DZone

DZONE'S 2019 GUIDE TO

Big Data

RESEARCH PARTNER SPOTLIGHT

unravel™

Key Research Findings

BY JORDAN BAKER
CONTENT COORDINATOR, DEVADA

Demographics

For this year's big data survey, we received 459 responses with a 78% completion rating. Based on this response rate, we have calculated the margin of error for this survey to be 5%. Below are some of the basic demographics of the respondents.

- Respondents reported working in four main industry verticals:
  - 17% work in finance/banking.
  - 17% work for software vendors.
  - 11% work in consulting firms.
  - 8% work in e-commerce/internet organizations.

37% work for organizations headquartered in the United States, 34% work for Europe-based organizations, and 5% work for companies HQ-ed in South Central Asia.

- A majority of survey-takers work in enterprise-level organizations:
  - 29% work in organizations sized 500-9,999.
  - 22% work in organizations sized 10,000+.
  - 18% work in organizations sized 100-499.

Over half (58%) of respondents have 15+ years of experience in IT, 22% have 10-15 years of experience, and 12% have been in the industry for six to nine years.

- Respondents fill three main roles in their organization:
  - 33% are developers/engineers.
  - 22% are architects.
  - 22% are developer team leads.
- Respondents work to develop three main types of software:
  - 80% create web applications/services.
  - 45% develop enterprise business applications.
  - 22% work on native mobile applications.

Big Data on the Rise

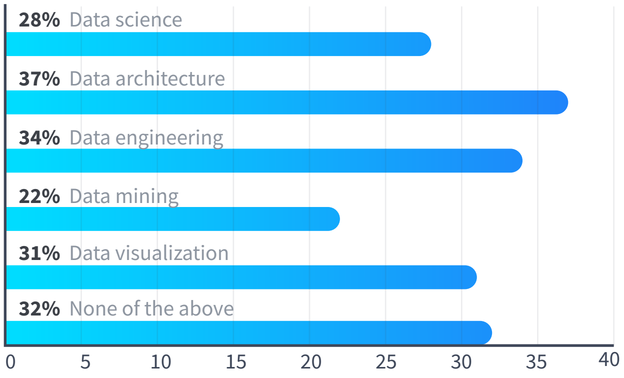
Over the course of the past year, our respondents have reported becoming much more data-driven. When asked what types of big data they tend to work with, 78% reported working with large volumes of data, 51% with a large variety of data, and 42% with data at high velocity. While the year-over-year change in the percentage of respondents working with large volumes of data and high velocity data fell within the margin of error for this report (a 4% increase and 2% decrease, respectively), these numbers, nonetheless, remain rather high. And those working with a large variety of data increased 7% year-over-year. Additionally, respondents' experience in all areas of big data increased considerably from our 2018 big data survey. Here's a quick breakdown of the percentages of respondents who had experience with a certain topic, comparing our 2018 survey data to the 2019 results:

- Data architecture
  - 2018: 26%
  - 2019: 37%
- Data engineering
  - 2018: 21%
  - 2019: 34%
- Data visualization
  - 2018: 21%
  - 2019: 31%
- Data science
  - 2018: 19%
  - 2019: 28%
- Data mining
  - 2018: 17%
  - 2019: 22%

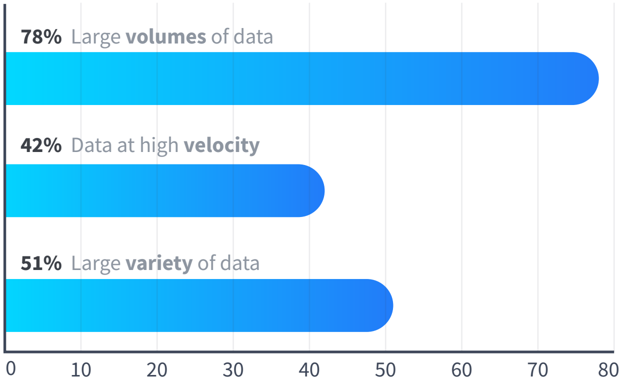
In addition to these impressive increases, the percentage of respondents

SURVEY RESPONSES

In which of the following areas of big data do you have experience?



What kind of big data do you work with?



reporting to have experience in none of those big data sub-fields fell from 45% in 2018 to 32% in 2019.

Given this increased interest in and experience with big data practices, it comes as no surprise that the adoption rates for big data-focused languages and frameworks also saw an increase over last year's survey. In 2018, 68% of respondents reported using Python for data science and machine learning; in this year's survey, 79% of respondents reported using Python. Spark and TensorFlow adoption also increased by 6%, bringing them to a 47% and 46% use rate, respectively, among our survey-takers. And, while we didn't see a dramatic increase over last year's survey, 51% of respondents told us they use R for the data science and machine learning projects.

For the rest of this report, we'll look at the various processes associated with each step of the big data pipeline (ingestion, management, and analysis), and see how they've changed over the past year.

Ingesting Data as High Velocity

When we asked respondents what data types give them the most issues regarding data velocity, two types saw noticeable increases over last year: relational (flat tables) and event data. In 2018, 33% of respondents reported relational data as an issue with regards to velocity; this year, that rose to 38%. For event logs, we saw the percentage of respondents reporting this data type as an issue go from 23% to 30%. Interestingly, relational data types seem to be a far bigger issue for users of R than for Python developers. Among those who use Python for data science, only 8% reported relational data types to be an issue when it came to data velocity. 30% of R users, however, told us they've had problems with relational data.

We also asked respondents which data sources gave them trouble when dealing with high-velocity data. Two of the issues reported fell drastically from our 2018 survey. The number of survey-takers reporting server logs as an issue fell by 10%, and those reporting user-generated data fell from 39% in 2018 to 20% in this year's survey. Despite these positive trends, respondents who said files (i.e. documents, media, etc.) give them trouble rose from 26% last year to 36%.

The tools and frameworks that data professionals and developers use to deal with data ingestions processes also witnessed interesting fluctuations over the past year. To perform data ingestion, 66% of survey-takers reported using Apache Kafka, up from 61% last year. While Kafka has been the most popular data ingestion framework for a while now, its popularity only continues to climb. For streaming data processing, Spark Streaming came out on top, with 49% of respondents telling us they use this framework (a 14% increase over last year). For performing data serialization processes, however, respondents were split between two popular choices. 36% told us they work with Avro (up from 18% in 2018) and 30% reported using Parquet (also up from 18% in 2018).

Managing a Large Volume of Data

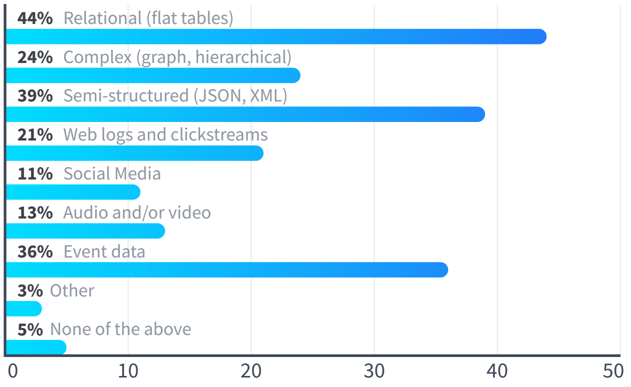
The basis of any data management plan is data storage. According to our respondents, there is a shift going on from cloud-based solutions to on-premise and hybrid solutions. 29% of respondents reported that their data typically resided in the cloud (down 10% from 2018), 31% told us they use a hybrid solution (up 7% over 2018's report), and 40% use on-premise data storage (another 7% year-over-year increase). In terms of the actual databases used to house this data, MySQL proved the most popular in both production (51%) and non-production (61%) environments, though its year-over-year adoption rate stayed rather static. PostgreSQL could be an interesting database to keep an eye on in the coming year, as its adoption rose in both production (42% in 2018 to 47% in 2019) and non-production (40% in 2018 to 48% in 2019) environments.

For filing big datasets, a vast majority of respondents told us they prefer the Hadoop Distributed Files System (HDFS). In fact, 80% of survey-takers reported using HDFS as their big data file system. While this large of a majority among respondents is impressive in its own right, HDFS also saw a 16% increase in adoption over our 2018 Big Data survey. The second most popular response to this question, Parquet, had a 36% adoption rate in our 2019 survey, up from 17% last year. Interestingly, even the least popular of the file systems reported, (O)RC File, saw an 11% year-over-year increase, rising to a 17% adoption rate.

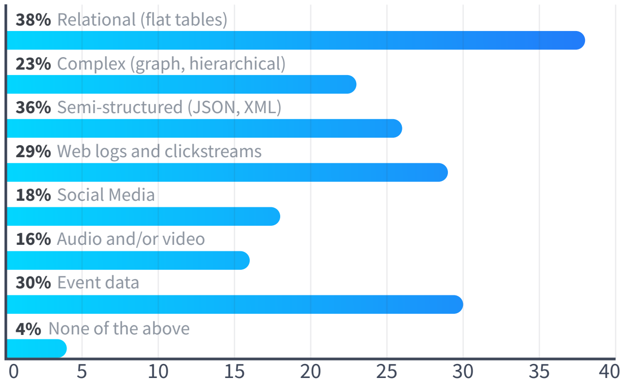
We also asked respondents about the issues they encounter when dealing with such large volumes of data. It turns out that normal files (such as documents, media files, etc.) cause the most headaches, with 49% of respondents selecting this option. Server logs also proved a popular answer, garnering 42% of responses. Data collected from IoT devices, however, saw the largest increase in developer frustrations. In 2018, 20% of respondents reported data from sensors or remote hardware as an issue;

SURVEY RESPONSES

What data types give you the most issues regarding the volume of data?



What data types give you the most issues regarding the velocity of data?



this year, 32% of survey-takers reported this type of data as a pain point. Surprisingly, despite user-generated data (i.e. social media, games, blogs, etc.) being one of the largest means of creating and ingesting new data, the difficulty this type of data gives to developers and data scientist seems to be decreasing. In 2018, 33% of respondents said user-generated data was a pain point in their big data operations; in 2019, this fell to 20%.

The types of data that gives developers issue when it comes to large volumes of data also witnessed a good deal of variability over last year. The data type that, according to respondents, causes that most issues — relational data — fell by 8%. Despite this decrease, it still registered 44% of respondents' votes. Event data also underwent a large swing, only in the opposite direction. In our 2018 survey, 25% of respondents said they had issues with event data; in 2019, this number rose to 36%. This increase in the number of respondents having trouble with event data is intriguing, given that user-generated data was reported as less of an issue than last year, yet much of the event data there is to be collected can be categorized as user-generated.

Analyzing a Variety of Data

Data mining is one of the most effective ways to sort through the immense variety of data an organization takes in. The two most popular tools for working with data wining, were, in fact, languages, specifically Python and R. Unsurprisingly, Python proved the most popular tool for data mining operations, garnering 80% of responses. Over half of respondents (51%) also selected R as a preferred data mining tool. Both of these numbers are fairly significant increases over our 2018 survey. Last year, 62% of respondents reported using Python and 41% said they used R for data mining.

One of the reasons these languages prove so popular in the data science community is the ability that their syntax gives to data professionals and developers to write complex algorithms. Interestingly, despite the extreme fluctuations we've covered in this report thus far, the preferred data mining algorithms stayed relatively stable year-over-year. Classification algorithms remained on top, selected by 62% of respondents, followed by clustering (61%) and regressions (52%). The only data mining algorithm which underwent a marked increase in adoption rate was times series algorithms. In 2018, 41% of respondents said they use time series algorithms in their data mining, which grew to 48% in this year's survey.

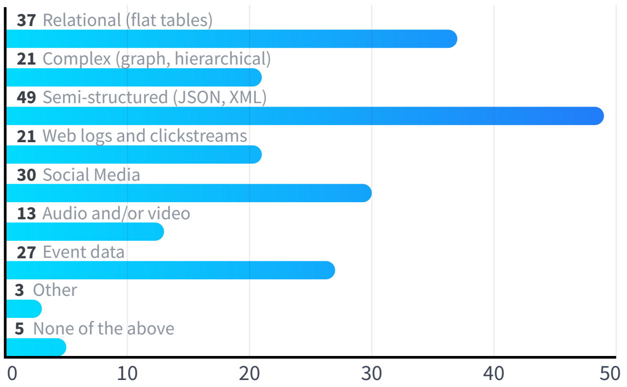
Among those respondents who work with data mining, 66% said that unsanitized data presents the largest challenge, up from 51% in 2018. This echoes the overall survey population, where 71% told us that data mining is the biggest issue in data science in general. Other important speedbumps in the data mining process that respondents delineated were the variety of data from different sources (65%) and a lack of knowledge or training (47%).

Given that well over half of all respondents reported that data variety is an issue, let's take a moment to examine the specific difficulties that come with highly variable datasets. When we asked survey-takers what data sources give them the most issues regarding the variety of data available, 52% said files (i.e. documents, media, etc.); 36% told us user-generated data; 25% reported ERP, CRM, or data from other enterprise systems; and 21% said sensor or remote hardware data. Of these five answers, only one saw any significant change over last year's survey (all the others falling within the margin of error for this report). When compared to 2018, respondents who are having issues with data from enterprise systems fell by 6%. While 40% told us that limited training and/or talent is an issue in the data science field, we nonetheless see an encouraging trend here, as the respondents who have issues with data sources due to data variety seems to have fallen over the course of the past year.

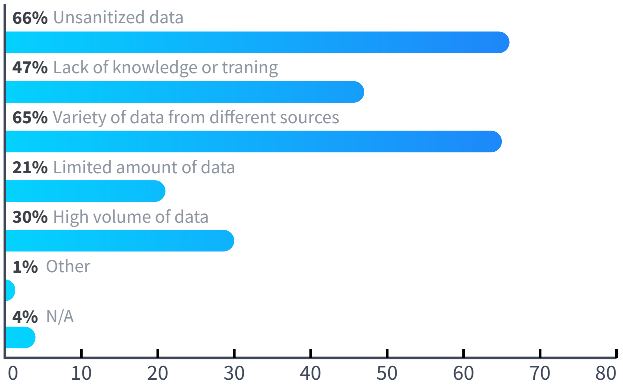
When we asked respondents about the data types that give them issues when it comes to the variety of data, we see a similar, positive trend. Respondents listed six main data types that prove challenging when working with data variety: semi-structured data, such as JSON or XML (49%); relation data (37%); data generated via social media (30%); event data (27%); complex data, such as graph or hierarchical data (21%); and web logs and clickstreams (21%). Of these six, only two saw a significant year-over-year change. In 2018, 41% of respondents reported complex data as an issue, while, as noted above, in 2019, only 21% of respondents reported this data type. This is a huge drop in the percentage of engineers and developers who have difficulty working with this data type. The second data type that underwent noticeable year-over-year change among our survey population was relational, which fell from 46% in 2018 to the aforementioned 37% in 2019. The fact that we are seeing stagnant or diminishing difficulty ratings across these six data types is, again, a positive sign. If developers and engineers continue to gain increased access to the proper training, expect these numbers to continue to fall.

SURVEY RESPONSES

What data types give you the most issues regarding the variety of data?



Which of the following challenges do you experience with data mining?



Diving Deeper Into Big Data

Twitter

@data_nerd

@revodavid

@mmarie

@JohnDCook

@victoria_holt

@caroljmcdonald

@karenchurch

@evdlaar

@GaelVaroquaux

@randal_olson

Podcasts

IBM Analytics Insights

Learn the latest in big data and analytics, as well as the implications of big data analytics for the enterprise from a range of experts in varying industries.

Data Stories

Get insight into the blurry lines between art and infographics through extensive coverage of data visualization projects.

O'Reilly Data Show

Dive into the opportunities and techniques driving big data and data science, as well as topics like graph databases, open source, and machine learning.

Zones

Big Data dzone.com/big-data

The Big Data Zone is a prime resource and community for big data professionals of all types. We're on top of all the best tips and news for Hadoop, R, and data visualization technologies. Not only that, but we also give you advice from data science experts on how to understand and present that data.

Database dzone.com/database

The Database Zone is DZone's portal for following the news and trends of the database ecosystems, which include relational (SQL) and nonrelational (NoSQL) solutions such as MySQL, PostgreSQL, SQL Server, NuoDB, Neo4j, MongoDB, CouchDB, Cassandra, and many others.

AI dzone.com/ai

The Artificial Intelligence (AI) Zone features all aspects of AI pertaining to machine learning, natural language processing, and cognitive computing. The AI Zone goes beyond the buzz and provides practical applications of chatbots, deep learning, knowledge engineering, and neural networks.

Refcardz

Temporal Data Processing

Download this Refcard to learn how to handle data that varies over time in relational databases using temporal tables.

Understanding Data Quality

This Refcard will show you the key places data derives from, characteristics of high-quality data, and the five phases of a data quality strategy that you can follow.

Getting Started With Apache Hadoop

Learn how Apache Hadoop stores and processes large datasets, get a breakdown of the core components of Hadoop, and learn the most popular frameworks for processing data on Hadoop.

Books

Data Analytics Made Accessible

In the 2019 edition of this highly rated book, dive into big data, artificial intelligence, data science, and R through real-world examples and use cases.

Data Science (MIT Press Essential Knowledge Series)

Get a concise introduction to the evolution of data science, how data science relates to machine learning, the ethical challenges that data science presents, and more.

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Learn from two leading big data experts what big data is, how it is changing lives, and how we can protect ourselves from its hazards.

unravel™

HELP YOUR APPLICATIONS DO DATA BETTER

Your Modern Data Applications, ETL, IoT, Machine Learning, Customer 360 and more, need to perform reliably. With Big Data, that's not always easy.

UNRAVEL MAKES DATA WORK

Unravel removes the blind spots in your data pipelines, providing AI-powered recommendations to drive more reliable performance in your modern data applications.

Greater Productivity

98% reduction in troubleshooting time

Guaranteed Reliability

100% of apps delivered on time

Lower Costs

60% reduction in cost

Don't just monitor performance – optimize it. LEARN HOW → UNRAVELDATA.COM

Will AI Finally Make Big Data Doable for the Enterprise?

Evidence of the success of big data is all around us. Amazon knows what we want based on the likes and wants of our million closest friends. Facebook will not let you forget that you once looked for a new lawn mower. Companies like LinkedIn, Amazon, and Google are all doing fine with big data and enjoying market caps that reflect the successful monetization of their data and associated algorithms.

But what about everyone else? The results are mixed and there is a palpable sense of disillusionment around the considerable cost, effort, and time needed to get a big data program off the ground. For data operations teams (DataOps), the margin between success and failure often hinges on two essential challenges:

- **Poor performance** achieving predictable peak performance of big data applications and their data pipeline components like Spark, Hadoop, and Kafka.
- **Scarce skillsets** acquiring and retaining the skills necessary to achieve and sustain success.

Unravel addresses both of these challenges

The key to solving both of these problems lies in the use of AI. Achieving peak performance on systems running hundreds of jobs across thousands of cluster nodes has become impossibly hard. So, we over-provision our datacenters or cloud resources and plan for the worst case. But we can apply AI algorithms to the operational metadata created during application runs to understand and predict application slowdowns and failures, and then prescribe corrective action.

To empower data operations (DataOps) teams to be able to scale their capabilities, AI can use operational metadata to diagnose underperforming applications and triage failures in a fraction of the time it would take if done manually.

These capabilities are built into Unravel.

WRITTEN BY DR. SHIVNATH BABU - CTO AT UNRAVEL DATA

PARTNER SPOTLIGHT

Unravel Data

Unravel radically simplifies the way you understand and optimize the performance of your data pipelines, with full-stack visibility and AI-powered recommendations.

Category Data Operations

New Release Unravel 4.5

Open Source? No

Case Study Unravel helps a top-20 global bank reduce big data application troubleshooting and tuning time, and speeds up response times for support teams serving application development and IT operations. The bank's multi-tenant data analytics platform is a very complex environment, and tooling has been an ongoing challenge for both app dev and operations teams. Unravel provides a 360º view of their modern data application portfolio, so that those teams can be proactive in their troubleshooting efforts and can support the SLAs required by the business. As a result of using Unravel's AI-powered auto-tuning and automated troubleshooting, the bank has seen a 70% reduction in support tickets and a 98% reduction in troubleshooting time for issues with their data pipelines. They have also seen a 60% reduction in resource costs through optimization and performance tuning.

Strengths

1. Captures and correlates performance and status data for Spark, Kafka, Hadoop, Hive, Hbase, Tez, Impala, and other data systems
2. For IT Operations - Reduce MTTI and MTTR with a unified, full stack view of your complex data pipelines with AI-powered alerts and recommendations
3. For App Developers - Write better code and optimize performance for cloud and hybrid environments with actionable advice, insights, and recommendations.
4. For Architects - Design production-ready data pipelines for current data needs and future hybrid cloud deployments
5. Cross-platform architecture supports Amazon, Microsoft, and Google clouds, as well as on-premises operations, hybrid, and multi-cloud environments.

Notable Customers

- Kaiser Permanente
- Autodesk
- Neustar
- TIAA
- Leidos
- Wayfair

Website	unraveldata.com
Twitter	@unraveldata
Blog	unraveldata.com/blog