LAST YEAR?



- Yes = 63%
- No = 37%

Data lake investment is on the rise—but not all data lakes live up to their promise.

Data is the lifeblood of a business. It captures a historical picture of the business while revealing opportunities for future growth. The more data your organization can leverage, the more it can innovate and flourish. Many organizations recognize the value of robust data analytics. They also recognize the struggle to glean analytical insights—but are working to correct this by increasing their investments in technology. In fact, when asked to choose between investing more time, budget, people, or technology to better solve existing analytics challenges, more than half (59%) of respondents selected technology—whereas no more than 13% of respondents chose time, budget, or people, respectively.

Unfortunately, not all tech tools are the answer to analytics challenges. In order for data to live up to its full potential, you need to be able to access your data quickly, but traditional data management systems and processes put up walls that prohibit this.

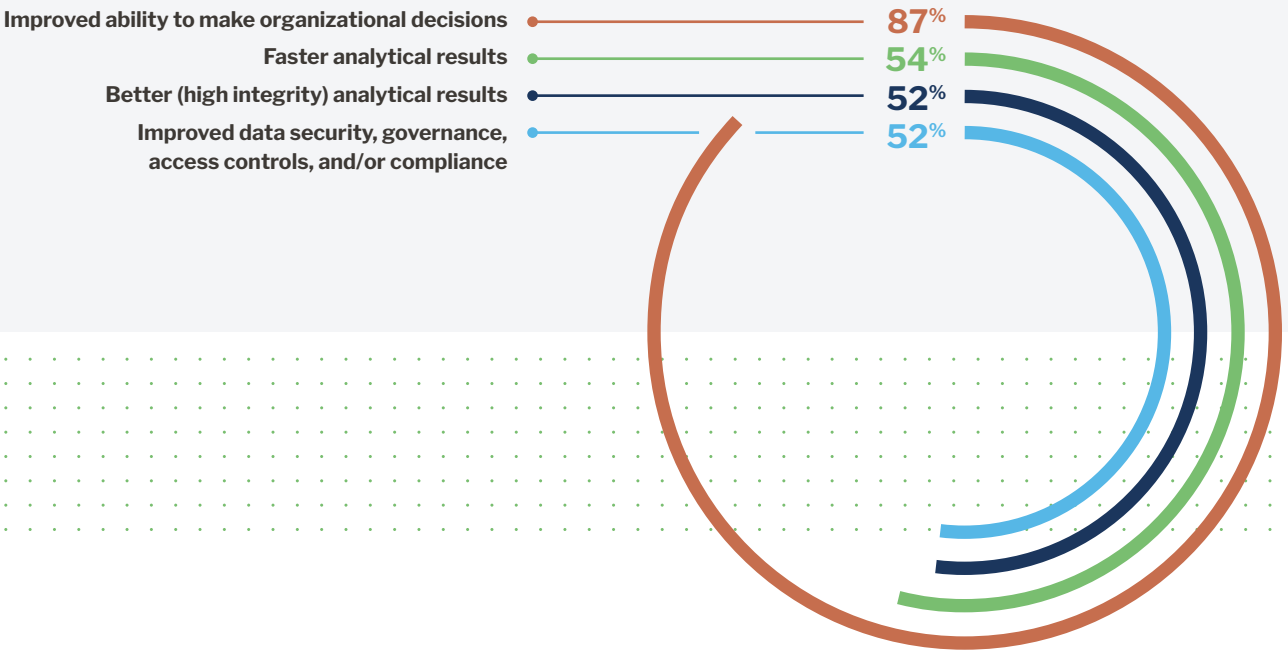
What's wrong with traditional systems? Data warehouses, data marts, and legacy databases only accept structured data, which makes them less flexible and scalable. In practice, organizations using these types of solutions end up accruing multiple of them to accommodate their multi-structured datasets, which makes it more difficult to source and analyze later.

Data lakes can cut through these siloes by ingesting all data types in their source formats. This gives organizations the ability to—in theory—store vast quantities of data from multiple sources in a centralized, secure, and relatively inexpensive location that scales with ease. With this data all in one place, end users like data analysts, business leaders, and others can quickly access the data they need to inform business decisions.

If your organization is like one of the 31% of respondents that hasn't adopted a data lake yet, you're limiting your data's potential. Respondents that have implemented data lakes are able to respond to data requests faster; 38% of respondents with data lakes are able to respond within an hour, compared to 24% without data lakes. Additionally, 20% of respondents without data lakes take days or weeks to respond, compared to 9% of those with data lakes.

69% of organizations have already implemented a data lake—and 87% indicate that the data lake has improved their organization's ability to make decisions.

WHAT BENEFITS HAVE YOU WITNESSED FROM USING A DATA LAKE?



Challenges surrounding data lake accessibility prevent them from fulfilling their promise.

Data lakes are supposed to deliver more organized, accessible data to support analytics—but not all data lakes do. The fact that the data is stored in its original form is one of the major benefits of data lakes, but it can also be one of its drawbacks when using immature platforms. Traditional data lakes aren't built for high performance analytics and therefore require too much data transformation upfront for it to be used for driving insights, which eats up data engineers', analysts', and scientists' bandwidth and leads to a longer time-to-decision. In fact, organizations with data lakes spend more time prepping data than those without (7.1 hours per week versus 6.6)—indicating that their lakes are requiring more data transformation to work than they should. The time spent prepping data can quickly undercut the effectiveness of the data's analysis; when insights are delayed because data isn't ready yet, businesses risk missing critical revenue opportunities.

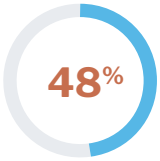
The most common timeframe it takes for IT professionals to respond to data requests from stakeholders is between 2 and 12 hours (37% of respondents). It takes 30% of respondents days or even weeks to respond.

Organizations with data lakes spend nearly as much time prepping data (7.1 hours/week) as they do analyzing it (7.2 hours/week).

Plus, if IT teams are spending as much time prepping data as they are analyzing it, it's unlikely they're using all their data because it's a burden to access more than what's needed for the immediate task. Only half (52%) of respondents say that either all or almost all their company's departments are applying analytics or BI to work they do on a weekly basis, meaning employees aren't making as informed decisions as they could be.

Again, one draw to the data lake is its seemingly limitless capacity, but with so much information for analysts to parse through, an organization's data lake can quickly devolve to a data swamp. Manual extraction isn't only time consuming—it's also costly. In addition to the cost of analysts' time, maintaining these data lakes can become sneakily expensive as data volumes scale, especially without the right in-house talent to maintain these solutions, since most platforms charge once users exceed their limited storage.

NOT EVERYONE IS EXPERIENCING IMPROVED ANALYTICS CAPABILITIES FROM THEIR DATA LAKE.



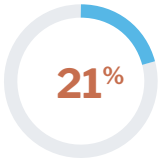
48% find it too hard or time-consuming to get data into and out of the lake

24% don't have enough data science or data engineering resources to support data analytics demands

30% don't have their end-user consumption/visualization tools directly connected to the lake, resulting in data duplication and data movement challenges

22% find it too hard to scale different types of data sets for use

29% struggle to operationalize/ take action on insights from the data lake

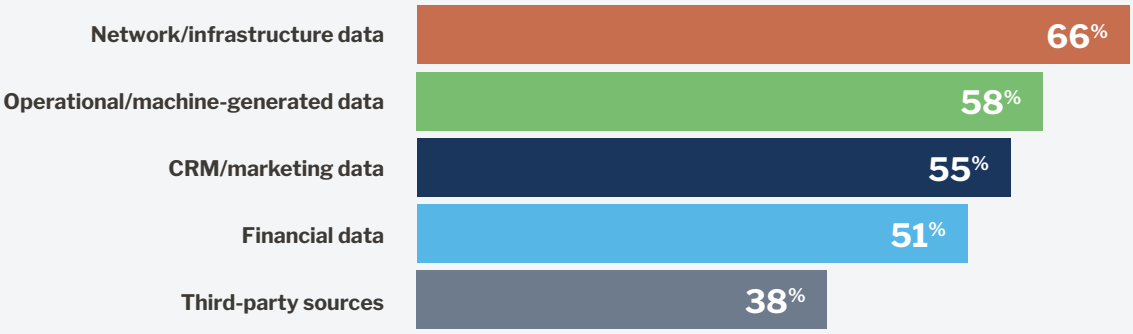


21% have a data lake that doesn't support multi-dimensional analytics use cases, such as full text search, relational analytics, machine learning, etc.

26% can't get access to the data they need to analyze in the lake

13% find it too expensive to scale data sets for use

WHAT TYPES OF DATA ARE DATA LAKE USERS ANALYZING TO INFORM BUSINESS DECISIONS?



CONCLUSION

How to overcome analytics challenges

Your organization needs an easy, cost-effective, and reliable way to access and analyze all the data at its disposal—a standard most data lakes aren't living up to.

The optimal solution should remove data silos and data pipelining requirements so that your organization's data lake(s) caters to any and all use cases, instead of a limited range of needs. Modern data lakes achieve this through, first and foremost, a cloud-based infrastructure that not only enables users to scale with greater ease, but allows users to couple their platform with other cloud-based analytics services and downstream software tools to deliver data indexing, transformation, querying, and analytics functionality. That's important; simply having a data lake doesn't ensure insights—you need to activate the lake with the proper tools and functionalities to derive valuable data analysis.

WITH THE RIGHT PLATFORM AND SURROUNDING TOOLING IN PLACE, DATA LAKES CAN:

- **Centralize data analytics to enhance governance**
- **Enhance log analytics to uncover potential security threats**
- **Cross-pollinate between relational analytics and search on the same unified dataset to generate new, deeper insights**
- **Improve access to more data to feed BI and ML analysis to uncover trends**

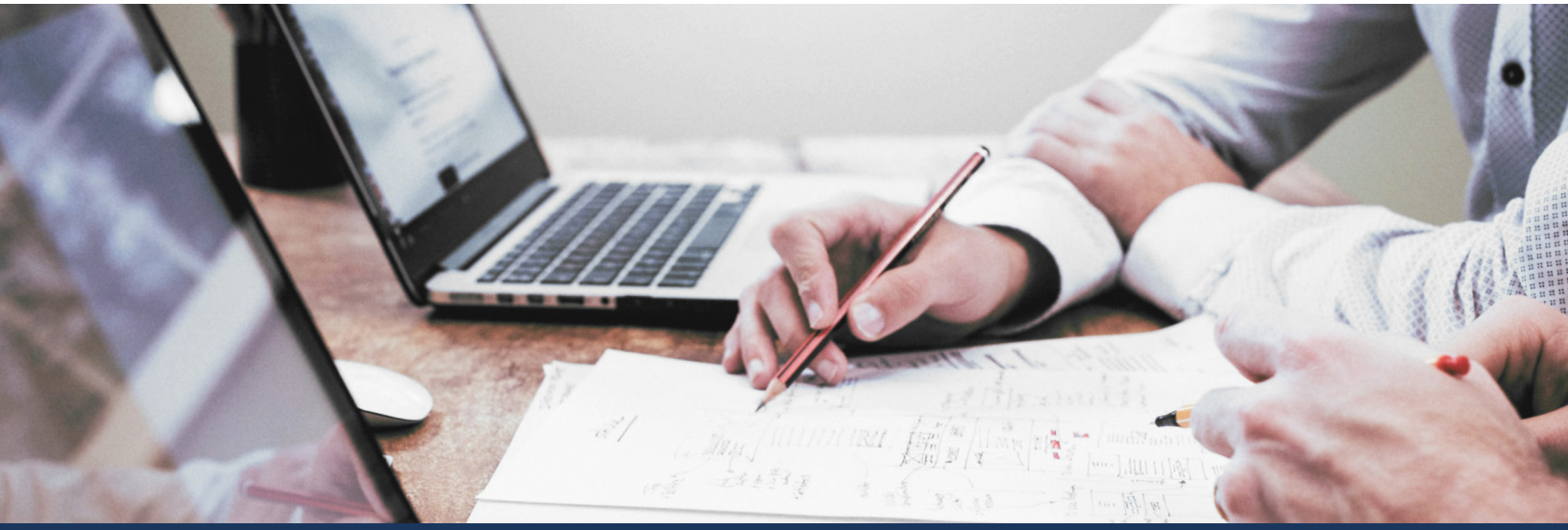
As businesses plan for 2022 and beyond, their data will be their strongest asset in making accurate yet agile business decisions. They don't have to look far to uncover data-driven insights that'll move their business forward, but simply uncovering them is proving to be a challenge for many.

Our research confirms that too many IT teams are still relying on dated analysis methods to deliver future-forward insights, costing them critical time-to-value on their data. Data management and prepping shouldn't overshadow data usage, and by relying on data lakes that are built on outdated infrastructures, their valuable insights are getting trapped in data swamps.

Limitless data capture empowers limitless innovation. The ability to retain greater and wider amounts of data for analysis unlocks new revenue opportunities, which is why organizations need to invest in modern cloud-based data lake solutions that can scale as the volume and velocity of their data grows, and that don't require upfront prep work from valuable IT talent.

SURVEY METHODOLOGY

ChaosSearch commissioned a survey of 1,020 U.S. IT professionals in firms with 500+ employees. The survey was fielded from September 28 to October 19, 2021.



ABOUT CHAOSSEARCH

ChaosSearch helps modern organizations Know Better™ by activating the data lake for analytics. The ChaosSearch Data Lake Platform indexes customers' cloud data, rendering it fully searchable and enabling analytics at scale with massive reductions of time, cost and complexity.

ChaosSearch was purpose-built for cost-effective, highly scalable analytics encompassing full text search, SQL and machine learning capabilities in one unified offering. The patented ChaosSearch technology instantly transforms your cloud object storage (Amazon S3, Google Cloud Storage) into a hot, analytical data lake.

For more information, visit ChaosSearch.io or follow us on Twitter @ChaosSearch and LinkedIn.

info@chaossearch.com | www.chaossearch.io