informatica

The Big Big Data Workbook

A practical guide to get your first big data project off the ground.

What's inside

Hint: Click to jump to section

Going big

Few technology trends have earned the kind of hype big data has.

But then again, few technology trends have offered enterprises so much potential for transformation. Ever since software began to envelope whole business processes at the turn of the century, it's been clear: Data changes the way we work.

Of course, with great hype comes great disillusionment. And in the case of big data, we've certainly seen both good advice and misinformation dished out in equal measure. Unfortunately, because this brave new world of infinite data is still so new, all the noise has left a lot of people confused.

This workbook aims to dispel that confusion.

It's about bulletproofing your strategy and executing pragmatically. Whether you're starting a localized, tactical initiative or planning a more foundational endeavor for the whole enterprise, this book will serve as a practical guide on your journey.

Let's dive in.

Part A: Getting ready

We've divided this book into three parts. In this first part, we're trying to sharpen your vision so you can pick the right project.

Back to what's inside <

What you need to know

Before digging into the specifics of your own project, here are some lessons most big data practitioners wish they'd known before they started their projects.

What you need to know

Why most organizations implement big data projects

When companies decide they're going to tackle big data, it's usually for one of these reasons.

They're trying to conduct better analysis and they realize it will take a significant increase in the amount of data being analyzed to get there. Usually, a business unit (like marketing) starts these initiatives.

They realize they can wrap their products in a service layer by providing (often real-time) analytics that can help their customers use their products more efficiently and effectively.

They want to do things faster, better, and cheaper by using big data to inform all the decisions being made about a specific business unit or process.

They realize big data is critical to every business unit in the organization and they attempt to establish the foundations for a data-centric view of their entire world.

They know they have to start getting good at big data before it's too late, but they haven't really figured out what to do with it just yet. The aim is to learn and experiment with it.

All of these are great reasons to get interested in big data. But if you're going to ensure your projects stand the test of time (and multi-departmental scrutiny), then you're going to have to be very clear about which of these reasons represents your interest best.

What you need to know

Why big data projects fail

A survey[1] found that 55% of all big data projects don't get completed, and many others fall short of their objectives. While this strike rate isn't atypical at such an early stage of a technology trend, it would be foolish not to learn the lessons these projects can teach.

Let's look at the four main reasons for big data project failure.

1

Vague goals

The reason for failure most commonly cited in the survey was 'inaccurate scope' of project. Too many companies aim for ambitious – but altogether much too ambiguous – projects with no clear objectives and then fail when they have to make hard choices about what's important and what's not.

Pursuing big data for the sake of having a big data project is a recipe for disaster. The complexities of these projects require a firm commitment to obtaining a certain outcome. Without certainty of goals, that isn't possible.

[1] www.informationweek.com/software/information-management/vague-goals-seed-big-data-failures/d/d-id/1108384

Why big data projects fail

2

Mismanaged expectations

All the hype around big data makes for some very dangerous assumptions about what your project can deliver. As tempting as it might be to make bold promises in short time frames, it's important that you maintain a realistic view of what can be expected from the project, how long it will take you, and the amount of effort it will take to get you there.

When the expectation of impact and insight is too high, you find yourself navigating terabytes of unknown unknowns looking for gold. When the expectation of delivery is unrealistic, you find yourself aiming for deadlines and budgets that aren't at all fair.

3

Project overruns and delays

Considering how new the discipline is to the enterprise, it shouldn't come as a surprise that most big data projects end up costing too much or taking too long. This is typically a mixture of mismanaged expectations and a misunderstanding of how to build a scalable architecture.

When rare, expensive Hadoop-Java developers are hired and then tasked with hand-coding mammoth implementations, companies soon realize it's impossible to move out of the sandbox environment without errors. As a result, big data projects end up languishing as a science experiment and never make it out of the lab.

What you need to know

Why big data projects fail

4

An inability to scale

It's hard enough finding five great Hadoop Java developers. But when projects grow and need to scale to 30 Java developers in a single year, things can hit a wall. The worst part isn't the opportunity cost of unused Hadoop clusters – it's the opportunity cost of lost momentum and time.

Too often, companies aim for short-term expediency over long-term sustainability. While we would be remiss to suggest you can always avoid making that trade-off, we can't stress strongly enough the importance of the long view. For your data to be appropriately secured and managed, you need to keep an eye on the long-term implications of your project.

The four reasons for big data failure are worrying and far too common. So let's now take a look at how you can avoid them and build a lasting implementation.

How to make your big data project work

If most big data projects fail for lack of clarity and inability to demonstrate the practicality of the initiative, then you should take it upon yourself to bring focus and proof to your project. Here are three useful tips to make sure your project hits the ground running and stays on its feet.

1

Set clear objectives and manage expectations

If you're unsure what your project should aim for, consider the objectives you've set for your existing data infrastructure.

If your organization already needs data for certain business processes (like fraud detection or market analysis), consider how big data might make those processes better or more valuable. Rather than tackling a whole new problem, you should aim to improve an existing process or project.

Without a clear focus and demonstrable value to business users, your project is doomed.

2

Define the metrics that prove your project's value

Clearly defined metrics that tie in to your objectives can save you a great deal of trouble. By setting yourself realistic targets that can be measured, everyone around you will be able to see the progress you're making.

More important, they'll know what you're aiming for in the long term. Ask yourself how you can measure the impact of your project in the context of your goals.

This is crucial because there will be short-term compromises that your business users will need help rationalizing and measurable goals help you prove you're delivering greater value than they realize.

How to make your big data project work

3

Be strategic about tools and hand coding

Avoid the temptation to manually hand-code everything directly in Hadoop. Remember, the goal here is not to build a working implementation with your bare hands from scratch – the goal is to deliver the value of big data to your organization.

Instead of attempting to hand-code every integration, clean every dataset, and hand-code all the analytics, you should look to tools and automation to help you accelerate through these processes.

More important, don't fall into the trap of wasting rare, expensive Java development talent on aspects that can't be scaled or transferred to other employees. Your role is to make strategic decisions about the deployment of scarce resources in a way that achieves your goals.

Adopt tools that can increase the productivity of your development team by leveraging the skills and knowledge of your existing ETL, data quality, and business intelligence experts, while freeing your Java superstars to work on specific logic for which tools aren't available.

Additionally, because technologies like Hadoop are evolving everyday, it's worth considering an abstraction layer that can protect you from the constantly changing specifications of the underlying technologies.

Above all, remember the skills you need are scarce – but tools are always available.

Choosing the right project

In light of the challenges you'll be facing, let's now take a look at how you should go about picking the right project for your organization.

What the right project looks like

If your organization is hungry for change and has already accepted that it will take a comprehensive data governance framework to improve the way they work, you can probably afford to skip this section.

On the other hand, if you're considering a localized, tactical project that can later be adapted for the wider enterprise, keep reading.

The right project has these four components.

1

Demonstrable value

The right project is one where the value is shared evenly between IT and the business unit you're trying to help. That means delivering clear value to a department, business unit, or group in a way that they can see.

2

Sponsorship

The executives who buy in to your vision are essential to the success of your project. Big data projects need champions and sponsors in high places who are willing to defend the work you're doing.

So if you know you can build superb analytics for logistics, but the only executive buy in comes from the CMO, you should think again. If marketing is your champion, build to serve marketing's analytics requirements. You can't force change on anyone. Follow the influence and make the most value you can from it.

What the right project looks like

3

A bowling pin effect

The strategic importance of your first tactical project is vital. Not only do you want to prove beyond a shadow of a doubt that big data can help the business unit you're serving, you also want to make sure that its value can then be easily communicated to the wider enterprise.

So when picking your first project, choose strategically.

Once you've demonstrated the value of big data to your marketing department, for instance, it will be easier to get buy in from the logistics teams who might otherwise have been reticent.

4

Transferrable skills

As we said in the last point, you want the value of your first project to help convince other departments in the enterprise. To that end, you need to ensure you can learn the right skills, capabilities, and lessons from your first project. More pointedly, you need to ensure you document all of them so that you can transfer them to your next project. Remember, if you're aiming for success, you're aiming for future projects.

So be prepared to scale so you can handle more projects in the future. This isn't just a question of scaling your cluster. It's about scaling your skills and operations. You'll either need to find more Java/ Hadoop superstars or find ways to get more out of the resources you already have.

Consider the impact

When you're choosing what your next project will be, you also have to consider how it's going to impact your organization. There are three broad aspects that should play a role in deciding whether you're pursuing the right big data project.

1

Cost and disruption

At the most basic level, the cost of your project is based on the time and money it will take to get it on its feet. In reality, you should also consider the potential disruption it will cause.

Sometimes the disruption is procedural – when business units are used to owning their data and they aren't comfortable giving up control of it to a centralized data governance framework.

Other times it's technological and skills-related – when you have to integrate new technologies into your existing infrastructure and reorganize or upgrade skills to do so.

Whatever the case, you should anticipate, acknowledge, and make sure you either minimize the disruption – or communicate why it's valuable.

Consider the impact

2

Timing of benefits and impact

When considering different starting projects, you'll naturally lean toward those that can deliver maximum business impact and improvement. But it's also important to consider the nature of the business impact. Will it deliver the bulk of the value in the short term or the long term?

More important, when will business users feel the business impact? For instance, you could introduce master data management to your data warehouse and drastically improve the efficiency of your business intelligence. But that value would only be felt once your business analysts realize they won't be cleaning financial data again.

3

Resources and restrictions

In light of your analysis of the previous two factors, consider the resources at your disposal. We'll be diving into this in greater detail later, but for now just bear in mind that you naturally want your project to deliver more bang than your invested buck.

Achieving that goal works both ways. On the one hand, you want to aim for maximum business impact. But you also have to be strategic about how you spend your budget. While it might be tempting to build a team of data scientists to match Google's, can you really afford to? Making smart choices between tools and headcount will be critical to the success of your project.

Tactical big data projects: Some examples

There are a wide variety of applications for big data. As exciting as this makes it, it also makes it a little daunting for people who aren't quite sure which project to start with. Here's a list of tactical big data projects we've seen our customers undertake.

If you're still not sure which project your organization should be starting with, consider the following examples for a better idea of what big data might offer your company.

Finance

- Risk and portfolio analysis
- Investment recommendations

Retail

- Proactive customer engagement
- Location-based services

Media

- In-game behavior tracking
- Cross- and up-sell options

Manufacturing

- Connected vehicle programs
- Predictive maintenance

Healthcare

- Patient outcome predictions
- Total cost of care
- Drug discovery

Public Sector

- Health insurance
- Exchanges
- Tax optimization
- Fraud detection

Tactical big data projects: Some examples

What some of our customers aimed for

Take a look at how specifically some of our customers describe their efforts. This is the kind of focus you should strive for.

- A large technology company in Silicon Valley aims to save over $10 million in growing data warehouse costs by using a combination of Hadoop and traditional data warehouse technology to decrease the growth in overall cost per terabyte.

- A large manufacturer in transportation means to reduce the fuel consumption rate on its vehicles by one percent over the next 10 years. It also aims to reduce its toxic carbon omissions by extending maintenance periods by 10 percent and improving its mileage by one percent.

- A manufacturer involved in locomotives intends to unlock an additional mile per hour on daily routes so its customers can save as much as $200 million every year.

- A global payment services company seeks to increase its digital business by 30 percent through increased customer personalization, all part of a big data strategy called 'retail omnichannel optimization.'

Those are some big wins for any big data team.

The foundational big data journey

If you're ready to build the foundations for an enterprise-wide approach to big data, the following three steps are going to be essential to your journey.

Indeed, even if you're targeting a handful of tactical big data projects, you should be considering these three steps. Each is crucial to the foundational integrity of your data-centric organization. In fact, for the most cost benefit, you should aim to follow these steps in order.

1

Data warehouse optimization

This entails choosing to store and process data on the most cost-effective platform. It often starts by moving raw or infrequently used data and ETL workloads off of expensive data warehouse hardware.

The aim is to avoid costly upgrades of your data warehouse and start using cheaper hardware and distributed computing frameworks like Hadoop so that you're prepared to handle the volume, variety, and velocity of big data.

The foundational big data journey

2

A managed data lake

A managed data lake is a single place to manage the supply and demand of all your data.
The operative word here is 'manage.' The aim is to convert your multi-structured mess into fit-for-purpose, reliable and secure information.

That means creating a data lake that refines, governs, and masters your data. It takes a great deal of foresight to get there though, since you'll need to incorporate rigorous, strategic data governance policies and processes. But without them in place, your lake will run the risk of basically turning into a data swamp.

3

Real-time operational intelligence

Here you're creating the technologies (analytics, data-hungry applications, engagement interfaces) your people need to access, analyze, and deliver all that data. The applications you build here have to be easy-to-use and deliver the information your users need.

This could be an interface for your customer service representatives that monitors customer behavior across multiple channels and identifies the customers most likely to churn in the next two weeks.

A three-step journey

As we've already said, for the most cost-benefit, we'd recommend following these steps in this order.

Data warehouse optimization

Lower your infrastructure costs and bolster the enterprise's architecture.

Managed data lake

Create a single place to manage the supply and demand of data.

Real-time analytics

Deliver cutting-edge applications that provide the information your people need.

The foundational big data journey

How our customers define their foundational goals

Even foundational projects must be specific about what they're trying to build. While the specificity here may not pertain to dollars and hours saved, it does apply to the boundaries of what exactly is being built. Consider the following examples of our customers' big data infrastructure projects.

- A global organization that conducts hundreds of millions of financial transactions in hundreds of countries has built an enterprise-wide data hub. The aim is to conduct big data analysis and identify key macro trends and patterns in customer interaction.

- A large technology company has built an enterprise-wide analytics cloud to drive faster time to market for data-driven products by including new datasets into analytics being used across business units.

- A global financial advisory organization has built a logical data warehouse infrastructure to ensure it can make consistent information available across all standard platforms (including Hadoop, operational databases, and traditional data warehouses) being used by the organization.

In short: Big plays make big impacts – but they demand the right foundations.

Part B: Your strategy

Now we'll get practical and look at the specific requirements for your next (or first) big data project.

Back to what's inside <

Defining your goals

Get your pencil out. As we've already covered, the number one cause of big data projects failing is a lack of clear goals. Now let's make sure the project you have in mind isn't going to drown in ambiguity.

The business goals

We'll start with the business because these goals have to take precedence over IT's if your project is to be fully appreciated.

Be as specific as you can be in laying down the goals you want your project to achieve for the business. And remember to aim for goals where the impact is measurable.

For instance, in the example of the customer service interface that predicts customer churn, the goals for that project shouldn't be listed as something vague like 'improve customer experience.'

The more crystal-clear your goals, the closer you'll get to achieving them. Five laser-focused goals are more valuable than one vague one.

The business goals

List, in order of importance, the goal(s) of your big data project as they pertain to the business and business users. (Feel free to enter fewer or more goals.)
e.g., Reduce customer churn

Write down a minimum and maximum amount of time for each goal to be achieved.
e.g., Three to six months

Now, for each goal, write down a measure of success that can be used to determine whether the goal was achieved. Ideally, these should be available metrics or calculations thereof.
e.g., Reduce average monthly churn rate by X%

How long should your big data project take?

Your big data project should take as long as is needed to deliver its full value. In our experience, the scope of the project dictates the time horizon.

We've worked with customers who've delivered tactical projects in less than three months. And we've worked with customers who've spent three years delivering foundational programs.

For lengthier projects, bear in mind that you should be aiming to demonstrate the value of your project every six months. If you adopt an agile approach to the project, then it helps to present the different phases and milestones as smaller projects.

One thing is clear – you shouldn't be guessing how long it's going to take. Estimate the time to deliver based on your experience and the experience of others who have undertaken similar projects before. If you're unsure who to call for guidance, you can always get in touch with us.

IT's goals

Now let's look at IT's goals as they pertain to your project.

(Remember, if your project is about helping IT work better or faster, you're going to have a hard time selling that to business users. As such, IT's goals should be communicated in conjunction with the goals your business users are already excited about.)

List, in order of importance, the goal(s) of your big data project as they pertain to IT. (Feel free to enter fewer or more goals.)

e.g., Establish processes for real-time collection, cleansing, mastering, and storage of aggregate customer data, credit card usage data, social graph data, and churn indicators

Stop, collaborate, and listen

We've written this workbook so you can start your big data project, whether you work on the business side or in IT. In either case, don't leave your goals to guess work. If you need to get specific guidance on what to aim for, grab a partner with the expertise you need and start collaborating now.

If your project is going to succeed, you can't do without strategic collaboration.

IT's goals

Write down a minimum and maximum amount of time for each goal to be achieved.
e.g., Two to four months

Now, for each goal, write down a measure of success that can be used to determine whether the goal was achieved. Ideally, these should be available metrics or calculations thereof.
e.g., Accurate churn prediction rate of X%

Defining your data needs

Now that we've outlined the specific goals of your big data efforts, let's get right to the heart of your project – the data itself. Whatever your project, you're going to have to think strategically about what information you need, what datasets speak to that need, how you're going to get that data, and how you're going to use it.

What data do you need?

First, let's look at the most basic purpose of your big data project; the information you're trying to provide to your organization. Answer the following questions as specifically as you can.

In order to achieve the business goals outlined previously, what do my business users say they need to know to make an informed decision?
e.g., What most valued customers are likely to churn and what behaviors correlate to churn

In order to deliver that knowledge, what data can be used?
e.g., Customer purchase history, reviews data, rate of purchases, abandonment rate, bounce rate, customer quality of service

What data do you need?

Which source systems contain these datasets?
e.g., Customer service records, product performance metrics, customer activity database, customer master data management

Outside of the data already noted, is there other information that might lend context or additional value to your analyses?
e.g., Customer service survey data, competitor analyses, weather data, social data

What data do you need?

What datasets that I currently don't have access to might contain additional contextual data?
e.g., Third-party social data, third-party market data, weather data

The hunt for dark data

When considering datasets you don't have access to, don't confine yourself to data outside your organization. Gartner has found that most enterprises only use 15% of the data inside the organization.[2] Appfluent, a company that does statistical analysis on data warehouse utilization, found that somewhere between 30% and 70% of data in a data warehouse is dormant.

The rest is hidden away in hard-to-reach, expensive-to-use, or difficult-to-find silos, legacy archives, and data stores. Which wouldn't be a problem if it weren't for the fact that you're already paying to store all this data.

When looking for the data you need, it's worth starting with a look at the data your organization already has.

[2] Gartner website: www.gartner.com/technology/topics/big-data.jsp

Five key data considerations

Once you've outlined the data you're going to be looking for, you'll have a clearer view of your specific big data challenges. In particular, there are five key elements you should consider before you go any further, as they will dictate what needs to be done for each dataset, as well as for your big data dataset.

1

Prepare for volume

You're going to have to get ready to deal with the 'bigness' of the data you need. Across dimensions, classify your data based on its value (say, customer transactions), usage (frequency of access), size (gigabytes, terabytes), complexity (machine data, relational data, video…), and who's allowed to access it (only your data scientists or any casual business user).

A thorough, organized inventory of your data will help you determine how to manage all of it. Assess your current storage and processing capacity and look for the most cost-effective and efficient ways to make it scalable.

Five key data considerations

2

Account for variety

The most challenging aspect of big data is the multitude of formats and structures you're going to have to reconcile in your analyses. You will have to integrate a number of sources if you want to include new data types and structures (social, sensors, video) with the sources you're already used to (relational, legacy mainframes).

Attempting to manually code every single integration is so cumbersome it could cost you all the time and resources you have. Make the most of available data integration and data quality tools to speed your process for more valuable tasks.

3

Handle velocity

The combination of real-time streaming data and your historical data usually increases the predictive power of analytics. So, some of the data you want may only be valuable if it's constantly pouring into your systems.

Indeed, most real-time analyses need to be based on streaming data – often from different sources, in different formats. Prepare your project with streaming analytic technology and a logical infrastructure to manage all the data.

Five key data considerations

4

Ensure veracity

No matter how important your analyses, it won't be worth anything if people can't reasonably expect to trust the data that's gone into it. The more data you analyze, the more important it is that you maintain a high level of data quality.

For your data to be fit for purpose, you'll need to know the purpose it's being used for. If a data scientist is looking for patterns in aggregated customer data, the preparation required will be minimal. On the other hand, financial reporting and supply chain data will need to be highly curated, cleansed, and certified for accuracy and compliance.

Create categories based on the amount of preparation needed ranging from raw data to a highly curated, mastered data store of cleansed, reliable, authoritative data.

5

Think compliance

The different datasets you deal with are going to come with different security stipulations and requirements. For each dataset, you need to consider what it will take to anonymize the data based on security policies.

Masses of your data will be proliferating across the enterprise in hundreds of data stores. Understand where your sensitive data resides and ensure you secure it at the source through encryption and then control who has access to it.

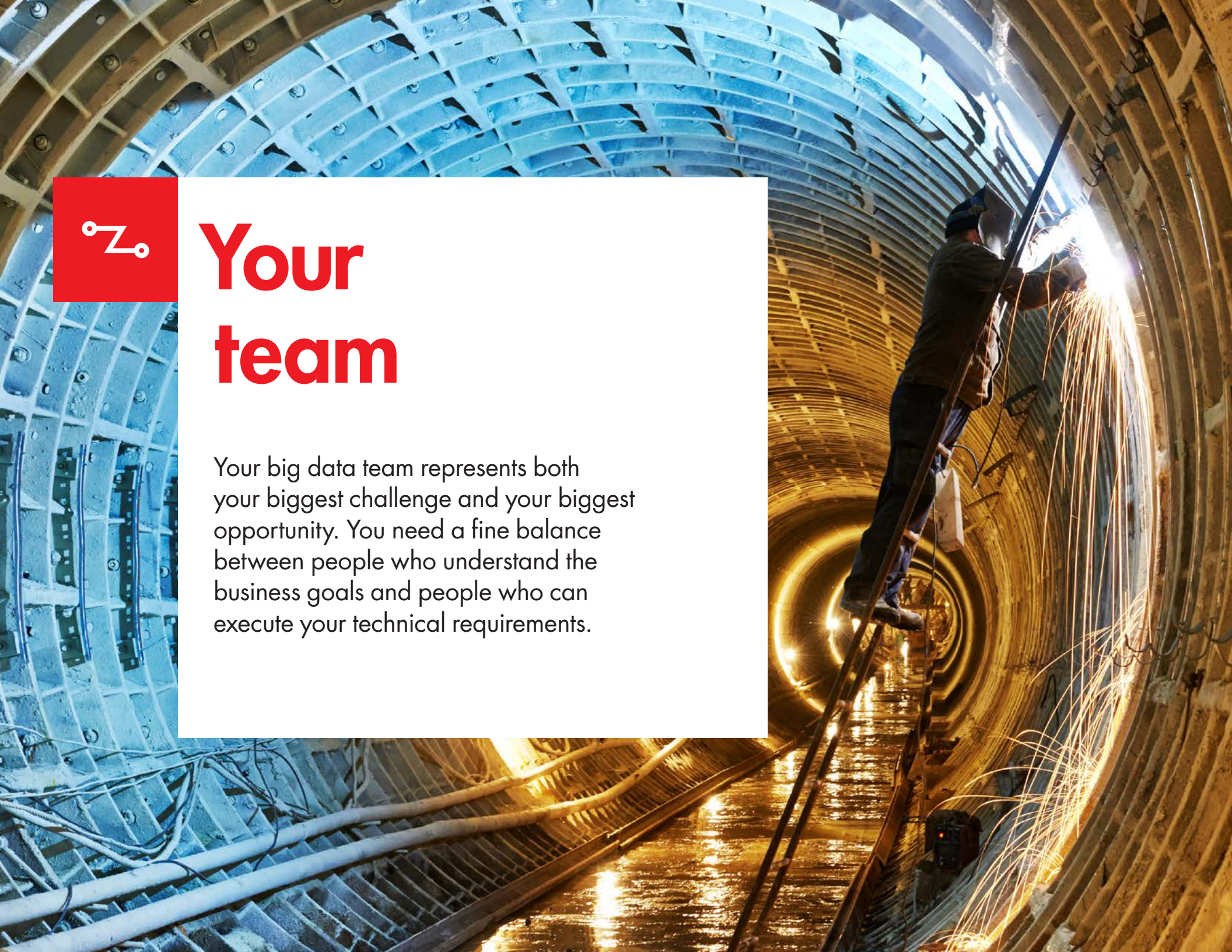
Even beyond secure, smart archiving of sensitive data, mask your data with predefined rules any time it migrates or enters your development and test environments.

Apply these five considerations to every dataset you deal with, and you'll be prepared for your big data challenge in a more realistic way.

Part C: Your lean big data supply chain

Traditional business intelligence and data warehouse methods do not scale to meet the needs of big data initiatives. Now we'll look at how you can scale your team, your processes, and your infrastructure.

Back to what's inside <



Your team

Your big data team represents both your biggest challenge and your biggest opportunity. You need a fine balance between people who understand the business goals and people who can execute your technical requirements.

Five key team-building lessons

Most organizations underestimate the level of skill needed to successfully apply new technology like Hadoop.

Distributed data frameworks are just plain hard to manage. From the Java skills required to develop on Hadoop to the new data science skills for which you'll have to hire, you're going to have to bring in a lot of new skills for your project to really fly.[3]

When you start building your team, be sure to incorporate the following lessons into your hiring strategy.

[3] "Hadoop, Python, and NoSQL lead the pack for big data jobs," InfoWorld, May 5, 2014: www.infoworld.com/t/it-jobs/hadoop-python-and-nosql-lead-the-pack-big-data-jobs-241884)

Five key team-building lessons

1

Use the skills you hired your people for

One of the biggest mistakes companies make when they hire data scientists and quantitative analysts is to make them do the dirty work. When your most skilled resources spend all their time hand-coding data integrations and cleaning data, you don't just leave them frustrated – you fail to take advantage of the skills that were so difficult to find in the first place.

Concentrate rare skills on the tasks that really need those skills. You don't want your best people to leave and you certainly don't want them wasting time doing work you could comfortably do with tools instead.

2

Think strategically about the composition of your team

If things work out, your project will grow in both scope and resources. Think strategically now and save yourself the harsh realization that you can't scale certain processes quickly enough because there are a finite number of people with the skills you need – even in Silicon Valley.

If your project grows in scope, what skills can you reasonably expect to find in time to match your needs? For instance, data scientists are infinitely harder to find, train, and hire than developers.[4]

The balance of your team is crucial. You're looking for the right mix of hard earned data management experience and an enthusiasm to learn new tools. Additionally, you need to strike a balance between people with technical skills and people with the domain expertise needed to build the right models.

[4] "Big Data's High-Priests of Algorithms," Wall Street Journal, August 8, 2014: http://online.wsj.com/articles/academic-researchers-find-lucrative-work-as-big-data-scientists-1407543088

Five key team-building lessons

3

Align the goals of your project early and then communicate them

One of the most common errors companies make when hiring new people is to forget to communicate the true goals of the project. From the first interview through to the actual job itself, it needs to be made crystal clear what you are trying to deliver to business users. Leverage your executive sponsorship to evangelize the mission and share success stories as well as issues.

Without a firm grasp of the business value of your project, your new hires run the risk of thinking they only have to think about IT's goals for the project.

4

When your team grows, the need to manage them grows, too

Unlike new technology that can be deployed, implemented, and then integrated in an objective fashion, new people have to get used to where they're working, what they're doing, and why they're doing it. Whether it's you or someone else, someone needs to embrace the management challenge that a new team calls for.

Elements like culture and cohesiveness cannot be underestimated. Think long and hard about how you integrate new hires into your processes. You may not be able to train them for skills, but you can certainly help them be better team members.

Five key team-building lessons

5

Your team cannot afford to stand still

Big data technologies are emerging everyday. And the ones that already exist are evolving rapidly. It's a hugely exciting time for the companies brave enough to adopt best practices early. But it also represents the defining challenge of having a head start on your competitors.

Your people need to evolve their skills as rapidly as the world around them is changing. The good news is nothing motivates the best people more than the challenge of staying ahead of the curve. The challenge is in delivering the training and discussion they need to keep growing their abilities and yours.

The importance of being strategic

An important choice you'll be making over and over again is whether to build your capabilities using automated tools or manual integrations.

Hand-coding offers you complete, precise control over what you're building. Often this is invaluable and necessary if, for instance, you're writing a complex script to extract metadata in a way that isn't possible yet.

Tooling, on the other hand, offers you greater agility and the ability to sustainably repeat the same process. For tasks like data integration and data quality, this is crucial because it means you aren't forcing your super-intelligent analysts and scientists to do the dirty work.

Be realistic about your resources. If you can't build a team as big and brilliant as Google's, don't waste your scarce resources trying.

Setting up data governance

If (and hopefully when) you're setting up a more foundational big data endeavor, you're going to have to put in place the procedural framework for data governance. In fact, even if your big data project is aimed at delivering value to a single department, you might consider building a miniature data governance board so you can learn how to cope with the unique challenges such a body presents.

Essentially, your data governance board is the formal body of executives meant to oversee the enterprise's approach to data. But it also includes the need for data stewards – functional or department-specific people who are tasked with managing the data coming from a specific business unit.

(In fact, some of our clients assign data stewardship roles based on the data domain. That means one person is in charge of product data, while another is in charge of customer data, and so on.)

Setting up data governance

You should aim to create processes that ensure your data governance framework helps more than it hurts. Actively work to ensure it doesn't turn into a bureaucratic burden by ensuring everyone is committed to achieving the same goals in the same time frames.

Your data governance framework should aim for the following five characteristics.

1

Cross-functional

A data governance board comprising different people with similar roles will be ineffective. The aim is to create a body capable of representing the unique views and needs of each business unit your big data project is meant to serve.

2

Communicative

Without good communication across functions, departments, and domains, your project is likely to drown in bureaucracy and misunderstanding. This happens far too often. Ensure all concerns are either abated or appropriately addressed.

Setting up data governance

3

Efficient

Your cross-functional process shouldn't feel like a barrier. It will take great agility for your big data project to succeed. So build in automation and exception reporting rules wherever possible and adopt collaboration tools to keep the lines of communication open and expedient.

4

Committed

Be sure to communicate the primary goals of your project effectively and make sure everyone involved in your data governance framework is dedicated to achieving those goals. Common goals should guide your governance thinking and decision-making.

5

Centralized

The biggest challenge with a data governance framework comes when you have to prioritize the goals of one business unit over the others being represented on the board. Ensure your decisions are for the long-term benefit of the whole board even if it means short-term benefits being felt by one business unit.

The skills you need and the skills you have

Time to get your pencil out again. Now that you can see the various subjective pitfalls and opportunities your new team will present, let's figure out what this team will actually look like.

The following page lists big data roles based on the jobs for which we've seen our customers hire. Based on the personnel currently available to you and the amount of time you expect your project to take (as entered in the section beginning on page 24), list how many people you need to hire.

The role	Can anyone perform this role already?	I need to hire for this role	Based on the amount of time I have, I need to hire X people
Data scientist	✓ or ×	✓ or ×	
Domain expert			
Business analyst			
Data analyst			
Data engineer			
Database administrator			
Enterprise architect			
Business solution architect			
Data architect			
Data steward			
ETL (data integration) developer			
Application developer			
Dashboard developer			
Statistical modeler			
Other			
Other			
Other			
Other			
Other			

The need for integrated thinking

When you go out looking for new team members, don't limit yourself to people with the right qualifications. Make no mistake – finding people with the right qualifications is a challenge in and of itself. But you also need to look for people with a willingness to synthesize business goals and technical capabilities.

Over and over, we hear from customers about how important it is that the people who join their big data projects are able to understand business realities and execute complex data science. This type of integrated thinking is huge and hard to find. It's worth training for and worth rewarding.

Your tools

As we've discussed a number of times already, the tools you use have a strategic role to play in the execution of your big data project. In this section, we're going to look at the tools you have and those you need.

Understanding big data tools

In our experience, the following tools are essential to the architecture needed for big data projects (we discuss the architecture in greater depth later). Of course, your goals and your resources should determine the technology mix needed for your specific project.

Go through this list of tools and put an ☒ against the ones most important – and most strategically relevant – to your specific project.

Data Ingestion

The process of consuming the data you need appropriately, efficiently, and methodically.

- ○ **Batch load**
  Can you access all types of data you need and efficiently scale the performance of batch loading into your data stores?

- ○ **Change data capture**
  Can you capture the changes made to data in your source systems without impacting the source systems?

- ○ **Data streaming**
  Can you reliably collect real-time data and stream it into your data stores?

- ○ **Archiving**
  Can you archive and compress data that isn't used frequently while ensuring easy access to archived data as needed?

Understanding big data tools

Go through this list of tools and put an ☒ against the ones most important – and most strategically relevant – to your specific project.

Data Management

All the policies, processes, and practices needed to properly manage the efficacy, accuracy, reliability, and availability of your data.

○ **Data integration**
Can you prepare and consolidate various structures and sources into one cohesive dataset for analysis?

○ **Data quality**
Can you reliably cleanse your data, de-dupe, and remove errors?

○ **Data security**
Can you discover and secure your data across all data stores by assigning rules about usage, access, and permissions?

○ **Virtual data machine**
Can you create an abstraction layer for your data that loosely couples data processing from the underlying deployment environment?

○ **Master data management**
Can you store a consolidated, complete, authoritative version of the truth for various data domains?

○ **Distributed data framework**
Can you use technology like Hadoop to cost-effectively scale your storage and processing needs?

○ **Data warehouse**
Do you have data warehouse technology that can withstand the performance, usage, and scalability requirements for big data analyses and integrations with Hadoop infrastructures?

Understanding big data tools

Go through this list of tools and put an ☒ against the ones most important – and most strategically relevant – to your specific project.

Data Delivery

The process of sending the data you have to the systems and applications that need it.

- ○ **Batch load**
  Can you efficiently scale batch loading of data between source, analytic, and operational backend systems?

- ○ **Real-time streaming**
  Can you deliver stream data in real-time to the applications, analytics and back-end systems that need it?

- ○ **Data integration hub**
  Can you make your data available using an approach like the publish-and-subscribe model to avoid the proliferation of point-to-point integrations?

- ○ **Data virtualization**
  Can you deliver data from your systems without overloading them?

- ○ **Event-based processing**
  Can you detect, analyze, and respond to threats, opportunities, and other business critical events in real-time?

Understanding big data tools

Go through this list of tools and put an ☒ against the ones most important – and most strategically relevant – to your specific project.

Analytics

The tools and processes that turn raw data into insight, patterns, predictions, and calculations about the domain you're analyzing.

○ **Visualization**
Can you present your data and findings in ways that are easy to digest and understand?

○ **Advanced analytics**
Can you apply cutting-edge analytic algorithms to your datasets to conduct complex calculations?

○ **Machine learning**
Can you apply sophisticated machine learning algorithms to identify patterns and make predictions at a level that you don't have the manual bandwidth to handle?

Among these tools and technologies, some tools like data integration, data quality, and master data management are so fundamental to your big data journey that they really aren't worth re-building. The amount of time and resources it takes to build those capabilities with your own hands isn't worth the precious skills and man-hours of your big data project.

Remember the goals of your project, and that they didn't include custom-building everything.

Your processes

Let's dive in to the actual processes you'll need to tackle big data. Your specific processes will be unique to your goals and requirements, but this section should give you an overview of what to expect and what you'll be learning.

The big data eight

From experience, we can say that agile methodologies are an excellent approach to big data projects. They ensure you manage expectations, learn from mistakes, and iterate your way to the best processes. That said, the approach to your project depends entirely on you and your situation.

In any event, the following eight steps will prove crucial to your big data supply chain. However you go about it, ensure you and your team establish effective processes for these steps.

1

Accessing the data

Your first challenge will be to acquire all the data you need. In some cases, this will entail capturing streaming data, and in others it will mean extracting data from a database. Set up repeatable, manageable processes to ensure this data can then be stored according to the ways you're going to use it.

2

Integrating the data

Big data's most complex challenge involves the variety of data structures and formats. For your analyses to be sustainably conducted, you'll need to set up a process for integrating and normalizing all this data. Ideally, this should take up as little manual processing as possible.

The big data eight

3

Cleansing the data

For your analyses to be reliable, you have to ensure you're cleansing your data to remove duplications, errors, inaccuracies, and incomplete data. Your process should ensure your most qualified analysts and scientists aren't spending all their time effectively 'doing the dishes.'

4

Mastering the data

One way to maintain a reliable source of clean, integrated data is to establish a process to master your data. The aim is to create a rich collection of consolidated data, organized by domain (such as products, customers, etc.), and enriched with big data insights, that can then feed all your other systems.

5

Securing the data

Here, you'll be establishing two basic processes. The first will be about defining the security rules and practices that each dataset calls for. The second will be about detecting sensitive data and masking it in a persistent or dynamic way to ensure those rules and best practices are enforced consistently.

The big data eight

6

Analyzing the data

Your process for analysis will depend on your analysts, your analytics tools, and your requirements as they pertain to your goals. The mindset of iterative discovery and continuous improvement will play a crucial role here as you want this process to get better, faster, cheaper, and more scalable with time and experience.

7

Analyzing your business needs

This step is both critical and almost always overlooked. Set up a clear process for the analysis of business needs even while you're analyzing your data. This is crucial because if you take your finger off the pulse of the business, you risk isolating your efforts and ensuring the business impact is minimized.

8

Operationalizing the insight

As we discussed early on in this workbook, the business impact of your big data project needs to be felt. Create automated pipelines for the answers you find and deliver them to the business users who need them most. For instance, data about customers most likely to churn should be made available to your customer service agents via a dashboard. Be sure to incorporate a feedback loop as well so you can see how the insight is received.

The importance of documentation

Aim to master these eight steps and your big data project will be heading in the right direction. The aim is to establish clear, repeatable, scalable, continuously improving processes. To that end, the documentation of these processes and the ensuing improvements are vital to your team.

The skills, capabilities, and lessons of your big data project must be made transferable and communicated frequently.

Your architecture

For your big data supply chain to be lean and effective, you have to ensure the architecture is solid and strategically built. In this section we'll look at what the ideal big data architecture should look like and how to go about deploying yours in a phased approach.

First steps: Your sandbox

When you start building the architecture for your big data project, the most logical starting point is to set up a sandbox development environment in which you can use test data to ensure your architecture is feasible. In doing so, make sure you consider the following lessons.

Start small

By starting with a well-defined sandbox over which you have complete control, you'll be able to iterate your way to the most successful implementation. Get up and running as soon as possible and document the lessons learned through each iteration.

Size matters

The key difference between the sandbox and your actual implementation is the fact that your production environment will be far larger. It will require automated processing to ingest, integrate, cleanse, and distribute the output. As such, it will take a far more robust structure and proven components and processes to be truly reliable and flexible in a live production environment.

First steps: Your sandbox

Mask before you test

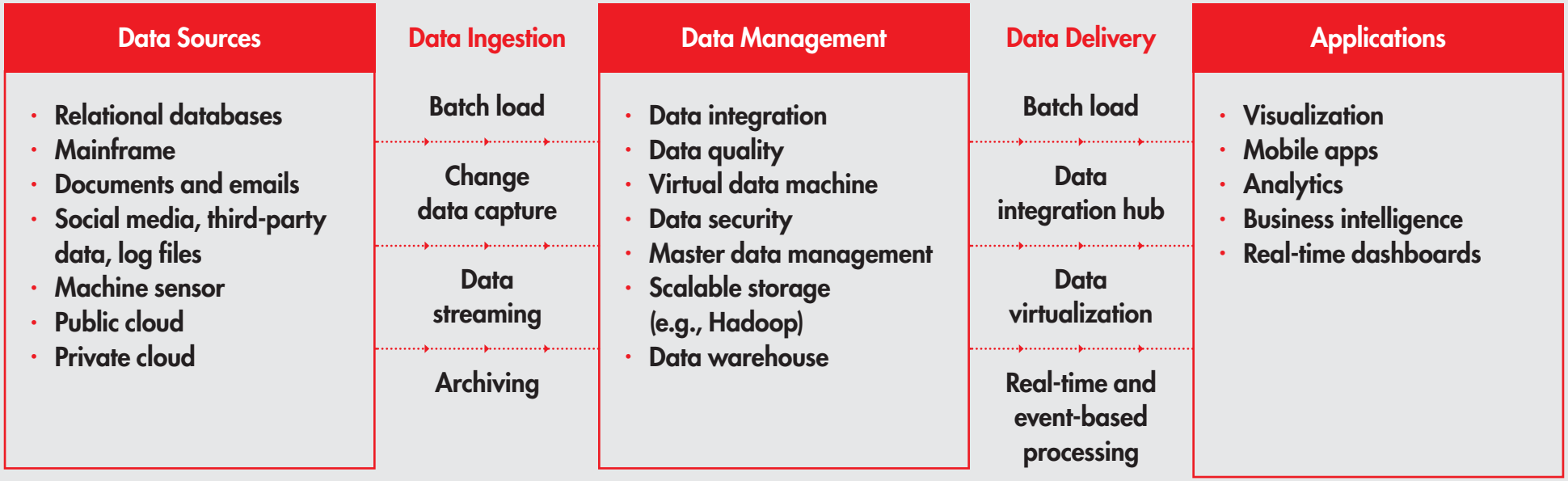
When organizations use test data, they usually use a variant of their live, production data to ensure the formats and structures represent the live environment. Unfortunately, a failure to mask this data appropriately can leave sensitive data exposed in an entirely unsafe test environment.

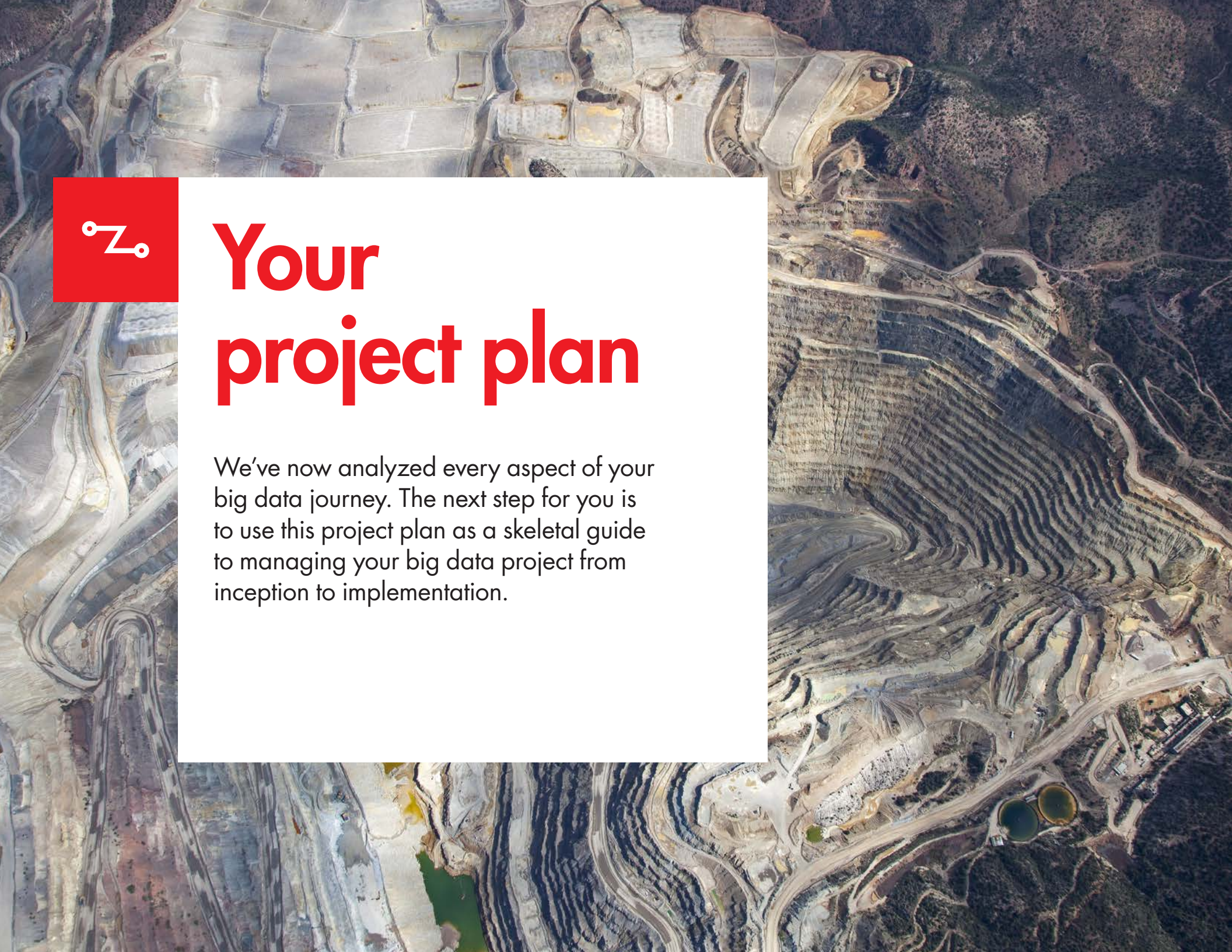
Don't get lost in translation

One of the most common sources of project overruns and costly delays in big data projects stems from the fact that hand-coded errors that were missed in the sandbox come back to haunt the team when the architecture goes live. So if you hand-code significant parts of your architecture, expect to re-factor a lot of code to meet production-level requirements and manage expectations accordingly. Alternatively, use productivity and automation tooling to avoid the need to re-factor your code and errors in the first place.

The ideal big data architecture

The following diagram represents the way we recommend building the ideal big data technology and process architecture.





Your project plan

We've now analyzed every aspect of your big data journey. The next step for you is to use this project plan as a skeletal guide to managing your big data project from inception to implementation.

Your project plan

Use this project plan template as a framework to document the details and various elements of your big data project. Then use the compiled document as a way to get the buy in you need from the rest of your organization. It will also come in handy when you approach external partners.

Stage 1: The strategy	Stage 2: The data
Identify the business and IT's goals	**Identify the information you need**
Define your measures of success	**Identify the data and sources to deliver it**

Your project plan

Stage 3: The supply chain

The people
- Assessing the skills you need
- Assessing the skills you have

The process
- Access data
- Integrate data
- Cleanse data
- Master data
- Secure data
- Analyze the data
- Analyze the business needs

The tools
- Distributed computing (e.g., Hadoop)
- Data quality
- Data integration
- Master data management
- Data masking
- Visualization
- Streaming analytics
- Analytics
- Machine learning

Stage 4: Operationalize the insight

Develop dashboards

Automate processes for data delivery

Set up a feedback process

Getting going

Use the checklists, principles, and guidelines we've outlined in this workbook to bring the potential of big data to your organization. Whatever the size of your project, by now, we're certain you're better equipped to deal with the many challenges this project is bound to spring on you.

Remember to be strategic about your resources and maintain a laser focus on developing processes and skills that are transferable, scalable, and constantly improving. If you maintain a view of the long term while undertaking this project, you'll be setting up your organization for better analyses and more informed decisions for a long, long time.

In many ways, your first big data project is going to be one you'll never forget. From the errors you're bound to make to the team you're going to build, you're about to set off on a journey of immense strategic value to your company.

By navigating and avoiding the many pitfalls we've discussed and maintaining a strong commitment to your vision for this project, you'll be able to change the way your organization works.

This is going to be big.

Next Steps

Ready to start applying the things you've learned?

If you're an Informatica developer, you can be a Hadoop developer. Our Big Data software trials, connectors, and services will get you going on the right path.

informatica

About Informatica

We help companies manage their data so they can gain measurable business value from it. And we're helping some of the biggest companies in the world navigate the most common data management mistakes to succeed at scalable, repeatable big data projects.

Let's talk.

IN18-1014-2730