Nexla

THE DEFINITIVE DATA OPERATIONS REPORT

About Nexla:

Nexla is a data operations platform that helps teams create scalable, repeatable, and predictable data flows for any data use case. Analysts, business users, and data engineers across any sector including e-commerce, insurance, travel, and healthcare can use Nexla to integrate, automate and monitor their incoming and outgoing data flows. The end result is predictable and reliable data access inside and outside the organization.

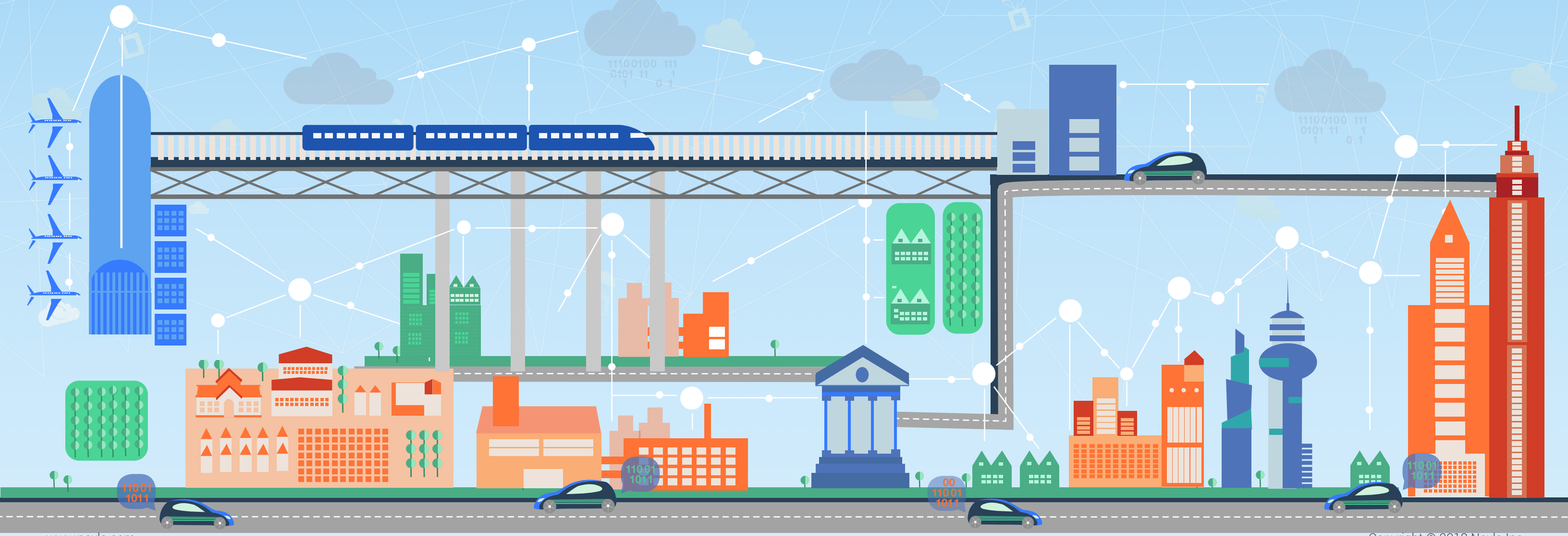


TABLE OF CONTENTS

DataOps controls the flow of data from source to value.

INTRODUCTION

Data operations is an emerging discipline that seeks to maximize the value of data across the entire organization. DataOps isn't an IT trend—it's a business trend. As the number of workers who need to use data in their everyday jobs increases, it no longer makes sense to relegate data to the purview of IT only. DataOps is an organization-wide data management practice that can help more stakeholders drive more value from data.

There are five key trends that are making DataOps critical for companies today:

1. Consumerization of enterprise and the rise of data fluency: More people need access to more data
Then: Can IT get this data for me?
Now: Why can't I do this myself?

2. Rise of the digital natives: Disruption from connecting data to software
Then: Sell technology to non-tech sector
Now: I will leverage data in this sector to change how things work

3. Buy, not Build: More B2B partnerships means more B2B data
Then: 6 Month CRM integration project.
Now: I need to connect data from these two services *today*

4. Big Data Maturity: Companies have built their analytics stacks
Then: What analytics system should I use?
Now: I need to feed more data into my AI models

5. More Data, More Problems: Data doubling every 18 months
Then: I have a few data sources
Now: I have a long roadmap of data I want to use. How can I get to that faster?

In this second-annual Definitive Data Operations Report, Nexla commissioned independent research firm, Pulse Q&A to survey hundreds of data professionals to understand how they are building data teams, what those teams are focused on, and where the challenges lie.

What is DataOps?

DataOps is an organization-wide data management practice that controls the flow of data from source to value, with the goal of speeding up the process of deriving value from data.

With DataOps, the outcome is scalable, repeatable, and predictable data flows for data engineers, data scientists, and business users. DataOps is as much about people as it is about tools and processes.

Tactically speaking, DataOps takes care of the grunt work typically placed on IT or data engineers. This includes integrating with data sources, performing transformations, converting data formats, and writing or delivering data to its required destination. DataOps also encompasses the monitoring and governance of these data flows while ensuring security.

A DataOps practice can open data access to more stakeholders within an organization, further increasing capacity for scale. The more "data leverage" you can create in an organization, the more likely you are to be successful.

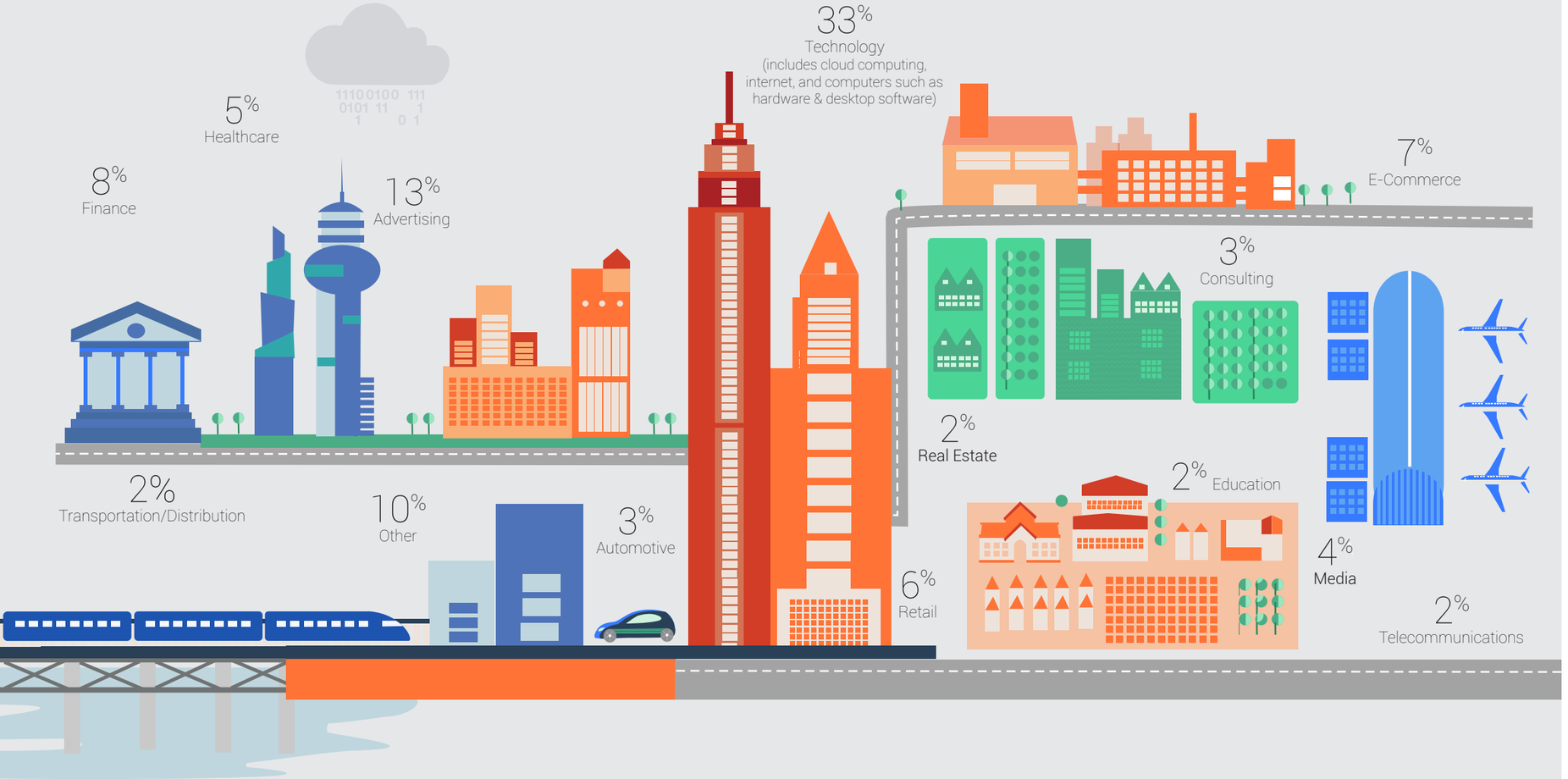
Ultimately, DataOps is not just about tools and processes. It represents a greater cultural shift that breaks down the silos between what has traditionally been viewed as "data backend" that produces usable data and "data frontend" that derives value from data. Only by enabling more users within their data systems can companies realize the economic benefits of becoming data-driven.

DataOps is a cultural shift that breaks down the silos between backend users who produce usable data and frontend users who derive value from that data.

METHODOLOGY

Pulse Q&A surveyed 266 data professionals from over 25 industries. The respondents included people working in tech companies as well as in e-commerce, advertising, finance, and more. The survey was conducted between May 3 - May 21, 2018.

WHAT IS YOUR COMPANY'S CORE INDUSTRY?

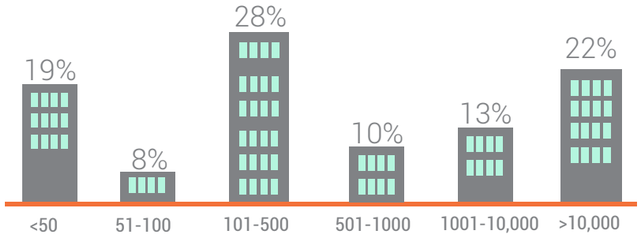


YEARS OF EXPERIENCE

Survey respondents had a wide range of experience, from new grads to data professionals of 10+ years.

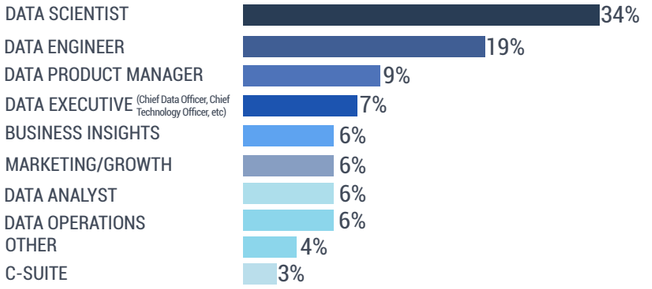
SIZE OF COMPANY

Companies included in this survey varied in size from less than 50 employees to over 10,000.



ROLES

From analysts through Chief Data Officers, only those who primarily work with data qualified for this survey.



DATAOPS AIMS TO SPEED UP THE PROCESS OF DERIVING VALUE FROM DATA.

KEY FINDINGS

- 85% of respondents say their companies have teams working on ML or AI. This is up from 70% last year
- 73% of respondents say their company has plans to hire in DataOps in the next year
- Data professionals are only spending 14% of their time on analysis. The rest of the time is spent on required but low value-add tasks like data integration, data cleanup, and troubleshooting. **Data engineers spend 18% of their time on troubleshooting. That works out to 9.3 weeks a year!**
- Data pros are longing for automation in their jobs. We asked data pros what tasks in their current role would benefit from automation:
  - The majority, 56%, unsurprisingly said that data clean up would benefit from automation
  - Analysis was the second-most cited task, at 47%
  - Data integration was close behind at 46% and building data pipelines at 41%

These findings highlight the need and desire for more scalable, automated processes to maximize value from data.

THE AVERAGE COMPANY...

has 1 data engineer for every

5 frontend business users

has data growth of 2.7 TBs a day

manages 4,300 data sets

ingests data from 4.4 partners

DATA ETHNOGRAPHY

Do you use real-time streaming data today?

YES - 58%

NO - 26%

NOT SURE - 16%

Will your real-time streaming increase in the next year?

13% No

24% Not Sure

63% Yes

What percent of your data infrastructure is in the cloud?

46% less than 50%

22% more than 50%

25% All of it - 100%

7% none - 0%

What are the data formats you use?

67% - JSON
59% - CSV
52% - TEXT
46% - XML
17% - PARQUET
11% - AVRO
9% - EDI
9% - OTHER
1% - ORC

What percent of your data is real-time streaming?

5% none

22% 1-10%

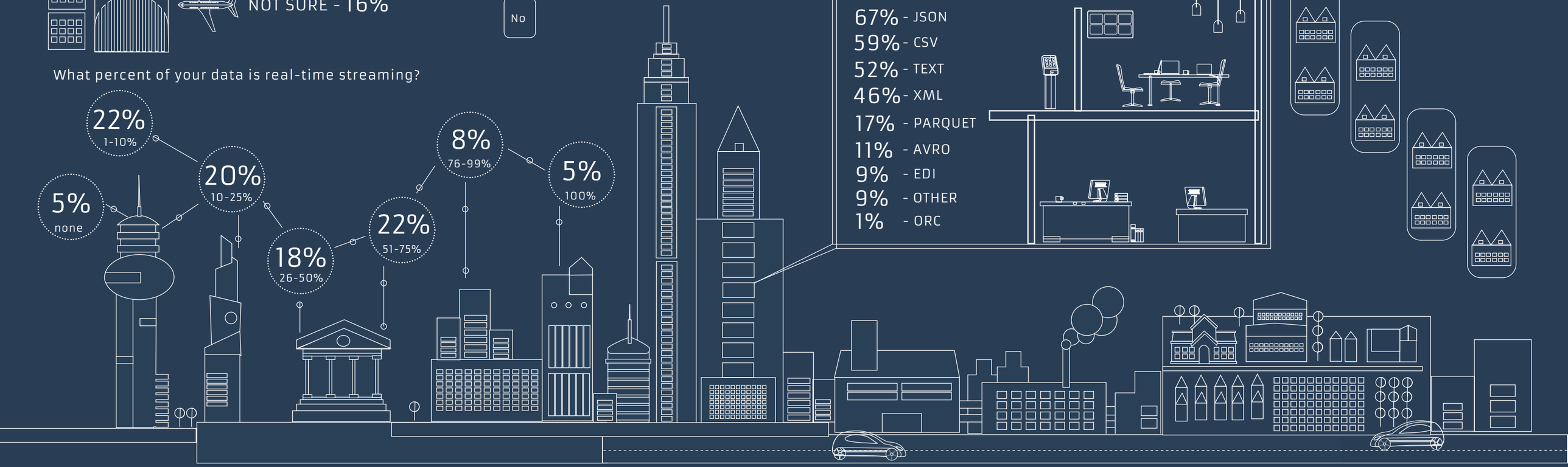
20% 10-25%

18% 26-50%

22% 51-75%

8% 76-99%

5% 100%

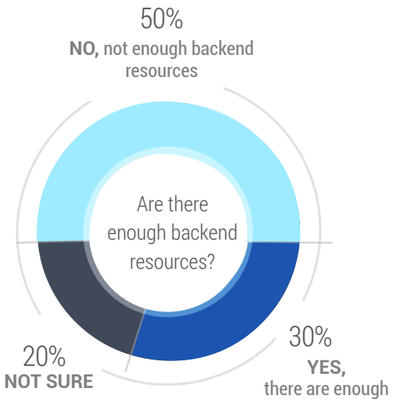


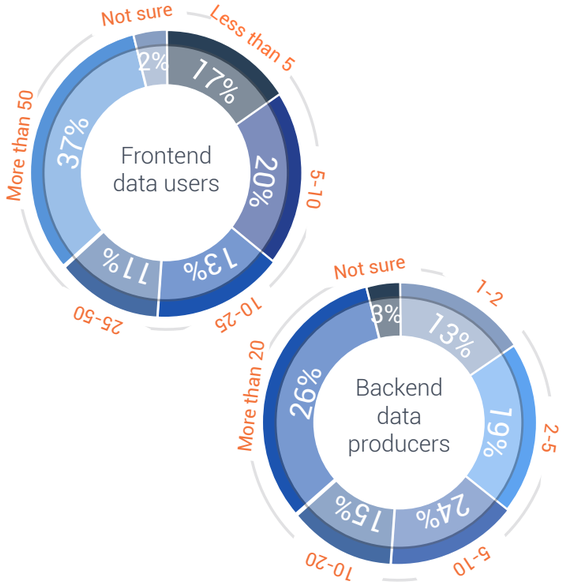
TEAM STRUCTURE

How big is the data team?

We asked data pros about the size of their data teams. They told us how many "frontend" data users, such as analysts, data scientists, and business users, they worked with and also how many "backend" or data engineering users, the team has.

We then asked data pros if they thought there were enough backend resources and data engineers to support data needs. 50% said no—there are not enough backend resources to support the company's data needs. Not surprisingly, respondents with more folks on the frontend team were more likely to believe they didn't have enough resources.





Frontend vs. Backend

We found the average data team has 5 backend engineers for every frontend professional that needs to use the data. There are outliers of course, with the minimum ratio at 0.5 (or two backend engineers for every frontend pro) and maximum ratio of 29. That's 29 frontend data users for every backend engineer— a ratio that is unlikely to be sustainable.

The more favorable ratios are found in smaller teams, where there are less than 10 front end users. It appears there is a minimum of 1 - 5 data engineers, even on the smallest teams. But as frontend users grow, the ratios get larger because the data engineering does not scale as quickly.

How DataPros spend their time

Data engineers spend 18% of their time troubleshooting. That's...

1 DAY each WEEK

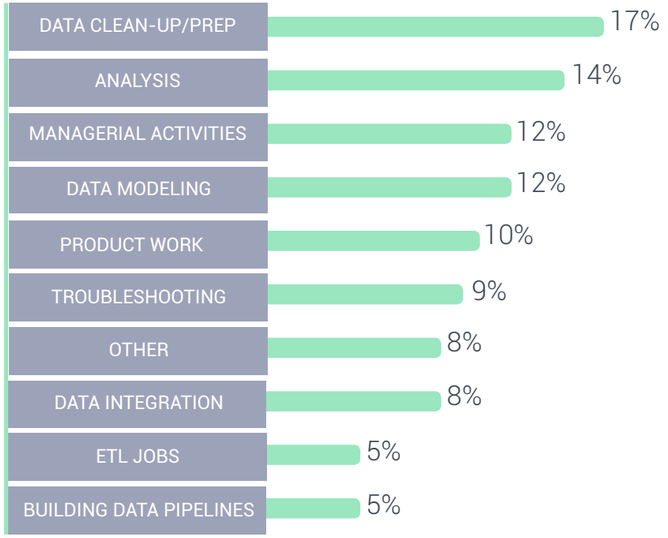
equals

9.3 WEEKS each YEAR

equals

7.2 YEARS of a 40 year CAREER

Average time spent in a month



As we did last year, we asked data pros how they spend their time in an average month. This year, we included activities like product work and managerial duties to gain a fuller picture of their workload. The classic data operations tasks of integration, building data pipelines, and ETL took up 18% of respondent's time. Data clean-up and prep was close behind, at 17% of time. As we expected, analysis was low on the list at 14% of time.

Interestingly, troubleshooting and fixing problems took up only 9% of the data pro's time. This figure seemed low, and is down 13% from last year. We looked deeper into the data and found when we restricted the sample to data engineers only, the number was 18%. Meaning, one full day a week of an engineer's time is spent troubleshooting. That works out to 9.3 weeks a year! What else could data engineers be doing with their time?

DATA ACTIVITIES

Least Enjoyable Activity

We asked data pros to tell us which activity in their job they enjoyed the least. It should come as no surprise that data clean-up and troubleshooting top the list, with 34% and 24% respectively. Some data pros would be happier as individual contributors as evidenced by 12% saying managerial activities were the least enjoyable.

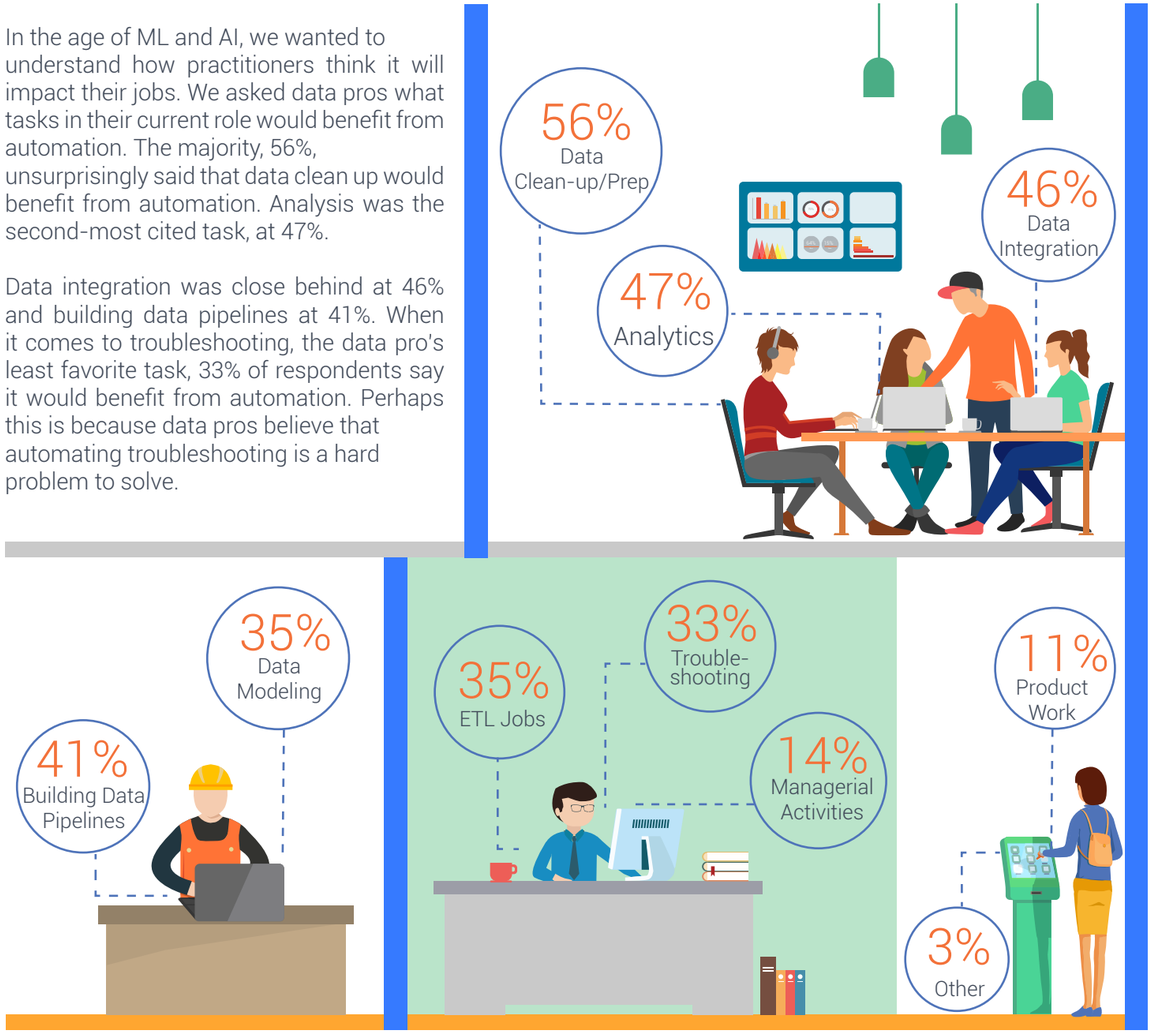
ACTIVITY	ALL RESPONDENTS	DATA SCIENTISTS	ENGINEERS, PRODUCT, EXECS
Data Clean-up/Prep	34%	44%	19%
Troubleshooting	24%	16%	35%
Managerial Activities	12%	8%	16%
ETL Jobs	6%	8%	3%
Data Modeling	5%	7%	3%
Building Data Pipelines	4%	5%	4%
Data Integration	4%	4%	5%
Product Work	4%	4%	4%
Other	4%	0%	10%
Analysis	3%	4%	1%

Data engineers find troubleshooting the least enjoyable activity. For data scientists, it's data clean-up.

Activities that would benefit from automation

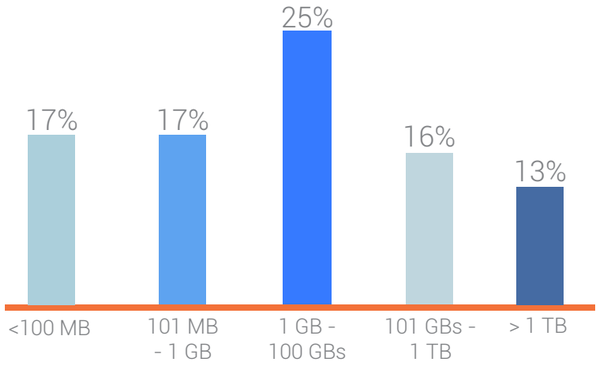
In the age of ML and AI, we wanted to understand how practitioners think it will impact their jobs. We asked data pros what tasks in their current role would benefit from automation. The majority, 56%, unsurprisingly said that data clean up would benefit from automation. Analysis was the second-most cited task, at 47%.

Data integration was close behind at 46% and building data pipelines at 41%. When it comes to troubleshooting, the data pro's least favorite task, 33% of respondents say it would benefit from automation. Perhaps this is because data pros believe that automating troubleshooting is a hard problem to solve.

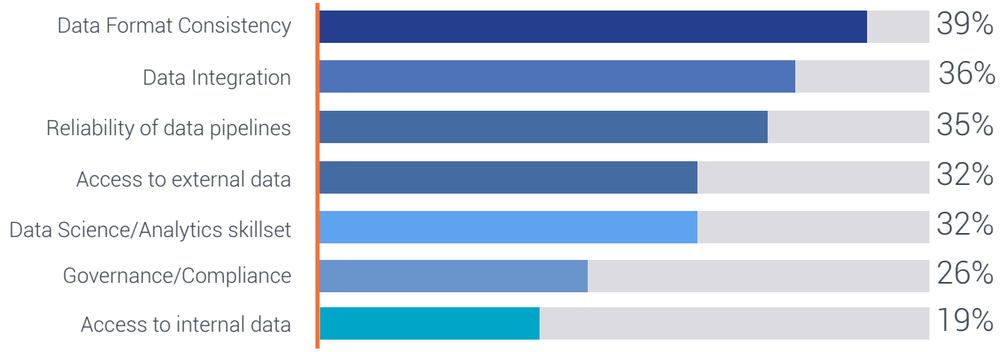


CHALLENGES WITH DATA

How fast is your data growing per day?



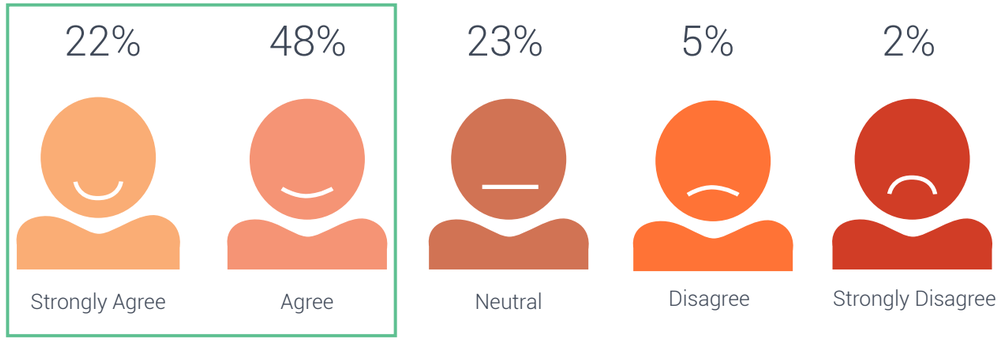
What are the challenges you face when working with data?



Given that data pros spend 17% of their time on data cleaning, it should come as no surprise that it tops the list of challenges they face. 39% percent of respondents said data format consistency is a challenge they face. This was followed by the usual DataOps suspects of integration at 36% and data pipeline reliability at 35%.

Data access continues to be a challenge for data pros. Interestingly, only 19% of respondents cite challenges with internal data access, suggesting efforts to break down data silos have been successful. By contrast, 32% cite access to external data as a challenge, suggesting inter-company data remains a challenge.

I wish I had more time to develop new solutions or improve current processes



70%

Now that we know what data pros are doing, what they don't like doing, and what they wish the robots would do for them, let's examine what they'd rather be doing instead. We asked data pros if they agreed with the statement: "I wish I had more time to devote to developing new solutions or improving current processes."
70% of respondents said they strongly agreed or agreed with that statement.

One thing is clear: Data professionals want to reclaim the time they're spending on manual DataOps.

Artificial Intelligence & Machine Learning

Almost all data pros report that their company is working on artificial intelligence and machine learning. This is up significantly from 2017, when "only" 70% of respondents reported their companies were working on ML or AI.

2018

2017

Hiring in DataOps

It should come as no surprise that the majority of respondents reported their companies have plans to hire in DataOps in the next 12 months.

73% Yes, we're hiring!

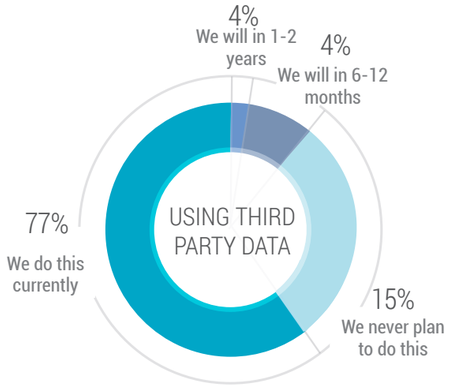
27% No, not now.

When looking at the 73% of respondents who said they are planning to hire, two-thirds reported they did not think there were enough backend resources. A perceived lack of backend resources seems to be a trigger for DataOps investment, which makes intuitive sense.

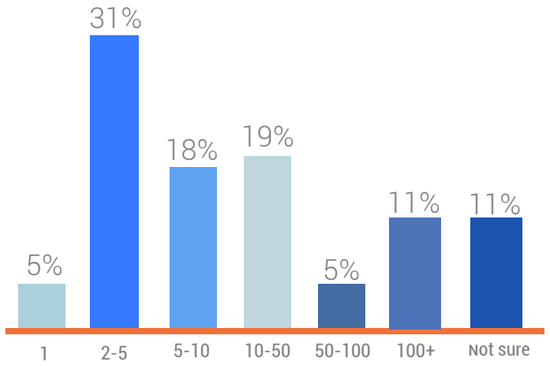
Receiving Data

Does your company currently, or have plans to, ingest data from third parties?

Companies are increasingly relying on data from outside their four walls. In 2018, 77% of respondents said their company currently ingests data from third parties. This is up from 60% last year.

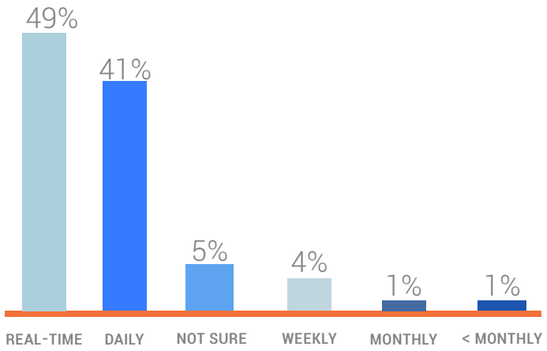


How many partners do you currently, or have plans to, ingest data from?



When we asked how many partners companies are ingesting data from, 54% said it was less than ten, and 35% said ten or more.

How frequently do you currently, or have plans to, ingest this data?

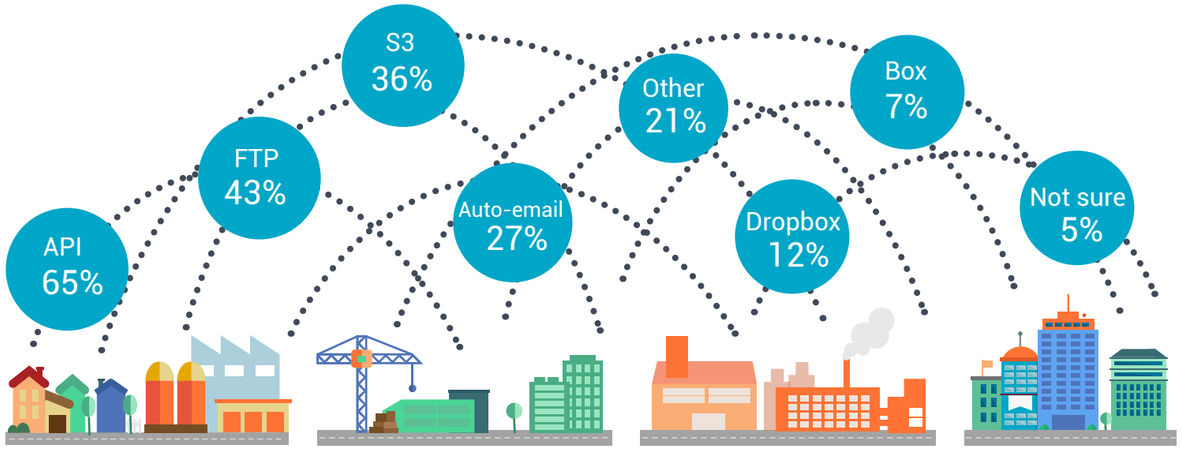


Third-party data ingestion isn't a once-in-awhile task. 49% of respondents said they are ingesting third party data in real-time while another 41% said they're ingesting it daily.

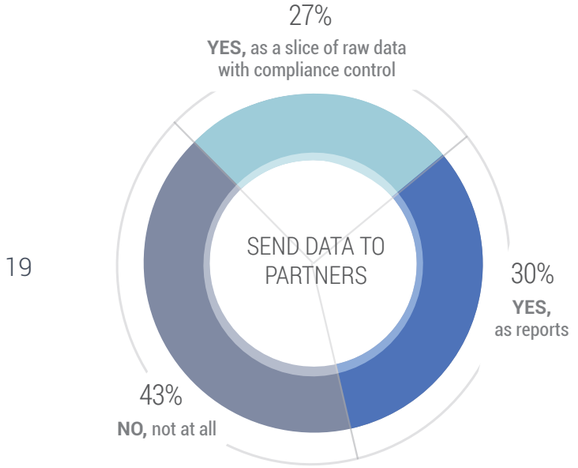
Sending Data

Which tools do you currently use to send data to partners?

Given this high frequency, it's not surprising that 65% of respondents say they send data via API. A still significant 43% say they're still using FTP to send data, while S3 and automated emails are used by 36% and 27%, respectively.

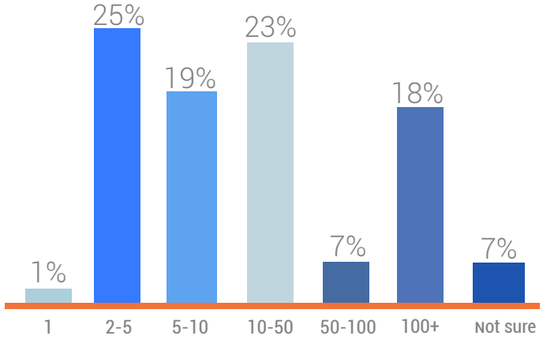


Do you currently, or do you have plans to, share data with third parties?



19

How many partners do you currently or have plans to share data with?



Of those companies that currently share data with third parties, 48% say they share with ten or more partners. 37% of these companies send their data in real-time, and 33% send the data daily.

ASSESSING YOUR DATAOPS

Based on our conversations with hundreds of companies, we've developed what we think is a handy matrix to understand your company's DataOps capability. To understand how the matrix works, it's critical to align on the right definition and criteria for assessing scalability and repeatability.

Factor One: Scalability

Scalability in this context is a measure of how easily a DataOps system can grow the volume of data, the number of data users, and operational complexity.

A DataOps infrastructure that is highly scalable can handle high volumes of data and process it in near real-time, while scaling with people. As companies ingest and send more data, the number of people who need to work with the data will only grow. Empowering them with tools is essential to the scalability.

This creates organizational leverage around data. The right tools, processes, and people as part of the DataOps solution can have a force multiplying effect.

Factor Two: Repeatability

Repeatability is a measure of how easily a system can automate or repeat tasks.

Sophisticated DataOps systems are able to maintain repeatability despite heterogeneous data types and sources. The ability to easily process data from multiple sources is key to repeatability.

DataOps is at its most repeatable when it can be smart about the source connections and easily integrate and transform data.

Where does your DataOps fall on the Matrix?

The first step to improving the data operations in your company is understanding where you are today. Depending on how your company rates on scalability and repeatability, your DataOps practice will fall into one of the four quadrants:

- State of the Art: High Scalability, High Repeatability
- Innovative: High Scalability, Low Repeatability
- Advanced: Low Scalability, High Repeatability
- Basic: Low Scalability, Low Repeatability

The DataOps Matrix

Scalable →

Repeatable ↑

Advanced

- Support and funding aimed towards DataOps
- Limited visibility into data quality
- Number of ad-hoc requests can be reduced through improved processes or tools

State of the Art

- You quickly integrate with many internal/ external data sources
- Systems can handle high volume and real-time data movement
- Ad-hoc engineering tasks are minimized

Basic

- Data integrations are scheduled and time consuming
- Low data velocity and batch processing
- Low visibility into data monitoring periods
- Errors and troubleshooting are significant projects

Innovative

- Data integrates smoothly through a set process with improvements on its way
- High volume, real-time data processing is in the works
- Resources allocated towards efficient data processes investment

CONCLUSION

With 73% of companies reporting that they plan to hire in DataOps in the next year, it's clear the shift to a DataOps data management practice is underway. This new approach is needed to handle the 2.7 TB of data (and growing) generated by the average company every day. Data complexity only increases when we recognize 85% of companies need to ingest some of this data from third parties, with 58% doing it in real-time.

That's why so many data professionals are hungry for automation in their day-to-day jobs, with 47% believing analysis could benefit from automation, and 46% saying integration could benefit. We know data teams are stretched since 50% of data pros reported that there are not enough backend resources to support their work. The average backend engineer is spending more than nine weeks a year just troubleshooting. DataOps offers a better way to accelerate time to value through automation and access.

Nexla is a data operations platform that helps teams create scalable, repeatable, and predictable data flows for any data use case. Analysts, business users, and data engineers use Nexla to integrate, automate and monitor their incoming and outgoing data flows. The end result is predictable and reliable data access inside and outside the organization.

DataOps is as much about people as it is about tools and processes.