2022 State of Workflow Orchestration

By GRADIENT FLOW

Executive Summary

Survey Details

The State of Orchestration Survey ran online from February 6 to April 4, 2022. A total of 581 respondents from a variety of industries participated. Respondents were recruited through online advertising and industry partners.

Key Takeaways

- **Usage of orchestration:** 43% of respondents stated that they use an orchestrator to handle over half of all recurring tasks. 68% of respondents who hold Data/ML Engineer roles reported that they use an orchestrator to handle over half of all recurring tasks; 91% of Data/ML Engineers said they use orchestrators for more than a quarter of all recurring tasks.

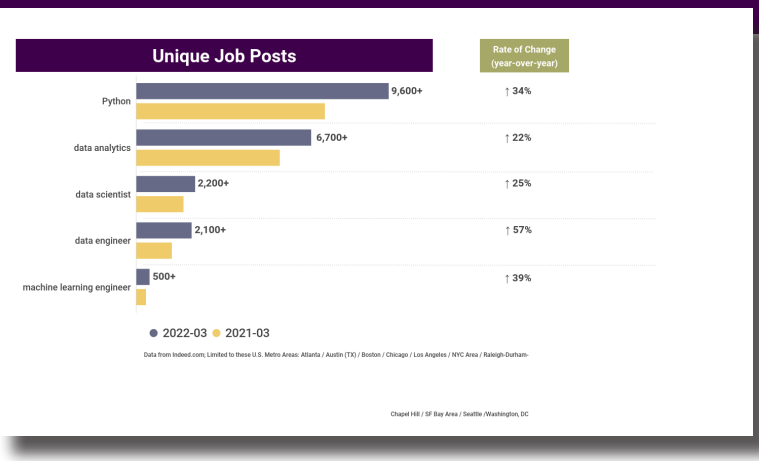
- **Key use cases:** 29% of all respondents chose data science as their primary use case for orchestration, making it the most popular use case overall. Data/ML Engineers designated the following top three primary use cases for orchestration: Data Movement (21%), Machine Learning/MLOps (16%), and Data Transformation (14%).

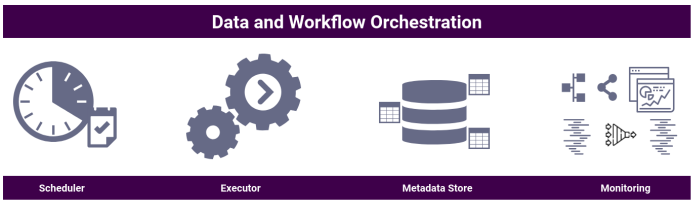
- **Orchestration frameworks:** The two most popular orchestration solutions are Apache Airflow (36% of all respondents) and Prefect (14%). Airflow was particularly dominant among survey respondents who hold Data Science / Analyst roles. Prefect (17%) and Dagster (12%) joined Airflow (32%) as the top three orchestration solutions favored by respondents who hold Data/ML Engineer roles.

- **Orchestration solutions for new projects:** Close to a quarter of all respondents (24%) and over a third (42%) of respondents with Data Science / Analyst roles signaled that they want to use Airflow in new projects. In comparison, Data/ML Engineers expressed stronger interest in trying Prefect and Dagster (over Airflow) for new projects.

- **Orchestration features:** The top three most important features were Ease of Use (38%), Caching (37%), and Monitoring (37%). The list of key features varied based on job role. Half of all respondents (51%) who hold Data Science / Analyst identified Caching as the most important feature for an orchestrator. In comparison, respondents who hold Data/ML Engineer or SWE/Dev/DevOps signaled that Monitoring is the most important feature.

Introduction





Numerous studies indicate that we are nearing a critical point in the evolution of analytics, machine learning, and artificial intelligence, where organizations are moving from small experiments to implementation. The global AI market will grow from $29.86 billion in 2020 to $299.64 billion by 2026, according to estimates. There are early signals that companies are accelerating their efforts to hire the type of talent necessary to deploy machine learning and artificial intelligence into their products and services following the global pandemic.

The huge growth in demand for data engineering talent indicates a heightened appreciation for foundational technologies. Early studies from McKinsey note that companies that excel at implementing and using AI have a clear data strategy that supports and enables analytics and AI. This includes investing in critical technologies such as data integration, DataOps, data governance, and data platforms. AI models perform better with better data, so it's no surprise that surveys through the years have shown that data teams spend the majority of their time gathering, cleaning, and enhancing datasets. More recently, machine learning researchers are rallying around "data-centric AI" — a collection of tools and techniques for cleaning, augmenting, and enhancing datasets to improve the accuracy of ML and AI models.

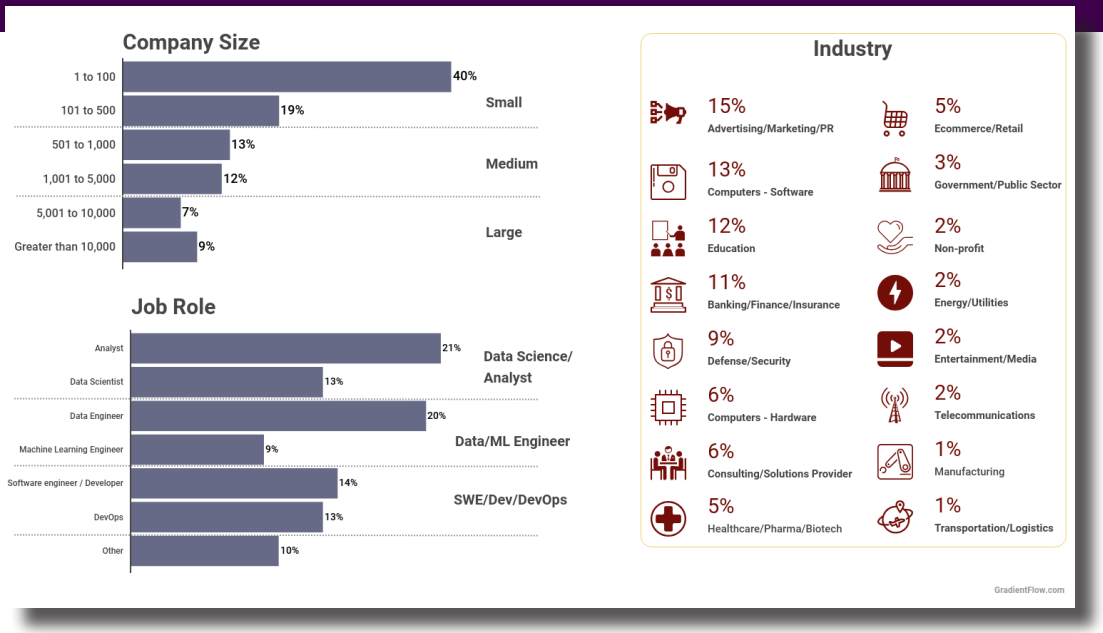
However, data unification and data integration challenges have gotten more complex as companies gravitate toward best-of-breed cloud services. Companies must now combine data from a growing number of data storage locations and cloud services, and make it available for downstream applications including data science and AI systems. The good news is that data engineering teams are now able to build modular data platforms composed of best-in-class tools (what many refer to as a "modern data stack").

Modernizing an organization's data infrastructure is increasingly difficult without an orchestrator. At a high-level, these are tools that enable developers to write, schedule, monitor, and manage pipelines. In the early stages of gathering and unlocking data assets, bespoke orchestrators may be sufficient. But as teams build and deploy more data and AI products and services, dataflow automation tools become essential.

As we note in this survey report, there are several popular open source projects that have spawned well-funded startups. Consequently, this category is currently a hotbed for innovation—new tools are on the way that will simplify and revise how companies build machine learning and data pipelines.

In this survey report, we identify key use cases for orchestrators as well as the volume of tasks for which they are used. We examine the specific tools companies are using today and which tools users want to try for future projects. We also identify and rank key features that respondents deemed critical for orchestration solutions.

Demographics & Key Segments



We group respondents based on their response to a question that measured their company's size:

- **Small:** up to 500 employees.
- **Medium:** 501 to 5,000 employees.
- **Large:** greater than 5,000 employees.

We also group respondents based on how they described their PRIMARY role:

Job Role	PRIMARY Role
Data Science / Analyst	Data Scientist or Analyst
Data/ML Engineer	Data Engineer or Machine Learning Engineer
SWE/Dev/DevOps	Software Engineer or Developer or DevOps

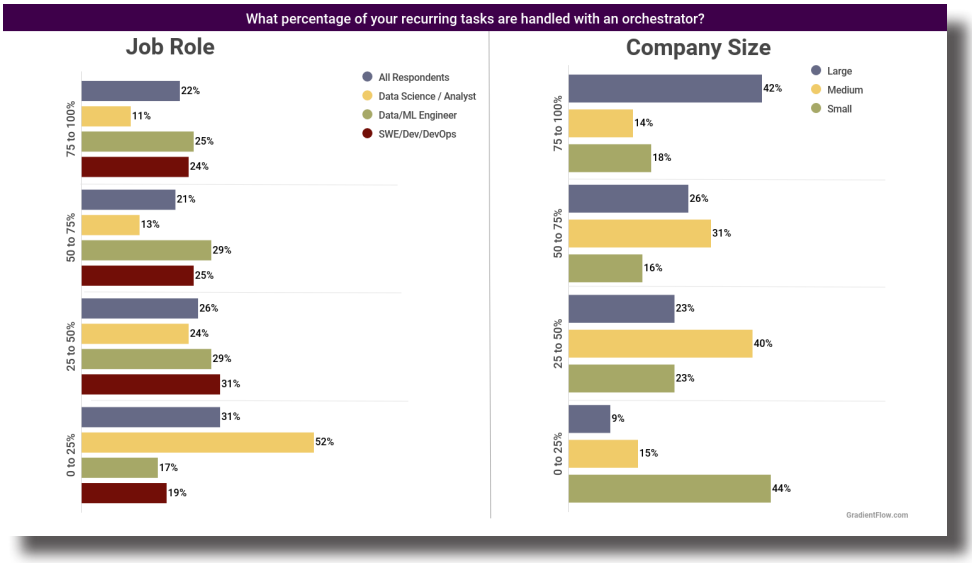
For the remainder of this survey report, we will frequently highlight responses based on segments derived from Job Role and Company Size.

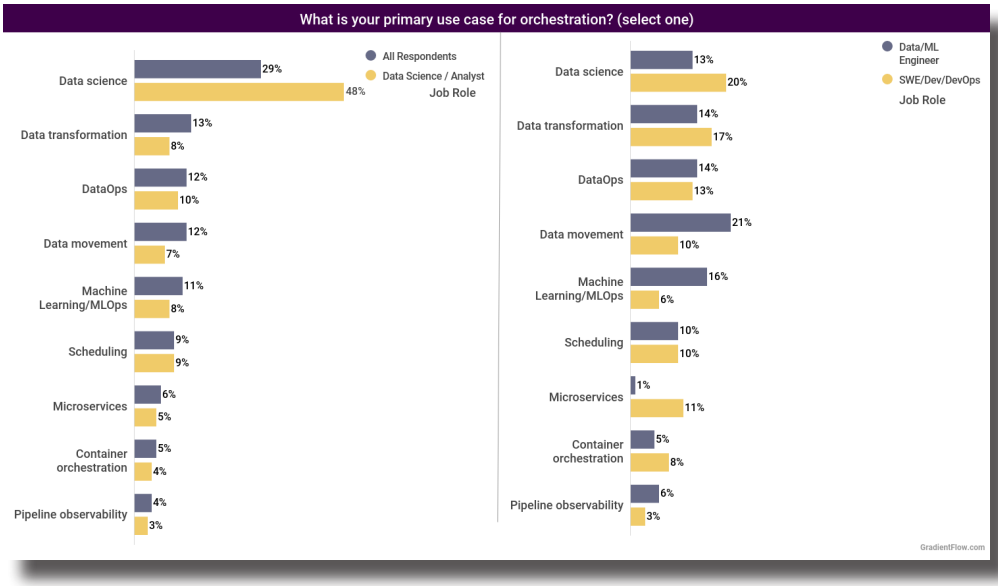
Usage of Orchestration

We began the survey by asking respondents what percentage of their recurring tasks are handled by an orchestration solution. Responses varied depending on job role and company size. Over half of all respondents who hold Data Science / Analyst roles signaled that they use orchestration for only 0-25% of recurring tasks.

In contrast, about a quarter of respondents who hold more technical roles (Data/ML Engineer or SWE/Dev/DevOps) stated that they use orchestration for 75-100% of all recurring tasks. 42% of respondents who work at Large companies also stated that they use orchestration for 75-100% of all recurring tasks.

29% of all respondents chose data science as their primary use case for orchestration, making it the most popular application overall. Not surprisingly, the picture shifts once we look at specific job roles. Data/ML Engineers designated the following top three primary use cases for orchestration: Data Movement (21%), Machine Learning/ MLOps (16%), and Data Transformation (14%). This is in line with expectations, as data movement and data transformation (ELT) are tasks very much associated with data engineering.



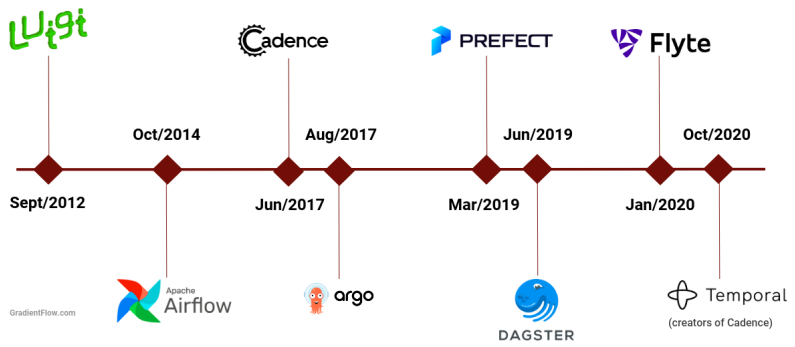


Orchestration Solutions

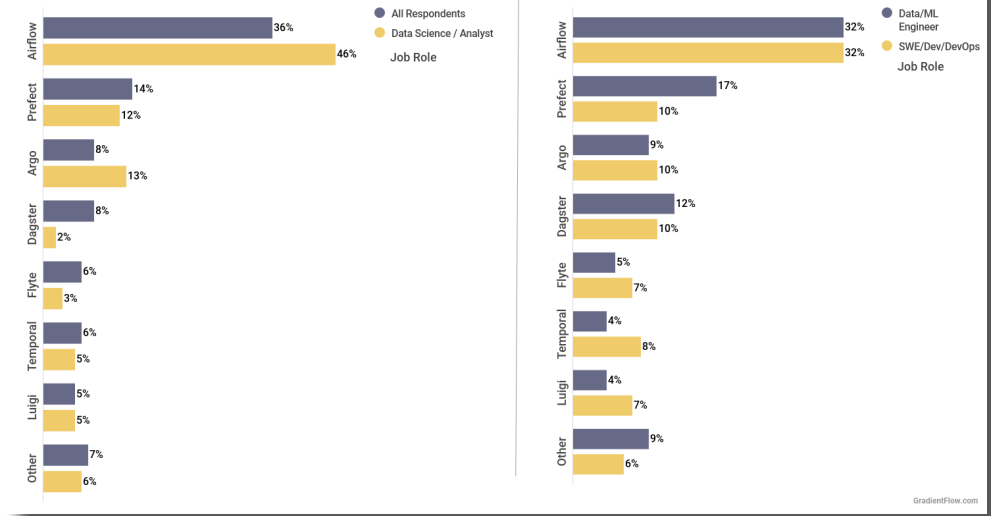
Over the past decade, several notable open source and SaaS orchestration systems have been released and a related group of startups have collectively raised over $450 million in funding. The rise of well-funded startups confirm that orchestration is currently a fertile ground for innovation. The hope is that we will soon have new tools that further simplify how companies build and manage data and machine learning pipelines. The core of this survey report examines which specific tools respondents are using and what features they value in their orchestration systems.

In line with results from our recent Data Engineering Survey, the two most popular orchestration solutions are Apache Airflow (36% of all respondents) and Prefect (14%). Long regarded as one of the first tools to make orchestration more manageable, Airflow is the leading choice across all job roles. Airflow is particularly dominant among survey respondents who hold Data Science / Analyst roles. Looking at more technical roles, Prefect (17%) and Dagster (12%) join Airflow (32%) as the top three orchestration solutions favored by respondents who hold Data/ML Engineer roles.

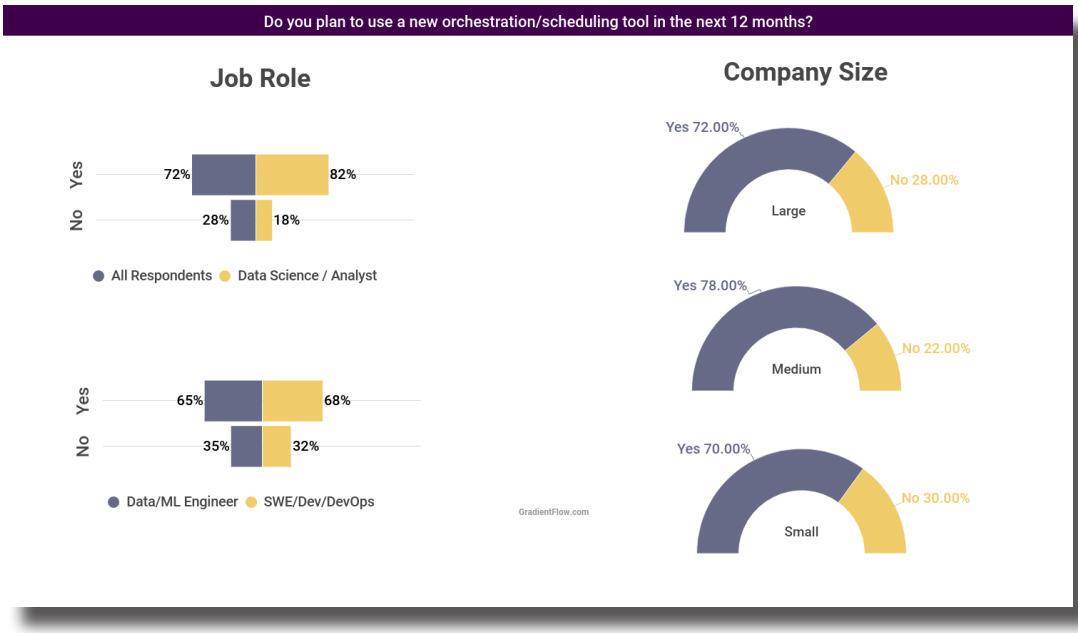
Launch Date of Key Orchestration Solutions



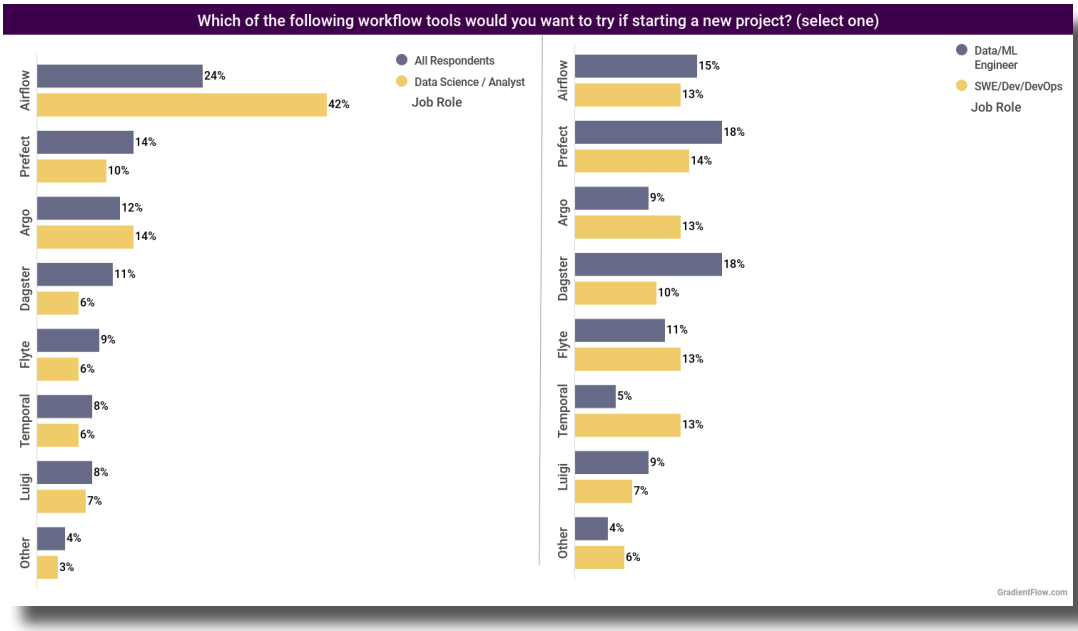
Do you use a workflow orchestration tool? If so, which? (select one)



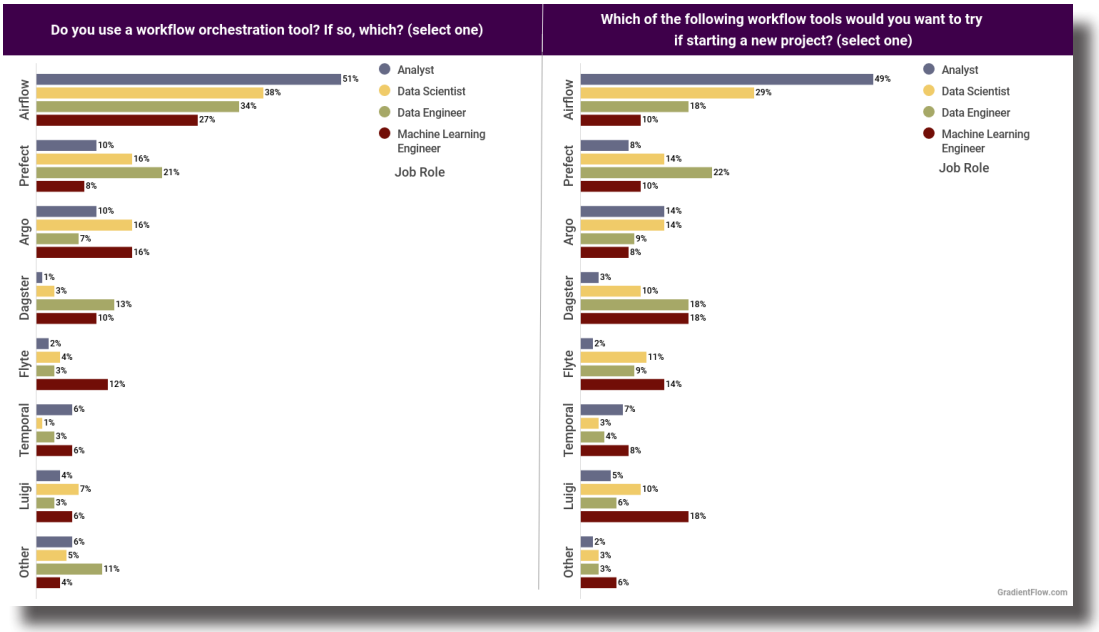
Survey respondents are keen on trying new orchestration tools: 82% of respondents with Data Science / Analyst roles indicated that they plan to use a new orchestration/scheduling tool in the next year.



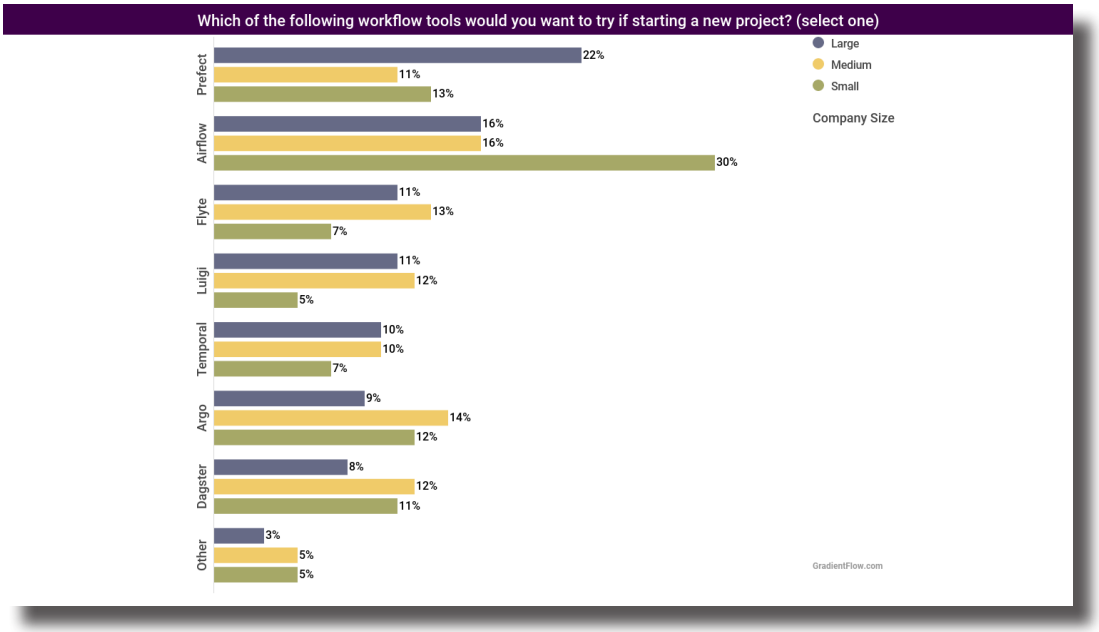
Next, we tease out which new orchestration tool respondents are interested in using for their new projects. Airflow is once again the most popular option, with close to a quarter of all respondents (24%) and over a third (42%) of respondents with Data Science / Analyst roles choosing it. In comparison, Data/ML Engineers expressed stronger interest in trying Prefect and Dagster (over Airflow) for new projects. Among SWE/Dev/DevOps respondents, Prefect (14%) was the top selection, with Airflow, Argo, Temporal (13%) all tied for second place.



We also examine tool preference for four specific job types associated with data and AI applications: Analysts, Data Scientists, Data Engineers, and Machine Learning Engineers. More than half (51%) of all Analysts use Airflow, but the share of use of Airflow is much lower for more technical roles (34% for Data Engineers, and 27% for ML Engineers).



When viewed through the lens of company size, more than one-fifth (22%) of all respondents who work at Large companies (more than 5,000 employees) expressed interest in using Prefect for new projects. As we'll see in a survey question on product features, respondents from Large companies prioritize features that Prefect excels at (Monitoring, Ease of Use, Scalability).

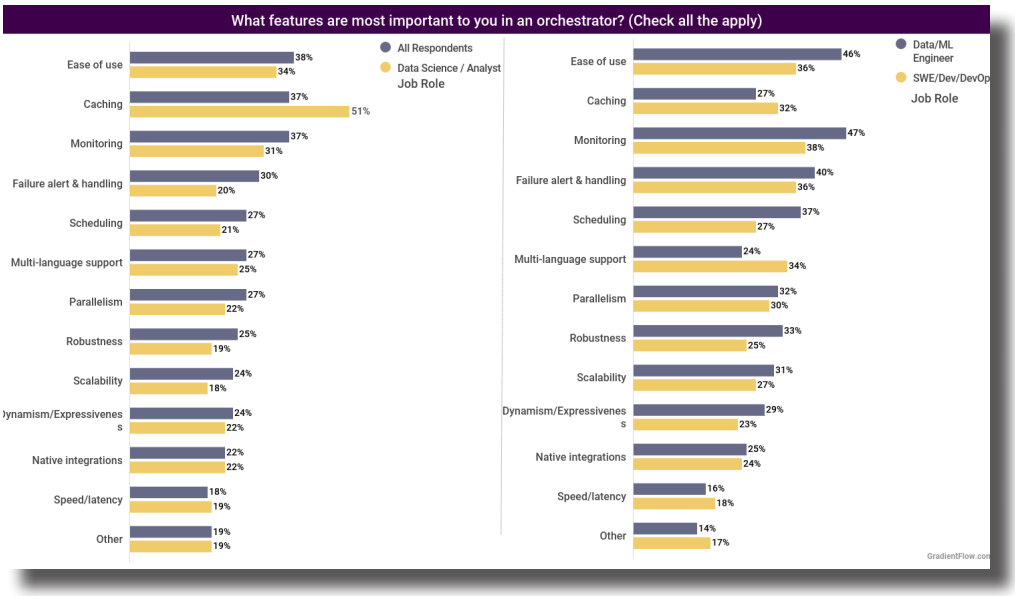


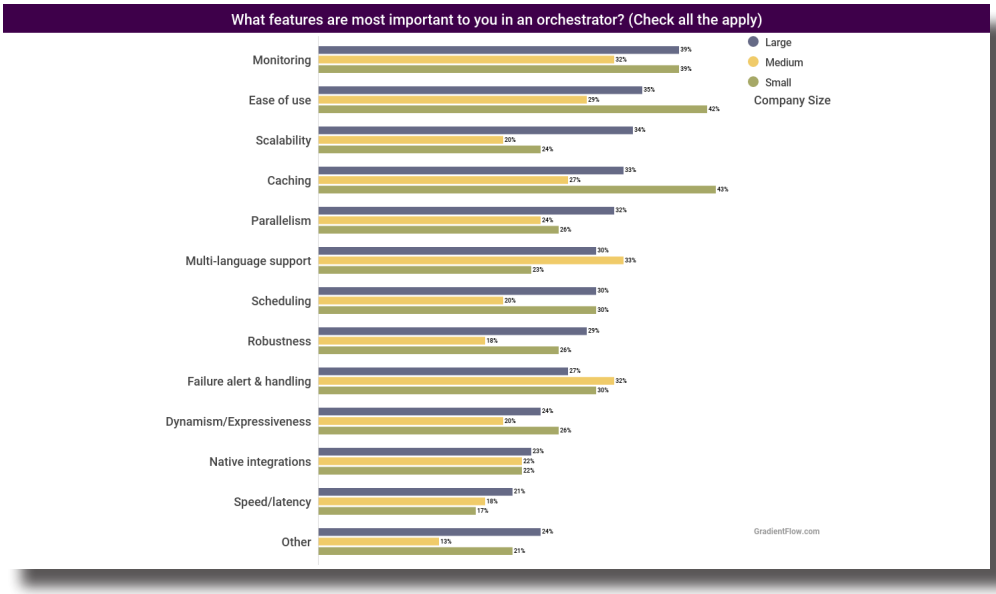
Lastly, we asked respondents what features they believe orchestration solutions should have. The top three most important features are Ease of Use (38%), Caching (37%), and Monitoring (37%). In light of the fact that Ease of Use is a top priority for respondents, it's encouraging to see that startups building orchestration tools are focused on user and developer experience (at least based on recent announcements).

The list of key orchestration features varied based on job role. Half of all respondents (51%) who hold Data Science / Analyst roles identified Caching as a key feature. Caching enables users to efficiently reuse previously retrieved or computed data, a prized feature in many data science and ML projects where interactivity and timely results are valued.

In comparison, Caching was lower down the list for respondents who hold Data/ML Engineer or SWE/Dev/DevOps roles. Among respondents who hold Data/ML Engineer or SWE/Dev/DevOps Monitoring emerged as the most popular feature. Failure Alert & Handling is the second most popular feature among SWE/Dev/DevOps respondents and the third most popular feature among Data/ML Engineers.

Drilling into results by company size, respondents from Large companies chose Monitoring (39%), Ease of Use (35%), Scalability (34%), Caching (33%), and Parallelism (32%) as their top five features.





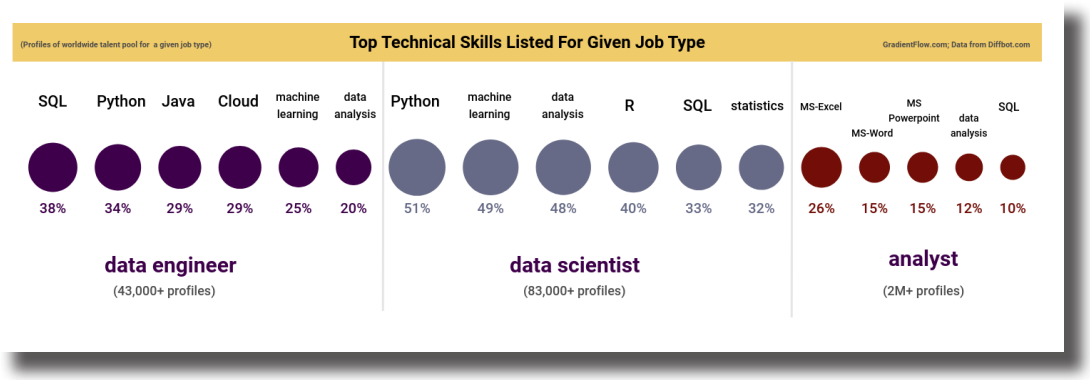
Closing Thoughts

As we noted in a recent post, data integration and data orchestration are active areas with well-funded startups building a wide array of solutions. Judging from the results of this survey, orchestration solutions have clearly arrived: 91% of Data/ML Engineers said they use orchestrators for more than a quarter of all recurring tasks.

Respondents (particularly Data/ML Engineers) stated that they consider features such as Ease of Use and Monitoring as particularly important. Recent announcements are encouraging for users concerned about usability. Major new initiatives from Prefect ("code as workflows"), Dagster ("software-defined assets"), and others are squarely aimed at helping users build and manage pipelines as seamlessly as possible.

Looking ahead, a couple of important trends are worth highlighting. Low-code/No-code tools are increasingly finding their way into data and AI, and data integration is no exception. Analysts vastly outnumber data scientists and data engineers. It will be interesting to see how tools evolve to accommodate less technical users.

Secondly, AI introduces new data challenges for platform teams accustomed to working mainly with structured data or text. Data platforms will need to handle more types of data and more sources as computer vision and speech applications become easier to build. As multimodal machine learning and AI applications become more common, users will need data tools and platforms that can handle different data types.



Acknowledgements

Thanks to Jesse Anderson and the Big Data Institute, Kathy Yu, and Jenn Webb for providing critical assistance. This survey was conducted by Gradient Flow; see our Statement of Editorial Independence.