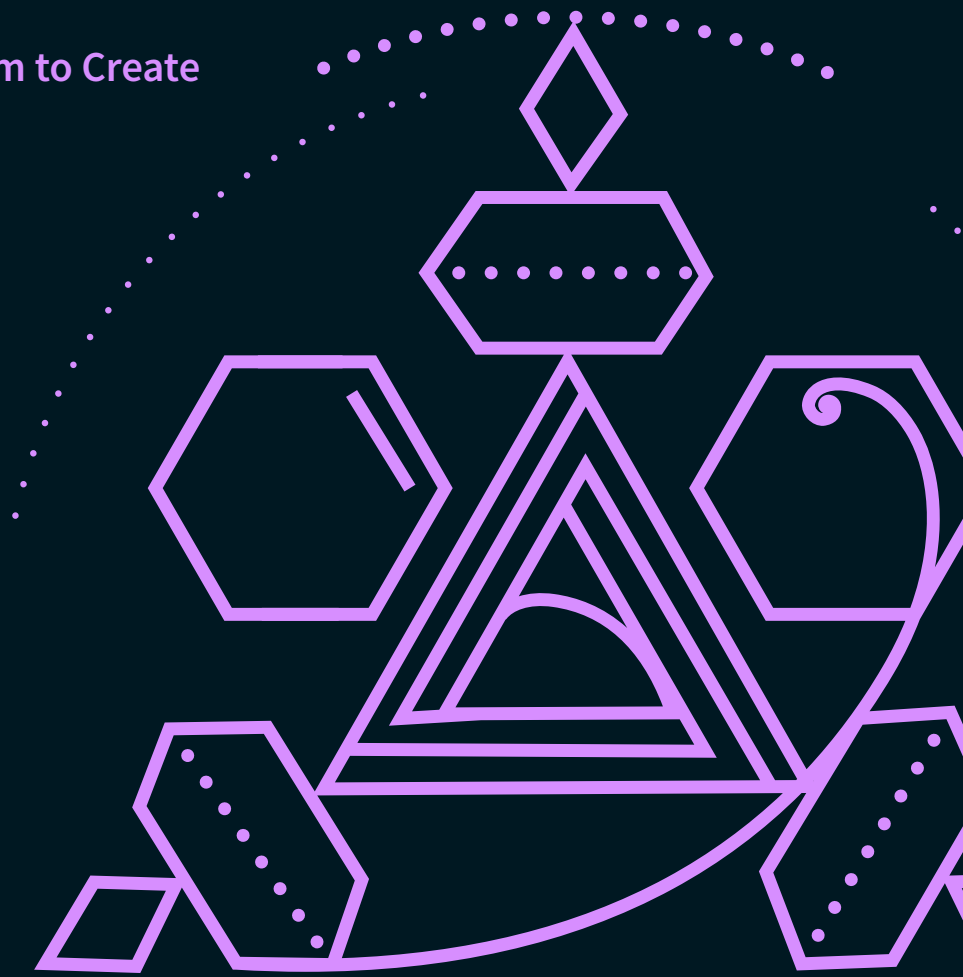
data iku

THE 5 KEY CHALLENGES TO BUILDING A SUCCESSFUL DATA TEAM

And How to Overcome Them to Create Better Data Products



WHITE PAPER

www.dataiku.com

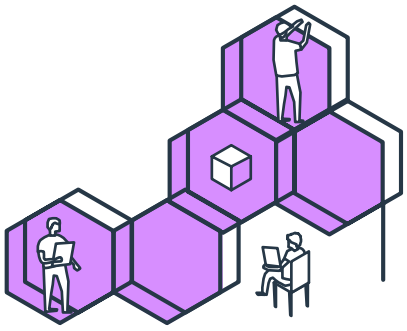
SUMMARY

INTRO DUCTION

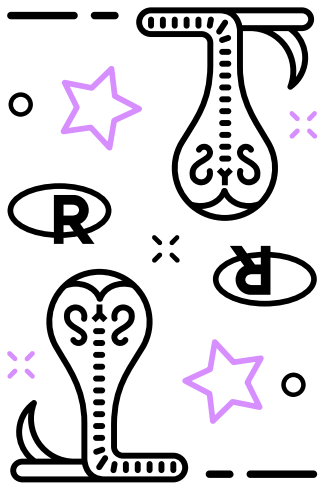
The age of data is here. Sensors, cameras, security monitoring systems, software, hardware, the internet, and even humans themselves all have one thing in common: data. ***And as data becomes ubiquitous, in parallel, business interest in using and learning from that data is on the rise.***

Enter big data, a holistic term that aims to encapsulate the sheer massiveness of the amount of information available from the growing number of data sources. As data storage capabilities have grown, IT teams have made a significant effort to collect data. But the reality is that many people (and in fact many organizations) still don't know what to do with all of that data. Once they get to the point where they're collecting massive amounts of data, teams struggle to answer the question: **now what?**

In an attempt to make use of big data, businesses have started to assemble or hire teams tasked with deriving insights from the sea of data they're collecting. Companies with data teams hope to answer business questions and address needs using this raw information. But when these teams are well organized and given the right tools and direction to succeed, they can do much more: **data teams can serve as research and development departments, experimenting with raw data to explore possibilities and solutions that the business didn't even know it had.** This can include anything from uncovering new insights about customer behavior to revealing business opportunities in new markets.



Successful data teams are innovative and creative but are also able to get past the experimental stage to actually tackle difficult business problems. But building this team isn't as easy as hiring a staff and letting them go to work - companies with high-performing data teams empower them to be a bridge between IT and business departments to evangelize creative innovation while also finding (and deploying) real solutions.



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.

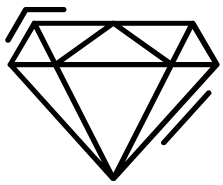
Many of these challenges are centered on this need for collaboration between IT and business profiles. Common mistakes, such as using static data or not thoroughly planning a solution's implementation, can trip up a young data team before they complete their first proof-of-concept. As data teams mature, 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.

Time-Based Challenges

One of the most alluring aspects of a career in data science is the level of confidence instilled by having all the answers in the data and at your fingertips. Data is at your command, 24/7/365, and it's simply a matter of molding and shaping it in the right way. **So often the real challenge is not having the data to answer a specific business question, but time.** Time can derail even the most well-planned of projects, and in many respects, it poses the greatest challenge to even the most well-oiled teams working with data.

For example, say you've been presented with a particularly malleable problem and your team has been given six months to discover its solution using all the tools at your disposal. **You have access to the raw data,** and your customer (perhaps another team at your company) has made the desired end results of the project clear. The members of your team (comprised maybe of data analysts, business analysts, data scientists, or some combination of the three) are intimately familiar with the overall process of slowly transforming raw data into highly relevant information with an applicable purpose at the business level: at times it can be laborious, but the end-result makes the effort well worth the work. Your team cleanses data, builds models, and refines information into visualizations according to the initial project parameters.

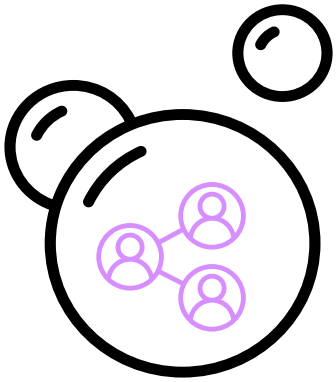


At last, six months later, the lump of data coal has been transformed into an exquisite crystalline diamond of information. The confidence is high, but it vanishes quickly as your customer bluntly states, "Great! Unfortunately, the original data has changed… can you re-do the same thing, but on the new data?"

You've wasted six months of effort and time, and now it's back to the drawing board with the updated data and repurposed problem. 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). "Oh, I thought we e-mailed you..." doesn't help matters. Rinse, repeat. The vicious circle has only just begun, and with no particular end in sight.

Welcome to the world of dynamic data, where the time-intensive nature of working with raw data meets the reality of project management chaos. The problem? **Data usability is thoroughly compromised due to project length and the lack of accessibility to accurate data.** It's like playing tennis while surfing — your foundation is in constant flux, changing in real time while your decisions are based on information that is simply not relevant anymore.

The core issue of this challenge is that the project length, and perhaps model build length, are not aligned with project expectations. You're running a marathon while your customer is doing sprints. Secondarily, the nature of data itself is a factor: data is not static; it changes constantly, but it needs to be used in the here and now.



Compounding the time issue is the fact that data projects are often driven by business needs, which means that the solution will serve as a basis for decisions made at a very specific point in time. Cost points, fee structures, and project scope often change overnight due to external factors.

The end result of this dilemma is that it is very difficult to reproduce the initial solution; repeated attempts become not only increasingly frustrating but also tend to prolong the process.

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

SOLUTIONS

Effective communication from the very beginning is a critical factor to ensure that projects based on dynamic data result in meaningful insights and results. This means aligning customer 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. Of equal importance is 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.

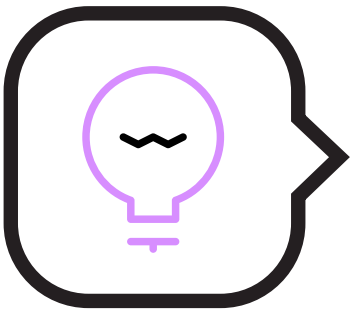
STEPS TO RESOLUTION

The common denominator here is communication: between the data team and the customer and, perhaps most importantly, internally among members of the data team itself.
This means not only the sharing of information but also of data and the role it plays in the modeling workflow.

The best approach is the implementation of a comprehensive solution that features a variety of tools designed to foster 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.

SOLUTION SUMMARY

- Implement a comprehensive solution that addresses both data accessibility and communication requirements.
- Establish a mutually agreed upon channel of communication.
- Ensure that your data team has access to real-time dynamic data.



Tip: The Data Source Scorecard

There are some pathways to data modeling that may be interesting to pursue but that don't necessarily add anything of value to the process. A useful tip when starting a new project is to look at data connectivity with an eye toward discarding irrelevant data sources.

There is no need to build a model on data that won't be available at decision-point. At the start of a data project, create a list of data sources and their related connectivity.

The list should address:

- ☐ Who is the internal provider of the data?
- ☐ How often do I need to update the data (i.e., anywhere from yearly to real-time) to make an accurate prediction?
- ☐ 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?

Workflow Reusability Over Time

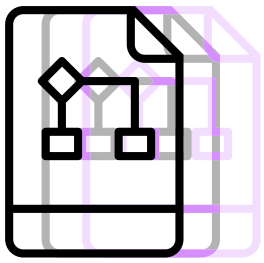
Whether it's six months or a year, time can wreak havoc on project planning when it comes to working with data. In the previous section, we discussed the challenges inherent in problem solving with a six-month timeline. Unfortunately, the same challenges exist when the timespan is extended to a year or more.

There is often a preconception about the permanence of data — if it's accurate now, then why not one year from now? **At times, data teams can work with blinders on, focused entirely on meeting project deadlines and not realizing that the operating premise itself has become flawed; the foundation is not stable.**

Again, dynamic data rears its ugly head as changes in technology, languages, and stacks invalidate the usability of time-sensitive solutions. One specific example is the development of a solution preceding a holiday with the intent to reuse the solution annually (i.e., for subsequent holidays). The solution may work perfectly in 2016, but when 2017 rolls around, it may have completely lost its relevance.

The model is unable to adapt to the multitude of technological and business changes brought throughout that year. The end result is that solutions need to be continually reproduced, costing time and money.

SOLUTIONS



A key solution is to 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 the lapse of time, the focus should be on a workflow*** *that can be replicated in the future as needed.*

A reproducible workflow coupled with up-to-date data ensures that both the method used to reach the results and the data itself is relevant. Additionally, data teams must be able to quickly test findings in production without having to go back to square one. **A reproducible workflow guarantees that the process itself remains constant, essentially negating the need to reproduce the solution.**

STEPS TO RESOLUTION

Workflows enable the rapid implementation of data analysis without the need to recreate a process. ***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.*

A well-organized data team combines the skill sets of multiple employee profiles to determine exactly how to proceed with a given problem: data scientists, business analysts, and software engineers work together to narrow the scope of a task and frame questions in terms of quantitative clauses.

For example, a customer may start with the question, "How can our website earn more revenue?" and, after discussion and probing questions, the data team may start work with the premise, "How can we increase advertising bids per impression by 10 percent without reducing our cost per click rates?"

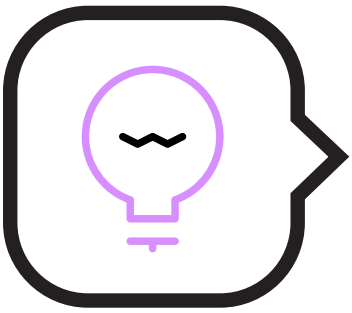
The process of data moving through the workflow now begins as datasets, and the team determines sample sizes, cleanses the data, and formats it properly.

A robust workflow should enable the data team to break off from the process and communicate insights with the customer (or internal stakeholder) in order to explore more possibilities, such as additional datasets or modifications to the original problem. This feedback cycle enables the data team to dig deeper into the data analysis process and to refine the model accordingly.

These iterations may repeat frequently, yet they should not disrupt the overall progress of the end-to-end workflow. Eventually, viable models and visualizations will enable both the stakeholders and the data team to move forward with putting the solution into production.

SOLUTION SUMMARY

- Create data projects based on a reproducible workflow.
- Facilitate flexible workflows that support collaborative feedback iterations.
- Data teams must be able to quickly test findings during production.



Tip: Choose the Right Tool and Adapt Project Management

In order to achieve efficiency on a team working on data projects, you should favor technological stacks and tools that simplify reproducibility and reusability. When selecting a tool/stack, ask yourself three questions:

- ☐ **Does the tool favor building a workflow instead of static analytics?**
- ☐ **Would the workflow be simple enough to build so that my business or data analysts would actually use it?**
- ☐ **Would the tool be elaborate enough so that my data scientists will choose to use it instead of their favorite scratchpad or notebook?**

In addition, you should adapt project management to position reproducibility as a relevant milestone:

- Data connectivity and setup
- Model building / proof of concept
- Reproduction of models on new data
- Testing
- Production

Collaboration – or Lack Thereof

We've all heard of the left-brain and right-brain argument: logic / analysis vs. creativity / emotion. In reality, there are few absolutes within any population; the interests and skill-sets of humans vary widely and with significant input from our genetic make-up, experiences, and education. In short, it's not a black and white world… there are plenty of gray areas.

There are times, though, when these differences can clash or be especially pronounced. One such example is technical vs. business profiles, sometimes (not-so-lovingly) called the oil and water of the tech world. While there can be areas of conflict, often exacerbated by the fact that both parties look at things a bit differently, there are still plenty ways to build bridges of understanding, collaboration, and increased communication.

CULTURAL COMMUNICATION BARRIERS

It isn't that those with technical profiles don't get along with people that have more business-oriented backgrounds and skills; in fact, they can work together quite well. **The issue is that both professions 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. And when business analyst and data science colleagues on the same team operate from independent silos of knowledge, the root of the problems can grow even deeper.

The reality is that collaboration between IT and business is vital to the success of data projects. Business and IT teams need to understand the problem's challenges from each other's perspectives. Do IT teams understand the business requirements, such as scope, cost, deadlines, data types, and visualizations required? Conversely, do business teams 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?

By understanding the big picture — the project's holistic view — each team will be more inclined to find a collaborative middle ground of understanding. Given this, it's imperative to facilitate collaboration between teams in order to ensure the success of any data project.

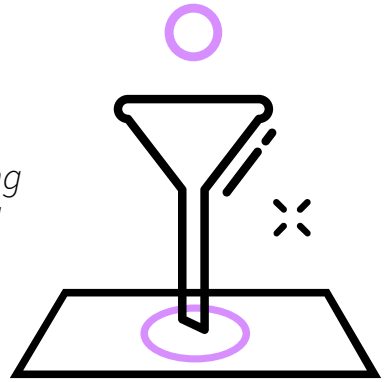
Some early symptoms of cultural communication barriers include:

- Communication on an as-needed basis instead of consistent communication
- Lack of flexibility in terms of both approach and project planning
- Absence of a collaborative work environment and tools
- Failure of leadership in terms of proactively fostering cooperation and collaboration between teams
- Meetings with no concrete results (i.e., each party espousing their own views, assignment of blame, little interest in understanding each other's challenges, etc.)
- Leadership shows favoritism for either IT or business teams
- Lack of strong leadership in finding common ground for the project and respecting its deadlines

The key message here is that if IT and business teams cannot work together, then the data project will never come to fruition; it will be lacking in either completeness or accuracy (or both).

Each profile has its own unique contribution to data projects as a whole — business teams can contribute insights that IT teams would never have considered, and IT teams have the capability to use technology to build meaningful data products that less technical profiles would not consider. Business profiles need to understand what can be done from a technical perspective, while IT profiles need to understand what needs to be done from a business perspective.

SOLUTIONS



There is a saying that "The fish stinks from the head down," meaning that problems originate from the top (i.e., leadership) and progressively manifest themselves in the rest of the organization. This is particularly true when it comes to communication between IT and business teams.

The fact is that healthy communication, cooperation, and collaboration between all parties can most effectively be facilitated from a singular source. **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.

This means implementing collaborative environments and tools that all parties can use to facilitate both discussion and knowledge sharing. 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.

The project itself needs to be fully transparent and agreeable to all major stakeholders. Data sharing using collaborative channels is also a critical aspect, as the organic nature of data needs to be accessible to all parties to ensure relevant outcomes.

STEPS TO RESOLUTION

To put it bluntly, products created in non-collaborative and non-communicative environments are doomed to fail. **Successful data projects leverage collaborative tools that enable all participants to work together toward the same goal.** This means that all team members can use their unique skill sets and levels of expertise to advance the project through all workflow stages from project inception to delivery. Whether business or IT, novice or expert, all colleagues can realize project success when they are using the same solution as a basis for collaboration.

SOLUTION SUMMARY

- Collaborative, workflow-centric tool that is accessible to all team members.
- Evangelization of communication across teams from senior management.
- Project transparency and accessibility supported by a tool that facilitates inclusion of all team members regardless of skill set and experience level.

Technological Communication Barriers

A common theme among mature populations is that everything in the past was better. In the modern age, how many people do you know that insist on being tied to their desktop applications when robust web-based applications are available from any location in the world?

There are a lot of people who firmly believe that old equals better and, sometimes, they are probably right… but there are other times when they're just plain wrong. It's human nature to use what's comfortable and familiar. After all, getting out of one's comfort zone is a challenge and requires effort. For every early adopter, there is a staunchly rigid supporter of a technology that may end up doing more harm than good.

One area where outmoded solutions tend to persist is communications. **In the world of data teams, outdated communication solutions pose a significant risk on a number of levels.** From policy breaches to communication channels that aren't secure, using the wrong type of solution can have some drastic consequences. A very common scenario is when employees use email to share all kinds of information from login credentials to data for analysis.

What can go wrong? Let's take a look from the innocuous to the serious:

- **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.

- **Non-collaborative:** Using email and/or chat applications limits the participants to only those specified. If some team members are not included, they will be excluded from the knowledge sharing. It's hard to enable project success when everyone is not on the same page.

- **Corporate policy breach:** 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.

- **Malicious code:** Email, chats, and external forums are all subject to data compromise from malicious users and/or organizations. Using these methods can severely compromise the integrity of an organization and can result in the theft of sensitive data, typically via botnets, advanced persistent threats, and malicious code (e.g., trojans, virii, rootkits, etc.).

Traditional communication methods, such as email, definitely have their place within any team, including data teams. But when it comes to sharing knowledge, they fail miserably. Some common reasons why employees are reluctant to change is that they have a strong connection to the status quo. Another common reason for using inappropriate communication channels is a lack of evangelization from senior management. There may be no comprehensive mission statement with regards to establishing standards and transitioning from unsafe data management practices to collaborative, yet secure, practices.

SOLUTIONS

The goal for any data team should not necessarily be banning traditional communication methods; as previously mentioned, email has its place and is highly effective for one-on-one communication. **But 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.

The centralization of data access enables administrators to implement security measures such as user access control and air-gapped intranet deployments. Both data science and development processes can also be easily audited, a particularly valuable feature in highly regulated industries like health care and insurance. A centralized environment also enables management to track user activity and enforce data security policies more effectively.

STEPS TO RESOLUTION

Your data team will certainly not stop functioning if you use only email. In fact, many data teams today rely solely on traditional communication methods. The point is that your organization may not be reaching its full potential by neglecting web-based collaborative environments. **Email by itself does not promote collaboration and may even introduce security-related issues.** Indeed, many of the challenges discussed in this whitepaper originate from a lack of collaboration and engagement between different teams.

Collaborative environments provide more bang-for-the-buck, addressing not only communication issues but also empowering your organization to take control of its content via data security, version control, and policy implementation.

SOLUTION SUMMARY

- Data sharing and communication via a collaborative, real-time, web-based tool.
- Tool that supports the centralization of data, process auditing, version control, and data security.

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.***

OLD VS. NEW TECHNOLOGY

Teams that work with data frequently implement older technologies for statistical analysis such as SAS and SPSS. These solutions 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.

SOLUTIONS

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

1. Abandon Old Technologies and Switch to New Technologies

   In this situation, the data team changes its approach to development by abandoning older technologies in favor of newer options. **This enables data teams to hire new data scientists that can onboard quickly and become productive with little downtime.** But changing the core architecture of a data science team has its own ramifications to both existing employees and the development process as a whole. **By catering to newer technology, existing employees face the challenge of updating their skill sets.**

   Data team growth (particularly in companies with rapid market impact) often mandates this change from using old technology to its newer equivalent because the teams hire experienced data scientists in the beginning (well-versed in classical tools) but need to quickly hire younger data scientists (fresh graduates) in order to meet demand. This results in the dilemma of whether or not to train new hires on the older technologies (costly and tedious) or allow them to use newer technologies like R, Python, Spark, Pig, Hive (inexpensive and fast).

2. Keep Old Technologies and Train New Hires

This approach represents the opposite of the first: the team keeps older technologies, and new hires are trained to use them. The problem here is that, as previously mentioned, these older options are both complex and robust. They've simply been around longer and, consequently, require an advanced skill set in order to gain proficiency.

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.

With this option, new hires would always require a significant investment of both time and money. And not just the time for the new employee, but your experienced employees' time as well. The hiring process may also need to be extended in order to comprehensively implement best hiring practices. Given the amount of training required, you would need to ensure that your new employees are keen on contributing for the long term.

Generally speaking, technology improves over time in terms of efficiency. Older technologies obviously have the benefit of robust and mature support and, typically, a dedicated (though ever-dwindling) user base. The benefits of staying the course and investing in new employee training would have to be weighed against the possibility of eventually becoming a data team that uses outdated technology. Your team may fail to innovate as quickly as your competitors, and you may have difficulty when recruiting.

3. Keep Old Technologies and Pursue New Technologies (Hybrid Approach)

A third approach is a combination of the above options: keeping the old and using the new technologies in parallel. **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.

STEPS TO RESOLUTION

Obviously teams needs to hire people in order to grow, so it's difficult to avoid this challenge. And when a company first establishes a data team, that team has to prove their value; once they have delivered on their first project, the demand (and need to hire new employees) will increase.

The solution is to implement a tool that enables all team members, regardless of skill level and expertise, to work together. In a competitive market, a team devoted to data projects can only survive if it can reliably deliver results. This means a tool that is workflow-centric while supporting meaningful collaboration between all employees.

SOLUTION SUMMARY

- Tool that enables all team members, regardless of skill level and expertise, to work together.
- Tool that enables quick and efficient onboarding of new more inexperienced hiresand data security.

Platform Incompatibilities

We're all human and are prone to making mistakes; after all, nobody's perfect. ***Some mistakes are simple and easy to correct, but others can cause complete chaos.*** *In 1999, U.S. space agency NASA lost communication with the Mars Climate Orbiter. The space probe took a wrong turn and disintegrated on an unplanned re-entry. The reason? The software controlling the Mars Climate Orbiter produced output in English pounds-seconds instead of metric Newton-seconds. NASA wasted more than $327 million due to a simple unit of measurement error.*

This is an example of how a team can waste significant amounts of time, effort, and money during the implementation phase of a project. Unfortunately, the same situation can hold true with data teams.

SOLUTION IMPLEMENTATION IN CLIENT PRODUCTION ENVIRONMENTS

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 implementation of the solution in the production environment is often only a minor consideration.** But not properly planning this final stage of putting data projects into production could be a costly mistake if not addressed properly.

In one scenario, 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.

Responsibility for aspects of data project deployment falls to the data team, so in this case, the onus was on them to address and facilitate the correction of compatibility issues (and absorb the cost of doing so). Not only does this cause understandable frustration for the customer, but it critically damages the reputation of the data team throughout the rest of the company. **Ultimately, the data team's failure to be thorough (especially when it came to deploying final products into production) made them look amateur and probably lose trust and clout with other teams.**

SOLUTIONS

Like a rolling snowball gaining size and momentum, an absence of communication between a company's data team and the customer (whether that customer is another team at the company or external) can quickly get out of hand and create more problems as the project progresses.

Some key questions to address before work begins:

- **Project planning**
  - 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?
- **Communication**
  - 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?
  - Is your team working in a web-based collaborative environment?

STEPS TO RESOLUTION

To avoid platform incompatibilities, the above questions should be addressed at a very early stage. If it's too late and the challenge already exists, then there is no option but to pick up the pieces and try to solve the compatibility issues. But mistakes like these should be a one-time occurrence; if they happen again, then senior management should step in to ensure that processes are fully defined and that the team is following those processes. **The world of data is growing rapidly and, in a highly-competitive market, there is little tolerance for organizations that fail during the initial planning phase.**

SOLUTION SUMMARY

- Ensure that your data team has access to the client's production environment (or can replicate it).
- Ensure that your team has access to actual up-to-date data.
- Thoroughly research all aspects of the project's scope and parameters at an early stage.
- Establish communication channels between the data team and its customer.

Growth

The day may come when your data team, despite its growing pains and obstacles, overcomes its primary challenges and becomes well-established within its domain. As with any business, though, even data teams have more opportunities to pursue as they grow. But the pitfalls and challenges of being a part of a mature company can be just as difficult as those faced by industry newcomers.

REINVENTING THE WHEEL

As established data teams enjoy the fruits of their labor, there is often the temptation to develop new solutions to address problems for which there is already a viable solution.

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 unchartered 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.

SOLUTIONS

If your data team wants to pursue customized software development, then be prepared for a significant investment. Your team's success (or failure) in this arena will be highly dependant on your development team. Do not rely on a single individual but, rather, implement a common framework that can be quickly adopted by developers with different levels of experience.

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.

STEPS TO RESOLUTION

Developing solutions in an attempt to fill in the gaps within your data team is not an efficient use of time. Additionally, the cost and effort involved in creating tools that probably already exist is simply a waste of resources. **Your data team may find greater success in taking advantage of existing technology and customizing it to serve the business needs of established organizations.**

SOLUTION SUMMARY

- Do not develop for the sake of filling in the missing pieces of your own data team.
- Be aware of the technological landscape and plan accordingly - leverage existing technologies (such as open source software) to create solutions that offer competitive advantages rather than reinventing the wheel.

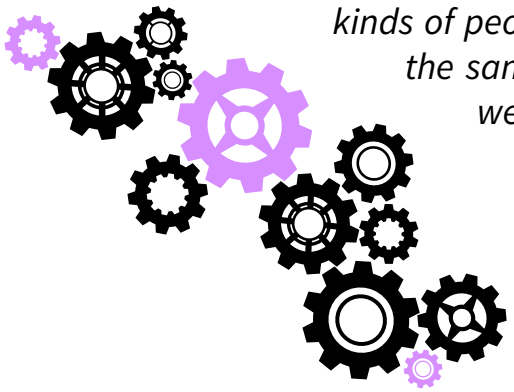
Tip: Deciding to Build

Standing at the growth crossroads may be a daunting situation for you and your data team, but an awareness of the business landscape in terms of leveraging technology will be beneficial.

Some other key points to consider:

- ☐ Does your initiative serve a generic need from an ecosystem or platform (e.g., Hadoop, Spark)? If so, if 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.

Missing any one of the key ingredients for a data team (like access to data or access to the production environment) has the capacity to completely compromise the integrity of your data science projects. Without thinking about each piece of the data project and all the parts working as expected, your models may be invalid before they're even created or your solution may not even be deployable on your customer's platform.

Another key component that may be missing from your data team is your capability 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.

ABOUT DATAIKU

Dataiku is the advanced analytics leader and preferred software solution in helping organizations succeed in the world's rapidly evolving data-driven business ecosystem. Guided by the belief that true innovation comes from the effective combination of diversity of cultures, of mindsets, and of technologies, Dataiku's purpose is to enable all enterprises to imagine and deliver the data innovations of tomorrow.

ABOUT DATAIKU DSS
(DATA SCIENCE STUDIO)

Dataiku DSS is a collaborative data science software platform that enables teams to explore, prototype, build, and deliver their own data products more efficiently. It is an open platform designed to accommodate rapidly evolving programming languages, big data storage and management technologies and machine learning techniques, and is conceived to accommodate the needs and preferences of both beginning analysts and expert data scientists. It also uniquely support:

Collaboration

Collaboration features make it easy to work as a team on ambitious data projects, to share knowledge amongst team members and to onboard new users much faster. You can add documentation, information or comments on all DSS objects.

Reproducibility

Every action in the system is versioned and logged through an integrated Git repository. Follow each action from the timeline in the interface, with easy rollback to previous versions.

Production Deployment

DSS lets you package a whole workflow as a single deployable and reproducible package. Automate your deployments as part of a larger production strategy. Run all your data scenarios using our REST API.

Governance and Security

DSS helps you create clearly defined projects and make sure your data is organized. And with fine grained access rights, your data is available only to the right persons.

Try Dataiku DSS for free by visiting www.dataiku.com/try