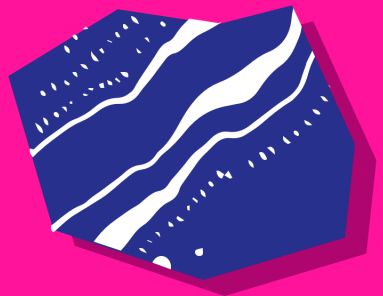


SODA:

Building Today's Data Team



Data Dream Team Series

Soda and Jesse Anderson have embarked on a series to explore building today's data team. With the seismic shift in how organizations think about data as the modern data value chain evolves, there's a new approach needed to align how the organization, the team, the people, are structured and organized around data.

In this guide Jesse explores how to build teams around data.



Jesse Anderson

**Data Engineer, Creative Engineer &
Managing Director of Big Data Institute**

He mentors companies all over the world ranging from startups to Fortune 100 companies on Big Data. This includes projects using cutting-edge technologies like Apache Kafka, Apache Hadoop, and Apache Spark.

He is widely regarded as an expert in the field and for his novel teaching practices. Jesse is published on Apress, O'Reilly, and Pragmatic Programmers. He has been covered in prestigious publications such as The Wall Street Journal, Harvard Business Review, CNN, BBC, NPR, Engadget, and Wired.



Soda

Data Reliability Tools & Observability Platform

The Soda platform and developer tools enable organizations to proactively manage data quality issues by providing both business and technical users with a single platform to discover, analyze, prioritize, and collaboratively resolve data quality issues. This approach brings everyone closer to the data and means that management and governance processes for data quality can quickly and easily form around data products.

Introduction

Smart companies are taking advantage of the unfathomable quantities of data being generated by transactions and mobile devices. Walmart can [optimize its use of staff](#) by keeping track of when people make purchases or pick up prescriptions. Delta Air Lines [helps customers find their connecting flights.](#)

Such data is available to smaller businesses, too, who are willing to analyze their needs and the missing information that can fill the gap. Catchphrases that are circulating around the industry, such as “digital transformation” and “data is a product” show that unprecedented possibilities are opening up through the proper use of data, and that businesses who grasp those possibilities will pull ahead. Along the way, they will improve their services to customers and [create a more sustainable environment.](#)

Introduction

Smart companies are taking advantage of the unfathomable quantities of data being generated by transactions and mobile devices. Walmart can optimize its use of staff by keeping track of when people make purchases or pick up prescriptions.¹ Delta Air Lines helps customers find their connecting flights.²

Such data is available to smaller businesses, too, who are willing to analyze their needs and the missing information that can fill the gap. Catchphrases that are circulating around the industry, such as “digital transformation” and “data is a product” show that unprecedented possibilities are opening up through the proper use of data, and that businesses who grasp those possibilities will pull ahead. Along the way, they will improve their services to customers and create a more sustainable environment.³

¹ 5 Ways Walmart Uses Big Data to Help Customers. [Read Article](#)

² Delta crew leveraging big data to personalise in-flight experience. [Read Article](#)

³ Opportunities for big data in conservation and sustainability. [Read Article](#)

01

Basics of Data Teams

Put simply, success with big data comes when value is created from your data. The data team is part of a larger process that must deal with questions such as: What new markets, services, or cost optimizations is your business embarking on? What data could help you make these decisions, and is it available from your data sources? Are managers providing requirements to the data team? Do you know what you will do with the results of data analysis and who is responsible for acting on the analysis?

Whether you are handing out data to your business analysts in spreadsheets or running thousands of services to process continuous data streams, you have to think about the team of people who protect and process your data. Making use of data is a team sport. It takes several different kinds of people to get things done, and in all but the smallest organizations, they should be organized into multiple teams.

To understand the scope of a team – particularly one dealing with modern data sets that go by the term “big data” – it’s worth thinking about data pipelines and data products. Briefly put, a data pipeline is a way of making data available: bringing it into an organization, transferring it to another team, and so on – but usually transforming the data along the way to make it more useful. A data product takes in a dataset, organizes the data in a way that is consumable by others, and exposes it in a form that’s usable by others.

More specifically, a data pipeline is a process to take raw data and transform it into a usable form by the next recipient in the organization. To be successful, this data must be served up by technologies that are the right tools for the job and that are correct for the use cases. As a result, the data is available in formats that reflect the changing nature of data and the enterprise demand.

The output of these data pipelines are data products. These data products should become the lifeblood of the organization. If this doesn’t happen, the organization isn’t a data-driven organization and is not reaping the benefits of its investment in data.

The Three Teams

To do big data right, you need three different teams. Each team does something very specific in creating value from data. Each team has specialized knowledge that complements the other teams. Each team has both strengths and weaknesses inherent to the individual's experiences and skills.

Data Science

When most managers think or hear of big data, it's in the context of data science. However, the reality is that data science is just one piece of the puzzle. Eric Colson, Chief Algorithms Officer of Stitch Fix, says, "Our data scientists are very well equipped with business context so they can judge for themselves how valuable something will be. They have access to the data to vet their ideas, and they have the business context to know how much impact a new capability can have."

The data science team consumes data pipelines to create derivative data. In other words, they take data pipelines that were previously developed and augment them in various ways.

Sometimes the augmentation consists of advanced analytics, notably the machine learning (ML) that is so hot nowadays. The members of a data science team usually have extensive backgrounds in mathematical disciplines like statistics.

In short: A data scientist is someone who has augmented their math and statistics background with programming to analyze data and create applied mathematical models.

Data Engineering

Going from data science in the lab to running data science at scale in the business isn't a trivial task. You need people who can create maintainable and sustainable data systems that nonspecialists can use. It takes a person with an engineering background to do this right.

The data engineering team creates the data pipeline that feeds data to the rest of the organization, including the data scientists. The data engineers need the skills to develop data products that are clean, valid, maintainable, usable at scale, and sustainable for additions and improvements. The data engineers are also responsible for choosing the data infrastructure to run on. Bas Geerdink of Rabobank,

a cooperative bank, defined the role of data scientists as "to build and to train predictive models on top of data." The data engineer's role is "to run those models in a production environment that goes from the data input to the data output."

According to Dmitriy Ryaboy at Twitter, placing the data engineering team under the engineering team made it easier to have the necessary conversations about how to make software changes and the effect on analytics. Instead of finding out about a change at the last minute or not at all, the data engineering team could be part of the conversation about rearchitecting or making major system changes from the beginning.


It's useful to have three levels of expertise in data engineering. At the lowest level are people supporting the projects on a daily basis. These may be new to the field of data engineering. Once someone has successfully carried through a major data engineering project, they can be considered qualified data engineers with greater responsibilities. And the highest level is called a veteran, who has seen it all.

In short: A data engineer is someone who has specialized their skills in creating software solutions around big data.

Operations

Running distributed system frameworks in production ranges from rock-solid to temperamental. You need an operations team to keep the ship moving and everything chugging along. Traditionally, that team does not touch the data engineer's code. They may have a hand in automation but not in writing the code for pipelines. An alternative mixes data engineering and operational functions. The same team is responsible for both the data pipeline code and keeping it running.

In either case, operations staff are responsible for the operation in production of the custom software written by your data engineers and data scientists (and maybe other people too). They keep the network optimized, fix hardware issues, install software and operating systems, etc. With big data, they have to understand cluster software and other big data technologies you have operationalized. They are familiar with the code being run and with the expected amount, type, and format of the incoming data.



In short: An operations engineer is someone with an operational or systems engineering background who has specialized their skills in big data operations, understands data, and has learned some programming.

02

Some Organizational Practices

Think strategically when building your data team:

- The data scientists create advanced analytics. A data scientist combines an advanced math and statistics background with programming, domain knowledge, and communication skills to analyze data, create applied mathematical models, and present results in a form useful to the organization.
- The data engineering team is responsible for creating data products and the architecture to build data products. On the architecture side, it is up to the data engineering team to create data pipelines. They must also choose the right technologies for the data and the use cases to accomplish this task.
- An operations team makes things reliable. They're responsible for putting both cluster software and custom

software into production and making sure it runs smoothly. The most common job titles in this position are operations engineer, systems engineer, and site reliability engineer (SRE).

This section includes some general trends you should adopt.

Automate As Much As Possible

If your organization does moderate to extensive data wrangling manually, it may be a symptom that the data engineers lack programming skills or a missing tool in the data pipeline. If a task is done manually more than twice, you should be looking at how to automate it. In general, look for every way possible to remove the humans from the loop. Automation will increase the reproducibility of results and infrastructure setup. However, due to a lack of tools and programming skills, a data wrangler may not fulfill this need. (Don't forget, reliable automation will require good data!).

Introduce Software Engineering Practices

Although occasionally a data scientist learns the software engineering rules and best practices that come along with

creating code, most data scientists don't understand what's needed.

Some engineering best practices are bare minimums, while others are good to have but dispensable. For example, I consider source control to be a bare minimum for a data science team. On the other hand, continuous integration and continuous deployment (CI/CD) are areas where the data engineering team can set up the process and show the data scientists how to use it.

Monitoring and Instrumentation

It's up to the operations team to ensure adequate monitoring and instrumentation of the programs in production to troubleshoot and recover quickly from problems. This interaction could entail the operations team examining the code during code reviews. The operations team will look for two main issues that other teams probably aren't looking after. First, exception handling: Is every significant potential error logged? Is the piece of information or the program's state logged to help the developers reproduce the issue? Second, process instrumentation: Do you know how many messages or objects are sent or processed and how long they took?

Security

Security is usually the purview of the operations team. Security in distributed systems can generally be broken down into encryption at rest, encryption over the line, authorization, and permissions.

The data engineering and data science teams often don't think about the security aspects of systems. They usually don't have an operational mindset and aren't considering all the way through the system design to the security portion. It's up to the operations team to ensure that the other teams adhere to security best practices in their code.

There is a fine line between keeping things secure and making things so secure that no one can even access the data they need. These overly secure environments tend to create shadow IT where people start to start up their own, less secure hardware that bypasses security restrictions. By being too secure and paranoid, some organizations become less secure.

03

Retraining Existing Staff

Modern data can improve the delivery of products and services while opening new markets and streamlining company operations. But before putting a data team in place, company staff from the chief officers down to the knowledge workers at their desks have to be prepared to use this data. The kind of data transformation discussed here is different from business intelligence – the transformation goes deeper.


Data literacy is a key part of decision making in today's era of data. Data consumers must be willing to make decisions based on trusted data and to override their intuition or experience. On the other hand, they must be willing to use that very intuition and experience to question the provenance and accuracy of the data, not relying blindly on the predictions of the analysts.

Some companies want or need to look internally for data engineers. More than likely, these people won't have all of the skills to be a data engineer from the beginning. Therefore, it is incumbent upon the organization to give these new data engineers the resources to succeed. These resources include learning new programming languages, new distributed systems, and new ways of creating data pipelines.

When doing internal promotion, look for people with a background in multithreading, multiprocessing, or client/server systems. From there, look for a desire to learn and expand their technical stack.

On top of all this, they need to show a love of data. This love manifests as having an interest in processing data and value creation from data. They may have already coded some analytics or data pipelines on a small scale.

It's much harder to retrain existing staff to become data scientists. For DBAs, the lack of programming skills (other than SQL) and software engineering skills forms a prohibitively high barrier. And the tools used by data scientists, such as Spark for streaming data and Kubernetes



for resource allocation, look nothing like the tools they've used. For programmers, math and statistics are tall barriers.

04

DataOps

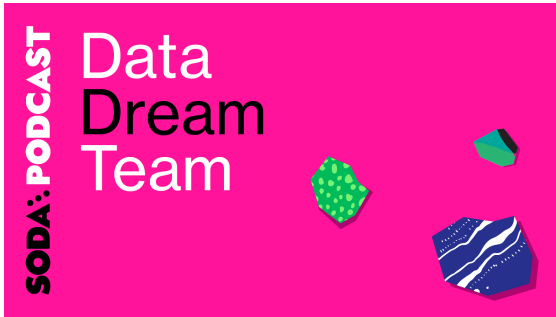
Some teams are discussing a fast-paced approach to data science that resembles the DevOps trend of the past 15 years. DataOps teams produce results in a time span of a few days, not weeks like most data science. The approach should not be used for all data science or data engineering, because DataOps does not produce robust enough results to use repeatedly for long periods of time. In other words, some robustness is sacrificed for speed.

The critical distinctions between DataOps and data engineering involve the speed of cycles. Data engineering – like most software engineering – works on more prolonged periods. DataOps realizes that business conditions change quickly. Each business competes with others to discover and respond to change; organizations with well-established data teams call for new models to be production-ready and deployed as fast as possible. To accomplish this, DataOps works on a much shorter cycle.

A DataOps team is cross-functional and is mostly – if not entirely – focused on creating value through analytics. This cross-functionality allows the team to create end-to-end value from the data because they can take raw data and create value from it. Thus, a DataOps team has data engineers, data scientists, data analysts, and operations on it.

DataOps should give the organization a way to deal with the issues – and issues will come up – while taking unreasonable burdens off of individuals' shoulders. The DataOps team should be focused on gradual and continued improvement until there isn't a need to burn the midnight oil anymore.

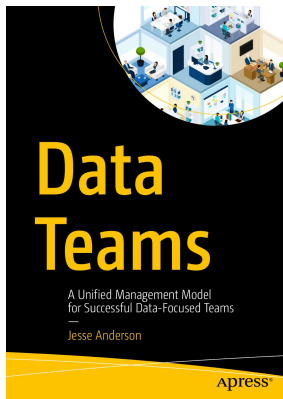
Our Podcast



This podcast series brings together amazing guests to talk about building data teams and share best practices from their data careers.

[Listen](#)

Jesse's Book



Data Teams: A Unified Management Model for Successful Data-Focused Teams

Successful organizations focus on both technology and management. Learn how to run big data projects, how to resource your teams, and how teams work with each other to be cost effective.

[Buy](#)



Want to learn more?

+32 2 786 43 65 (Europe)

+1 (646) 583-0116 (North America)

info@soda.io

soda.io

Let's Talk