



6 KEY CHALLENGES TO BUILDING A SUCCESSFUL DATA TEAM

And How To Overcome Them
To Create Better Data Projects



A WHITE PAPER BY DATAIKU

www.dataiku.com

THE CHALLENGES

The objective world of data can sometimes be fraught with highly subjective challenges, and companies that create teams to work on data projects face a number of obstacles as they ramp up their operations.

Many of these challenges are centered on this need for collaboration between IT and business profiles. Common mistakes, such as using static data or not thoroughly planning a solution's implementation, can trip up a young data team before they complete their first proof-of-concept. As data teams mature, the challenges do not go away but instead take different forms, like deciding whether to stick with older technologies (SAS, SPSS) or opt for newer approaches (R, Python, Spark).

This whitepaper aims to address these challenges and offer solutions that are applicable to all teams working on data projects whether they are just starting out or are already established.

THESE CHALLENGES WILL INCLUDE:

- [Real-Time, Dynamic Data](#)
- [Workflow Reusability Over Time](#)
- [Collaboration - Or Lack Thereof](#)
- [Skill Set Disconnect](#)
- [Operationalization Growth](#)



CHALLENGE 1

REAL-TIME, DYNAMIC DATA

One of the biggest mistakes data teams can make is attempting to answer a business question with old, static data. Today's data isn't just big, it's massive - and it's dynamic (i.e., constantly changing). Not taking advantage of this by creating data projects that can run on real time data for up-to-the-minute insights is a mistake.

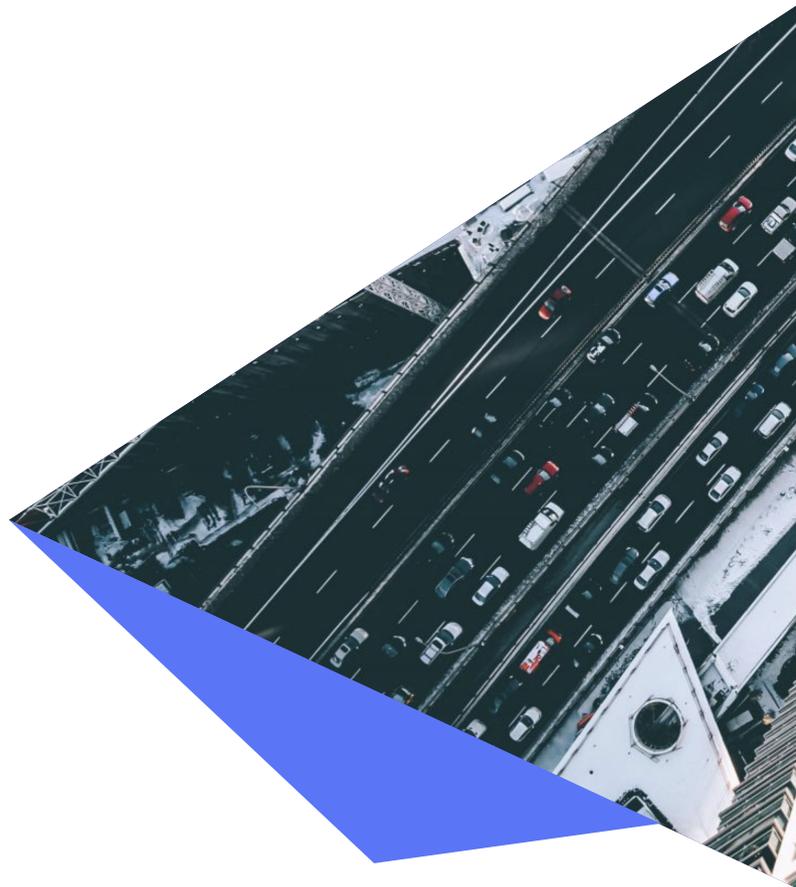
For example, take this (unfortunately) common scenario: let's say you've been presented with a business problem, and your team has been given six months to come up with a solution. You spend months cleaning data, building models, and refining information into visualizations according to the initial project parameters.

At last, six months later, you present your work to the business team. Your confidence is high, but it vanishes quickly as your customer bluntly states, "Great! Unfortunately, the original data has changed... can you re-do the same thing, but on the new data?"

You've wasted six months of effort and time, and now it's back to the drawing board. Perhaps four additional months pass as the data is refined and tweaked again, only to be told that the original project parameters have changed (yet again). Rinse, repeat. The vicious circle has only just begun, and with no particular end in sight.

The end result of this dilemma is that it is very difficult to make meaningful conclusions from static sources of data. The problem being solved is necessarily one of a dynamic nature, and so it is critical that the sources of data are able to capture that fact.

In other words, the expiration of data constantly invalidates your models' relevance. If your data team works in a bubble, then the solutions struggle to have relevance or provide value outside that bubble.



CHALLENGE 1: REAL-TIME, DYNAMIC DATA

SOLUTIONS

ESTABLISH CROSS-TEAM COMMUNICATION

This means aligning customer (in this case business team) expectations with the realities of data projects. Establishing such a channel of communication ensures that you and your customer have a means to discuss project scope and to get on the same page when it comes to managing both parameters and data.

MAINTAIN DATA VALIDITY

Your team must have access to real-time dynamic data. This ensures that your models and visualizations reflect your customer's unique situation; consequently, there will be no disconnect between parameters and data validity.

TIP: CREATE A DATA SOURCE SCORECARD

In order to ensure the quality of data being used for a project, at the start, create a list of all desired data sources and find out:

- Who is the internal provider of the data?
- How often is the data updated?
- Is the data source 100% reliable?
- Where does the data originate from (e.g., API, production database, data warehouse, etc.)?
- What is the expected delay to initially receive the data and/or get updates in the event of a data source change?

Implementing a comprehensive solution that fosters both communication and the data workflow process. Robust data connectivity, web-based access, and collaborative features in a familiar platform are some of the requirements that would address the realities of dynamic data while also supporting your data team's communication needs.

CHALLENGE 2

WORKFLOW REUSABILITY OVER TIME

If data teams are creating models in a vacuum, so to speak (that is, locally on their own machines or in an environment that lacks transparency), solutions will need to be continually reproduced, costing time and money. When companies do not operate with clear and reproducible workflows, it's very common for people working in different parts of the company to unknowingly be working on creating exactly the same solution.

Moreover, a lack of transparency and lack of workflow reusability generally mean there are poor data governance practices happening. Imagine if no one understands or has clear access to other members of the data team's work - in case of an audit, figuring out how data has been treated and transformed as well as what data is being used for which models becomes nearly impossible. With members of the data team leaving and getting hired, this becomes exponentially more complicated.



CHALLENGE 2: WORKFLOW REUSABILITY OVER TIME

SOLUTIONS

Create data projects that are based on a reproducible workflow. That is, the movement of raw data through different processes such as cleaning, enriching, modeling, and ultimately, the delivery of a new dataset. Instead of placing the emphasis on the data, which may or may not be relevant due to environmental circumstances, the focus should be on a workflow that can be replicated in the future on different datasets or for different projects as needed.

TIP: CHOOSE THE RIGHT TOOL

Technology selection for data teams should favor stacks and tools that simplify reproducibility and reusability. When selecting a tool/stack, ask yourself four questions:

1. Does the tool favor building a workflow instead of static analytics?
2. Would the workflow be simple enough to build so that business or data analysts could easily use and understand it?
3. Would the tool be elaborate enough so that my data scientists will choose to use it instead of their favorite notebook?
4. Does the tool touch on all aspects of delivering a data project, including data connectivity and setup, model building, reproduction of models on new data, testing, and production?

Ensure the data team can quickly test findings during production. This amplifies the power of reproducible workflows because they can then be rapidly implemented without having to constantly recreate the entire data flow. The use of reproducible workflows enables IT and business teams, particularly in Agile environments, to rapidly proceed through iterations within the workflow itself without disrupting the evolutionary development of the data.

CHALLENGE 3

COLLABORATION

In order to truly build a data-driven enterprise, both technical and business-oriented profiles must be involved - not just in their separate capacities, but working together for the best results.

For example, technical contributors should also have an understanding of the business requirements such as scope, cost, deadlines, data types, and visualizations required. But on the other hand, business profiles need to understand where the data is coming from, whether it's reproducible or not, the data workflow, and how frequently the data needs to be updated.

The challenge is that both profiles use different tools, different practices, and have different expectations. The world of Python, R, and Spark is quite foreign to the world of PowerPoint, Word, and Excel. And Agile, Kanban, and Lean live far apart from Six Sigma, PRINCE2, and OPM3.

In addition, the expectations of both parties can be markedly opposed. For example, an IT team may prioritize efficient functionality over usability while their business-minded brethren may focus on comprehensive reporting over a lean and efficient architecture.

These incompatibilities create a breeding ground for misunderstandings and, if not addressed, can develop into barriers that critically hamper productivity. The key message here is that if IT, data science, and business personnel cannot work together, then the data project will never come to fruition; it will be lacking in either completeness or accuracy (or both).



CHALLENGE 3: COLLABORATION - OR LACK THEREOF

SOME SYMPTOMS OF SERIOUS COMMUNICATION BARRIERS INCLUDE:

FIRST COMMUNICATION BETWEEN TEAMS AT THE END OF THE PROJECT

Successful data teams and data projects involve experts in IT, business, and data science from the start. Pulling in expertise at the last minute when most of the work is already done is extremely costly and is a sign of deeper, dysfunctional communication.

LACK OF STRONG LEADERSHIP

If team leaders don't support horizontal collaboration (between team members with the same profile or background, like between data scientists for example) as well as vertical collaboration (between different types of profiles, like between business and IT), data projects are doomed to failure.

PROBLEMS WITH TRACKING AND VERSIONING

It doesn't take long for email threads to grow in length. Using email to share files is a recipe for disaster when it comes to keeping track of content and for data versioning. Expect the loss of data and non-inclusion of key stakeholders.

MISCOMMUNICATION

Electronic communication does not convey emotions nor does it foster inclusion. Email or chat can cause problems when colleagues discuss important issues and a misunderstanding ensues. In addition, such methods are not inclusive — continually CCing relevant parties is not an efficient way to promote healthy conversation.

LACK OF STRONG DATA GOVERNANCE POLICIES

Organizations typically implement policies for the sharing of content and data protection. Emails are typically not safe and are often used by employees who are either unaware of communication policies or wish to use a workaround.

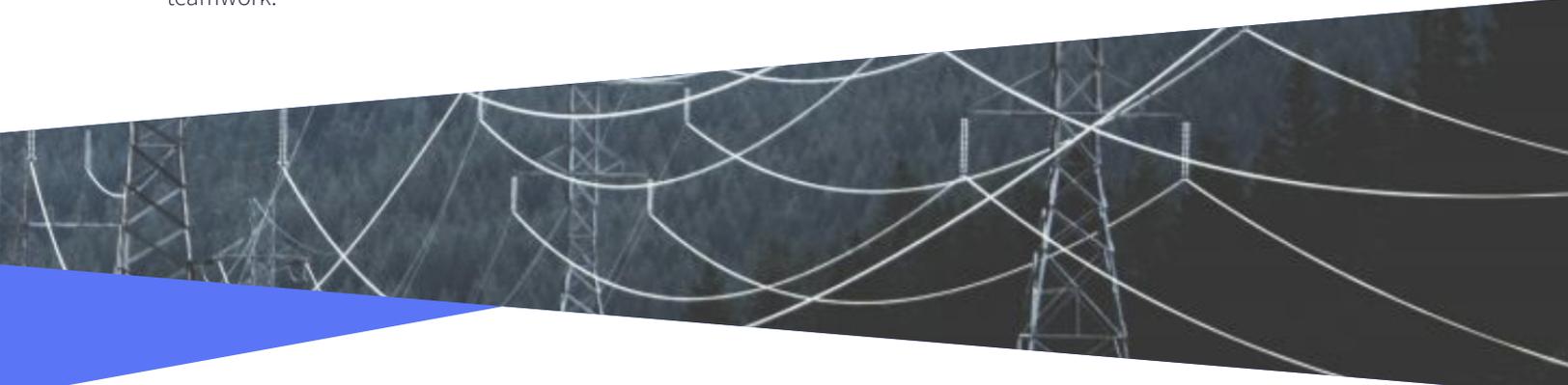
SOLUTIONS

COLLABORATIVE, WORKFLOW-CENTRIC TOOL THAT IS ACCESSIBLE TO ALL TEAM MEMBERS

For example, team members with different skill sets should all be able to contribute to the success of a data project as a whole: novice data scientists can clean/enrich the data and prototype basic models, while experienced data scientists can modify the models for improved results, and business analysts can add insights into the relevance of the model based on the project requirements.

EVANGELIZATION OF COMMUNICATION ACROSS TEAMS FROM SENIOR MANAGEMENT

A guiding organizational hand needs to be present in order to act as a bridge between parties. It's not necessarily about solving individual debates but rather more broadly emphasizing collaboration by creating a shared foundation of teamwork.



PROJECT TRANSPARENCY AND ACCESSIBILITY SUPPORTED BY A TOOL THAT FACILITATES INCLUSION OF ALL TEAM MEMBERS REGARDLESS OF SKILL SET AND EXPERIENCE LEVEL

The sharing of critical data and related conversations that revolve around that sharing should take place within a collaborative, real-time, web-based environment. Such environments empower users to communicate with each other in a shared setting — content is both transparent and accessible. Versioning and keeping track of important data, a must for any platform, enables users to gain a 360-degree view of their content in terms of change management.

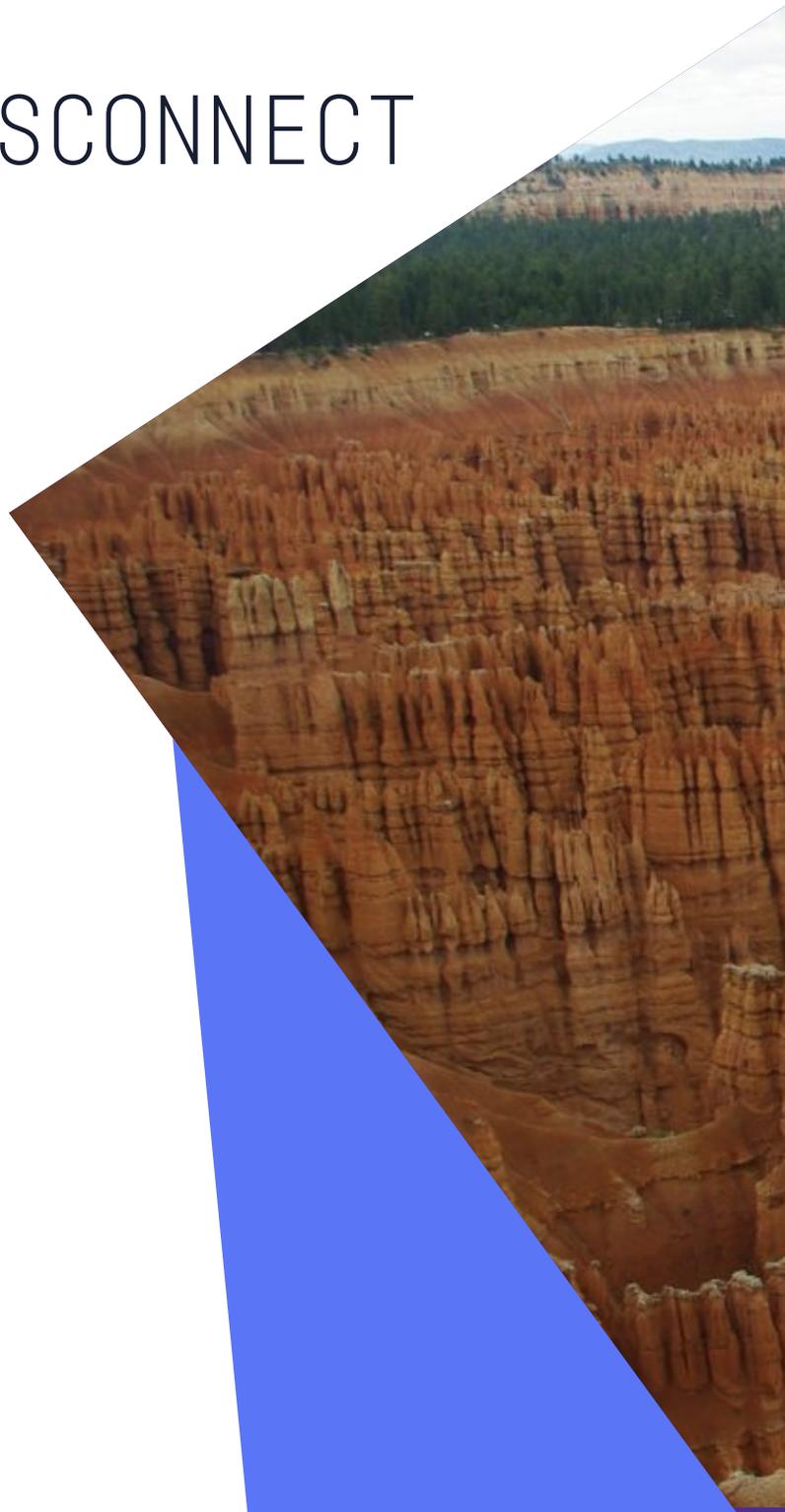
CHALLENGE 4

SKILL SET DISCONNECT

If finding data scientists for your team is a challenge, then finding data science talent with the right skill set to fit your organization may be a nearly insurmountable obstacle. The core issue here is that there is a disconnect between the knowledge traditionally used in data science compared to the skills taught in higher learning institutions.

For example, older technologies for statistical analysis (such as SAS and SPSS) were in place when established companies first launched data teams, and the learning curve for these older technologies (particularly given their age and complexity) is significant. Most new graduates branded as data scientists have completely different skills rooted in newer technologies like R, Python, Spark, Pig, Hive, etc.

The end result is two sets of data scientists, both representing different generations of statistical analysis methodologies. The challenge of old vs. new technology has exacerbated in recent years due to the growth of the data industry coupled with the need to hire new talent.



CHALLENGE 4: SKILL SET DISCONNECT

SOLUTIONS

From a human resources standpoint, there are essentially three paths available, each with their own respective pros and cons:

ABANDON OLD TECHNOLOGIES AND SWITCH TO NEW TECHNOLOGIES

This enables data teams to hire new data scientists that can onboard quickly and become productive with little downtime. On the other hand, changing the core architecture of a data science team has its own ramifications to both existing employees and the development process as a whole.

KEEP OLD TECHNOLOGIES AND TRAIN NEW HIRES

The immediate benefit of this approach is that, unlike switching to a new platform or new technologies, there is no imminent disruption to your data team's productivity. The downsides revolve around the new hire learning curve and the possibility of becoming an antiquated data team over time, not able to adapt to the newest technologies or hire top talent.

KEEP OLD TECHNOLOGIES AND PURSUE NEW TECHNOLOGIES [AKA THE HYBRID - AND RECOMMENDED - APPROACH]

A third approach, which is likely the most appropriate for the vast majority of companies and use cases, combines both of the above options: keeping the old and using the new technologies in parallel. This can happen through the implementation of a [data science tool](#). In this scenario, flexibility gives established employees the freedom to continue development using older technologies while new employees can develop using the new technologies. In other words, the data team doesn't have to make sacrifices -- both paths are pursued at the same time.



CHALLENGE 5

OPERATIONALIZATION

Data teams often face challenges due to poor project planning. In the beginning stages, the team spends a lot of time discussing the problem and how to arrive at the best solution. Yet the plan for the actual operationalization of the solution (that is, having the model acting on real-world, real-time data) is often only a minor consideration - a big mistake.

For example, let's say an advertising company had set up a new Hadoop cluster that enables analysts to run SQL queries against normalized project data. The purpose was to redesign the client's information system; but the client's production environment was not compatible with the data team's technology stack.

The end result was an unforeseen extension of the project coupled with the associated time and expenses. The most damaging aspect of this human error was that it put the data team in a very difficult position: they had already completed a significant amount work, but there was no way for that work and those projects to be implemented.

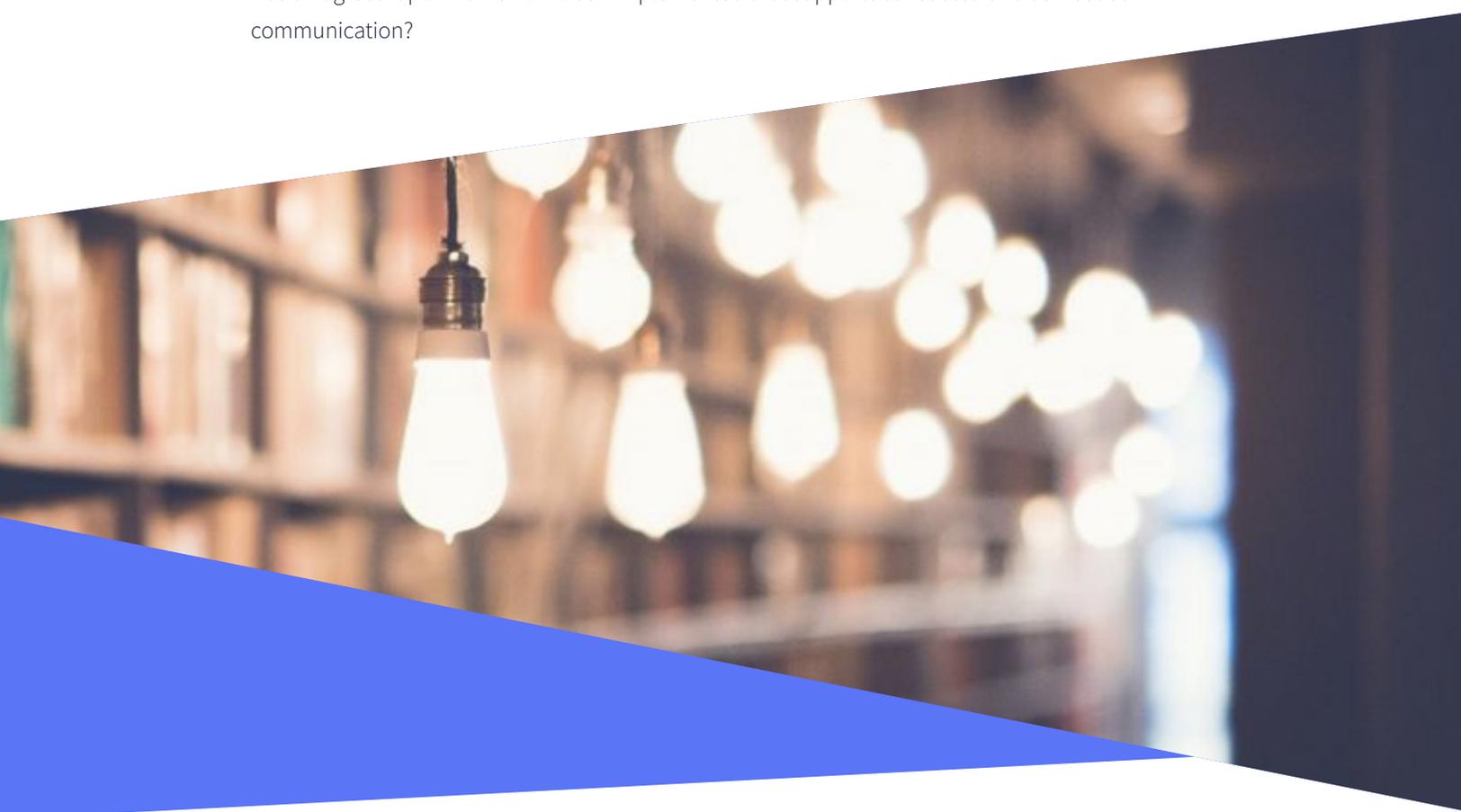


CHALLENGE 5: OPERATIONALIZATION

SOLUTIONS

FAILURE TO CONSIDER O16N CAN BE ADDRESSED BY ADDRESSING SOME KEY QUESTIONS BEFORE WORK BEGINS:

- Has the project been comprehensively researched from handoff to deployment?
- Does your team have access to the client's production environment?
- If not, can your team replicate the client's production environment?
- Does your team have access to actual real-time data?
- Is there an established communication channel between your team and the department who has requested the development of the data solution?
- Has an agreed-upon framework been implemented that supports scheduled and as-needed communication?



CHALLENGE 6

GROWTH

The day may come when your data team, despite its growing pains and obstacles, overcomes its primary challenges and becomes well-established. Unfortunately, the pitfalls and challenges of being a part of a mature company with a growing team can be just as difficult as those faced by up-and-coming data teams.

For example, a data team whose parent company is an investment bank may be focused on using predictive analytics to model currency fluctuations — their primary work revolves around producing accurate models and visualizations of potential international currency changes based on their parent company's business initiatives. Although the team is primarily focused on currency movements, they decide they would also like to develop a customized software solution that uses basic modeling to produce generalized financial advice (like asset and pension management).



The problem here is that the technology ecosystem is comprised of many moving parts, and there are a significant number of variables involved in developing solutions. If the solutions already exist, this data team would be reinventing the wheel. The temptation to extend a business enterprise into uncharted territory is often born out of a desire to produce a “comprehensive” solution that addresses all possible customer needs.

This particular road can be a long and tumultuous one, as the development of complex software from scratch is costly, time-intensive, and difficult to maintain over the long term. Pouring resources into a new enterprise may ultimately compromise the growth potential of the data team itself.



CHALLENGE 6: GROWTH

SOLUTIONS

Know that data teams pursuing customized software development are in for a significant investment. If your data team wants to implement a growth initiative, then perhaps instead consider investing in business applications that create competitive advantages as opposed to technical foundations. Building solutions from the ground-up can be a herculean task, so why not fill a need from an existing vendor using open source technology? Limit costs and save time by leveraging existing technologies to solve unique business needs within the vendor ecosystem.

Additionally, every data team should start small and build slowly. Once they are able to prove success with smaller projects, it can then become viable to roll out bigger initiatives/roll out those projects and processes more widely for a more gradual transition.

As your team begins to grow to take on other projects, consider:

- Does your initiative serve a generic need from an ecosystem or platform (e.g., Hadoop, Spark)? If so, is there an open source project or a commercial product that could serve this need? Could my organization sponsor such an initiative?
- Does my team have the necessary skill sets and experience for the product management of a technical framework?
- Connect to the open source community and the vendor's ecosystem. The technology ecosystem has a lot of moving pieces, so knowing the roadmap in advance will help you focus your platform's development in terms of what's really needed.

CONCLUSION

Data teams are complex, nuanced organizations with different kinds of people using different tools yet all working toward the same ultimate end goal. If the data team is not a well-oiled machine, the end goal (data projects) suffer. They may not ever be seen through to completion or they may be inefficient or ineffective. In other words, when one piece of the data team isn't working right (even if it seems minor), things can go awry.

Keep asking yourself about your potential for growth. Is your team limiting itself? Or continually reproducing the same types of projects over and over from the beginning? Then it may be time to examine your processes. Can your workflows be easily reproduced? Are your methods as efficient as they should (or could) be?

A successful data team also faces many obstacles that are not technical, one major example being how to overcome recruitment challenges. If your team and its methodologies, processes, or tools is stuck in the past, then you may be limiting your hiring profile to data scientists and data analysis with a very narrow skill set, potentially crippling your team's long-term growth. The challenge of handling future growth must be balanced with the reality of hiring team members whose skill sets are appropriate for your business model.

Some say identifying a problem is half the battle. So perhaps all that's missing in your data team is that one key component. We hope that this whitepaper has helped you to find what might be missing for you, whether it's a high-level collaboration platform or a simple reminder that communication is key.





Your Path to Enterprise AI

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to enterprise AI. Data-powered businesses use Dataiku to power self-service analytics while also ensuring the operationalization of machine learning models in production.

20,000+

ACTIVE-USERS

*data scientists, analysts, engineers, & more

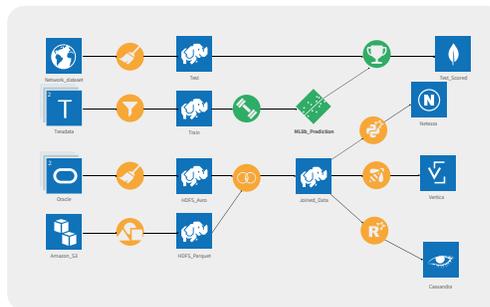
200+

CUSTOMERS



1. Clean & Wrangle

Name	Sex	Age
Morales, Mr. Owen James	male	22
Moran, Mr. James	male	26
Haskins		26
Kurowski, M		25
Allen, Mr. W		35
McCarthy,		35
Rowland, W		29



5. Monitor & Adjust



2. Build + Apply Machine Learning



3. Mining & Visualization



4. Deploy to production





WHITE PAPER

www.dataiku.com