BIG DATA INTO BIG BUSINESS

A PRACTICAL IMPLEMENTATION GUIDE FOR BIG DATA ANALYTICS

RAJ NAIR

CONTENTS

EXECUTIVE SUMMARY

Big data means big business. But unless you have the right foundation in place for success, just having lots of data doesn't guarantee business results. To ensure impactful outcomes, you need a proven approach to big data analytics.

An effective big data analytics plan should have three stages:

1. Exploration and Validation
2. Deployment and Governance
3. Optimization and Democratization

The use of big data analytics begins with the Exploration and Validation phase, identifying what problems it can solve in your enterprise. This process often involves investing in the construction of one or more use cases by way of a Proof of Concept (PoC).

The Deployment and Governance phase of your big data implementation is focused largely on solidifying architecture and developing standards, policies, and skills.

As the architecture of your big data environment takes shape, traditional questions including what kind of hardware will be required, what style of architecture will be used, and what business intelligence tools work best must be answered. But with data analytics being such a rapidly progressing field, new questions must also be addressed to keep up with the dynamic changes of big data itself.

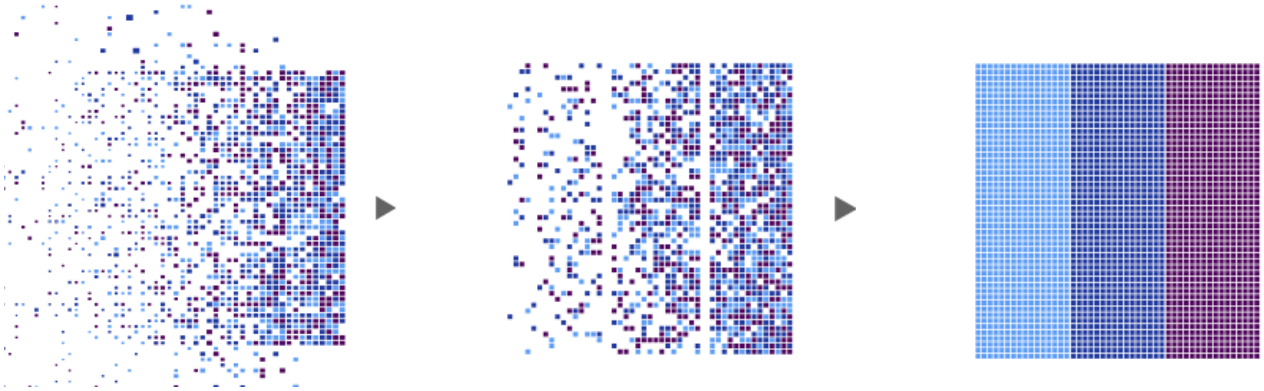
By the time you reach the last phase, Optimization and Democratization, automation reaches the forefront of your needs and must be successful in order to handle the massive operations that acquisition, distribution and scaling require.

The implementation of big data into your enterprise for analytical purposes is a marathon, not a sprint. But the long haul is well worth considering the results—your ability to win customers, sales and your own business challenges through the use of data science and predictive analytics.

INTRODUCTION

Creating value from data is done by using analytics to determine future conditions of your industry and dictate your company's actions based on those conditions, but there's a significant gap between gathering data and producing analytics that can define your business strategy going forward.

By properly leveraging big data, every organization has the ability to compete on analytics. To ensure impactful outcomes, you need a proven approach to big data analytics. The way forward is a three-stage process of big data maturity; a path each organization must follow as it collects, governs and exports its data into successful applications. These three stages are best described as:



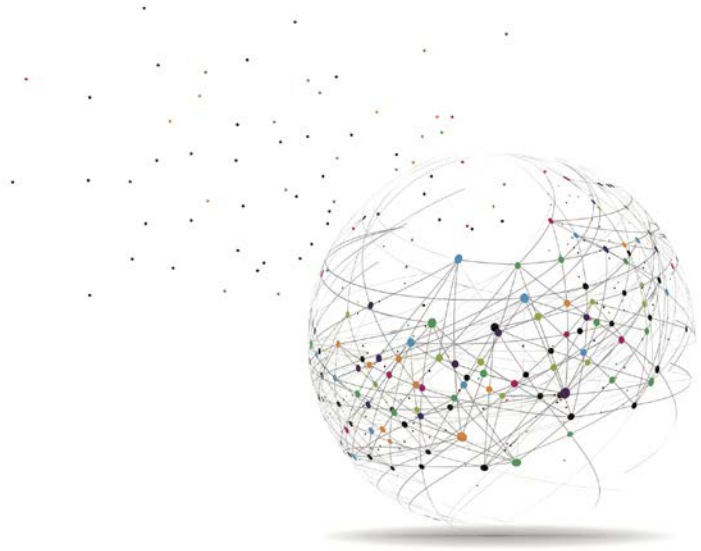
STAGE 1
EXPLORE AND VALIDATE

STAGE 2
DEPLOY AND GOVERN

STAGE 3
OPTIMIZE AND DEMOCRATIZE

This whitepaper will take a deep dive into all three stages to serve as a guide to implementing big data into your business with the end result of generating analytics and data science that will help differentiate your organization from the competition.

STAGE 1: EXPLORE AND VALIDATE



The world generates about 2.5 quintillion bytes of data per day but without structure and purpose, all that data does little for business. A data strategy must be explored early on to decide what data streams are important to your organization and how to harness it for future use.

The very first step in using big data is determining what business problem you are facing that can be solved through the use of data analytics. In other words, what value can big data add to your company.

This may not be obvious. Big data analysis can build out customer profiles for future sales. It can find weak points in your overall logistics. It can meticulously track marketing campaigns, and what level of ROI your company is generating via user experience (UX). Remember, these efforts should be led and dictated by business, not by your IT department. The architecture of a big data project will come much later. Initially the priority must be on what the business needs are.

STAGE 1: EXPLORE AND VALIDATE
BUILDING A BUSINESS CASE

Whether the challenges are obvious or not, you must build a business case before moving forward with a big data project. We then set into motion the execution of the PoC.

EXPLORING POTENTIAL BUSINESS OUTCOMES

We build potential outcomes by using a business value framework. The framework asks simple questions:

- What are the business's pain points?
- What are the desired outcomes?
- What impact on our business will this change bring about?

DISCOVER NECESSARY DATA

Now we consider how to incorporate data into answering these challenges:

- What data needs to be collected to address those needs?
- Is it data the company is already collecting or is it new data that must be gathered?
- If so, what is the process to collect and parse it?

HIGH LEVEL USE CASE

Once we have the groundwork and the necessary data in place, we formulate potential high level use cases to test the process. Out of the potentials, one is selected for the PoC. We then move to analyzing what technologies and approaches will work best in bringing the PoC to life.

NEXT STEPS

Only when the use case is completely formed do we move ahead with the PoC.

STAGE 1: EXPLORE AND VALIDATE
BUILDING A BUSINESS CASE

Resisting the urge to take your big data business case and dive right in is paramount. This is generally uncharted territory and needs to be mapped and explored accordingly.

Conducting one or more PoC tests can:

- Reduce risks.
- Reduce time to implementation.
- Give your senior management and project leaders a better understanding of the tools and technology used in Big Data ecosystems.

The architecture of the PoC will vary widely from company to company, but certain components must be followed to maximize the added value before actual deployment. These include:

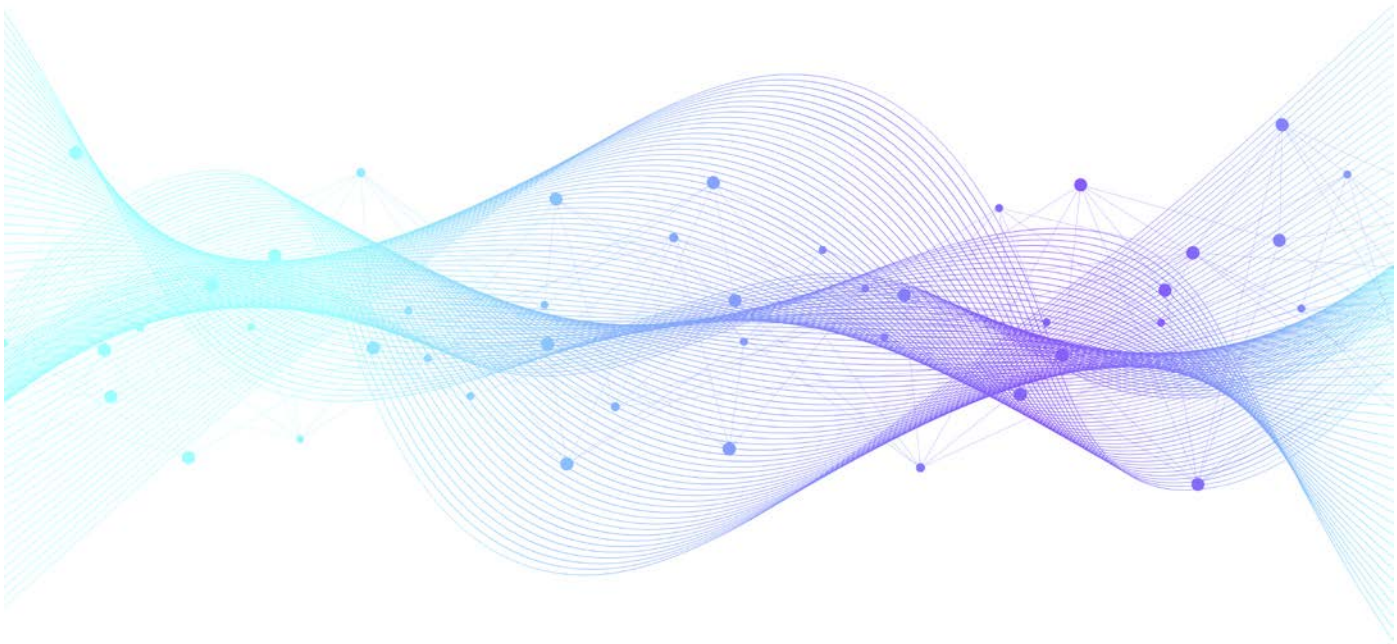
- Use real business data.
- Don't confuse a dashboard with business outcomes.
- Have a plan around governance.
- Set out to prove the value of the PoC.
- Be able to scale the PoC across the organization.
- Involve IT.

When analyzing results, keep in mind that multiple iterations might be necessary and that there may be more than one solution for each business case.

STAGE 2: DEPLOY AND GOVERN

When your firm has achieved PoC success, it can move on to the second stage of big data practical implementation. This is not a quick fix, nor is it a process of building on existing business foundations. It is a decidedly unconventional approach to implementing new business units. It will question everything that exists and likely ruffle more than a few established feathers as old practices are tossed aside. This process will take the longest amount of time of the three stages. Stage 2 is a process of decisions and questions, all of which need firm resolutions. What follows is a step-by-step guide to preparing your organization for deployment.

STAGE 2: DEPLOY AND GOVERN
EMPOWER A BIG DATA CHAMPION



The big data champion should not be a techie, but a C-suite level business executive who will both educate and inspire the rest of your company's business units to integrate more business cases and new kinds of data into the data analysis platform you are building. Because of the nature of the beast, CIOs and CTOs are frequently tabbed for the position but neither of those positions is inherently the best choice, because technology is not the driving force here—business is. In some cases, an additional position of Chief Data Officer (CDO) is created specifically to be the voice championing this new endeavor.

STAGE 2: DEPLOY AND GOVERN
SOLIDIFY DATA GOVERNANCE COMMITTEE

Data governance can be defined as the orchestration of people, processes and technology to enable an organization to leverage data as an enterprise asset. Because data governance has such a broad reach, the committee that manages it should be broad as well, including both IT and business unit heads. These committee members must approve and enact policies over four key areas.

DATA LINEAGE

This defines the life cycle of data, from its origins to analysis to where it is stored (or if it is stored) long-term. The lineage is represented visually and includes metadata and any transformation it undergoes.

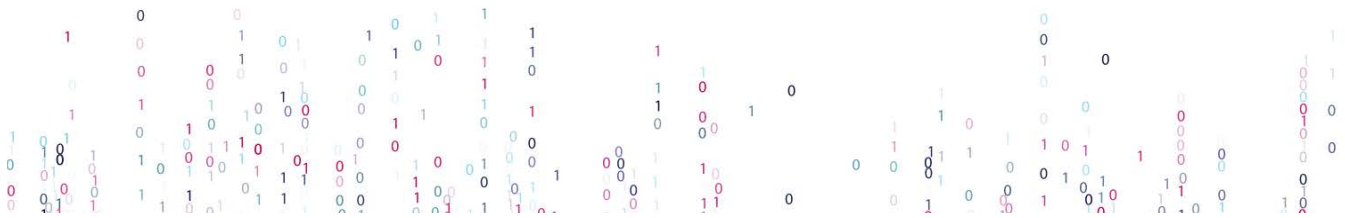
DATA STEWARDSHIP

An important aspect of data stewardship is assigning ownership to data. This helps build a data dictionary. Data stewardship has several goals, including:

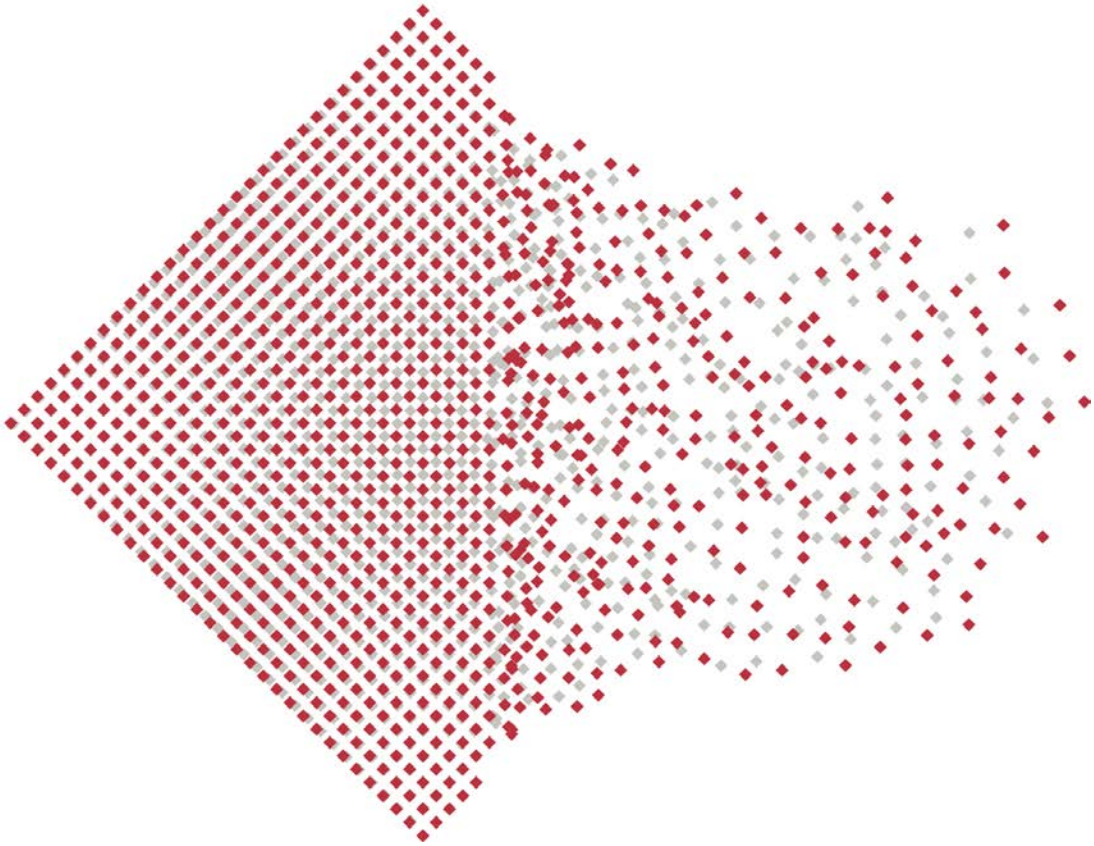
- Policies and procedures are in place and have become part of the corporate culture.
- Integration into the enterprise processes, such as project management and system development methodology.
- Designation and participation into every business function that owns data.
- Training for all involved parties in written form that is given on a regular schedule.
- Innovation in maintaining the vision of data quality and remediating data issues.
- Processes and procedures are written, approved and applied.

Your company's data stewards are in a position of unique power that must be weighed against what's best for the company. To this end, data stewards should:

- Secure the data, but only the data that needs securing.
- Control access to the data, without that control becoming a roadblock that could keep the data from being used when needed by the party that needs it.



STAGE 2: DEPLOY AND GOVERN
MODERN DATA ARCHITECTURE



With your people and policies now in place, we turn to the actual construction of your big data implementation. We start with where to house the data itself first, then turn to answering both traditional and new questions on how our data architecture should play out.

Finally, we'll play scientist for a few minutes and discuss why change is necessary from the traditional data model to the new structure being innovated and implemented daily by firms that want to use their data not only to compete, but also to win.

STAGE 2: DEPLOY AND GOVERN
DATA LAKE VS. DATA WAREHOUSE

Data warehouses are common items in enterprises and have been for decades. They can store vast amounts of data, but they do so in a very regimented format where all data have to fit predefined qualifications before being loaded. Data warehouses are synonymous with extract, transform and load (ETL) processes, because they're excellent at harnessing specific data and rejecting anything else that is entered incorrectly.

As data became more and more collectible and accessible, enterprises began seeing individual applications maintaining enormous amounts of data that had little-to-no value for other applications. This made for lots of information silos across an organization.

The first iteration of a solution was the data mart access layer over the central data warehouse which allowed greater access, but also caused more problems in terms of data governance, data ownership, and data accessibility.

The next step was the creation of the data lake, which was designed to:

- Store any form of data.
- Process that data so it can be analyzed and kept ready for use.

The best metaphor for the difference between data lakes and data warehouses can be found in the childhood toy, Legos. A Lego data warehouse would accept groups of Legos only after they had been arranged in a particular pattern, say a blue brick on top of a yellow brick on top of a red brick. Once that pattern is confirmed, the stack of Legos can enter the warehouse and made available for play.

STAGE 2: DEPLOY AND GOVERN
DATA LAKE VS. DATA WAREHOUSE

Conversely, a data lake allows you to dump all your existing Lego pieces into one giant container. While it might look like a mess from the outside, the container has an ad hoc scheme that allows users to define:

- How much data they need.
- When they need it.
- What types of data they need.
- What sources the data should come from.
- What analytics they wish to run from said data.

All these queries are tabulated by the data lake and resulting data is presented to the user.

The data lake life cycle encompasses four stages.

- **Data Acquisition:** The data acquired is as raw as possible, may exist in multiple forms and may need different mechanisms to be acquired.
- **Data Processing:** The data may need processing to derive valuable information. It may also be processed into an alternate model while still retaining its value.
- **Data Analysis:** The data is further analyzed to generate easier accessibility. The analysis requirements are driven by machine learning of information access patterns.
- **Data Storage:** Analyzed data is stored appropriately for both long-term health and ease of accessibility. A more modern pattern is called the "integrated analytics platform." It's designed with more structure for performing analytics and data science.

STAGE 2: DEPLOY AND GOVERN
TRADITIONAL VS. NEW QUESTIONS TO ANSWER

While data lakes and big data might be new terminology to your staff, data itself, along with its storage and functionality, should be familiar concepts to most enterprises. The following questions are key to any modern data architecture and must be answered to get your big data project off the ground before moving on to Stage 3.

WHAT STYLE OF ARCHITECTURE WILL YOU USE?

The distinction is between two schools of thought: Inmon vs. Kimball.

The table below reflects the strengths and weaknesses of each across a wide field of characteristics.

Characteristic	Kimball school	Inmon school
Primary audience	IT	End users
Role in organization	Key part of corporate decision-making	Transforms/retains operational data
Objective	Deliver sound technical solution based on proven methods	Deliver user-friendly solution making querying data easy with a solid response time
Data integration requirements	Individual business areas	Enterprise-wide integration
Structure of data	Metrics, performance measures, scorecards	Non-metrics and multi-purpose data
Scalability	Can adapt to highly volatile needs in a limited timeframe	Enterprise needs first app immediately
Staffing requirements	Small teams of general techs	Larger teams of data specialists
Time of delivery	Enterprise needs first app immediately	Enterprise has time to spare
Cost of deployment	Lower start-up costs; each successive project has a similar cost	Higher start-up cost; each successive project has a lower cost

STAGE 2: DEPLOY AND GOVERN
TRADITIONAL VS. NEW QUESTIONS TO ANSWER

WILL YOU USE DIMENSIONAL MODELS OR 3RD NORMALIZED FORM?

Dimensional models are optimized for online queries. They're comprised of two-axis tables with "fact" and "dimension" labels. Their best uses are reading, summarizing and analyzing numeric data. 3rd Normalized Form (3NF) is where no other columns from the parent table should exist without the reference table.

WHAT HARDWARE IS REQUIRED FOR YOUR DATA NEEDS?

Big data typically has big needs. If you're not addressing them with cloud computing solutions, there are three requirements to get your enterprise ready to handle Big Data.

- Storage: Large companies will use hyperscale computing environments, run frameworks like Hadoop and used flash storage to reduce latency time.
- Processing: Your servers must have enough processing power to support the data analysis without draining off your other resources.
- Networking: Sending mass amounts of data back and forth means a major upgrade for most businesses' networking hardware. If you don't have 10-gigabit connections in place, start researching them.

WHAT ETL TOOL SHOULD YOU USE?

As mentioned previously, ETL tools use batch processing to help business users conduct analysis and transform data to a database or a business intelligence platform. There is a wealth of different options. Major differentiators include error handling and what mode of transformation is used. Industry leaders (in terms of number of users) include:

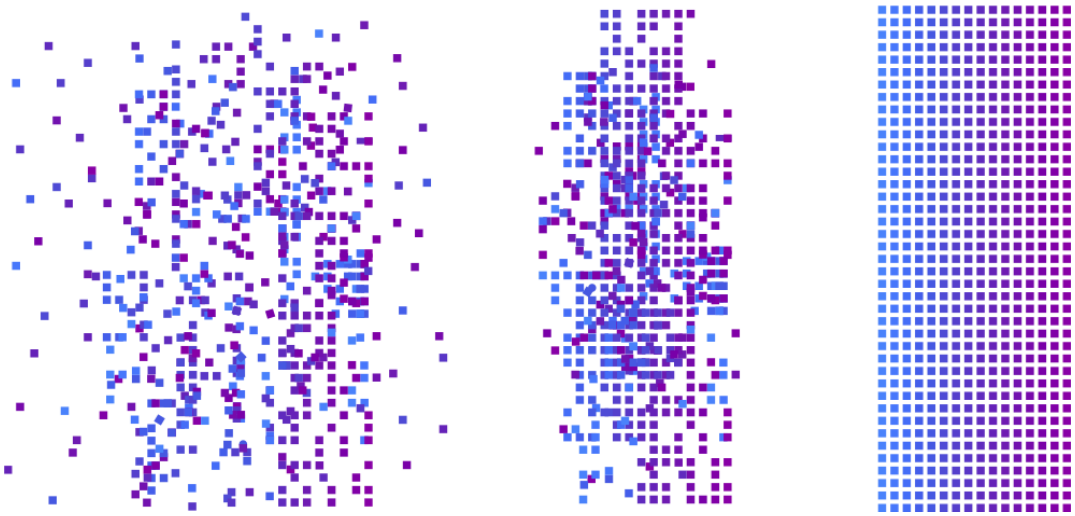
- Oracle Warehouse Builder (OWB)
- SAP Data Services
- IBM Infosphere Information Server
- PowerCenter Informatica
- SQL Server Integration Services (SSIS)

WHAT BUSINESS INTELLIGENCE TOOL SHOULD YOU USE?

BI tools are largely used to visually represent data so it can inform both business and non-business people. Factors to determine the best fit should include what hardware you're using and what forms of visualization improve comprehension. Some of the more well-regarded choices include:

- Tableau
- Datameer
- Pentaho
- Qlik

STAGE 2: DEPLOY AND GOVERN
TRADITIONAL VS. NEW QUESTIONS TO ANSWER



Data usage and technology are racing along the same accelerated path as the rest of digital technology. To push past traditional setups, ask these questions as you near the end of Stage 2.

DO WE NEED AN ENTERPRISE DATA WAREHOUSE?

The case for a data lake is presented earlier in this whitepaper, but the only clear answer of whether your firm needs one or both can only be answered by your organization. The traditional data warehouse has seen an evolution over recent years in order to keep pace with the demands of modern businesses. Traditional data warehouses were not built to process near real-time decisions. To stay relevant, data warehouses have to be modernized using repositories, virtualization and distributed processes. This allows the data warehouse to support workloads of relational and non-relational data.

DO WE NEED TO LEVERAGE HADOOP AND/OR NOSQL?

Hadoop is an open-sourced framework that supports the storing and processing of massive data sets. It works by distributing data sets across hundreds of parallel servers. Hadoop is weak when it comes to real-time processing of records, but is excellent at storing sensor data and processing it. It is particularly useful for industries such as financial services and manufacturing, where it is used to manage, store, and analyze data, or for feeding data into machine learning and AI devices.

STAGE 2: DEPLOY AND GOVERN
TRADITIONAL VS. NEW QUESTIONS TO ANSWER

WHAT PLATFORM SHOULD WE USE?

Choosing a platform for your big data enterprise work is a lot like buying a new car. If you don't know what you want going in, they're all going to start looking the same. Key decisions that need to be made up front include:

- **On-premise or cloud-based?** While more enterprises are choosing the cloud, mainly for scalability and ease of management reasons, on-premise makes much more sense if your firm has heavy regulations or compliance concerns.
- **Proprietary or open-source?** Open-source means there's a low cost of ownership and you can alter the platform to your custom needs. That sounds great, but it often isn't nearly that easy, which can lead to more expenses as your firm needs to bring in outside help to customize an open-source resource. With a proprietary platform, you know what you're getting when you make the purchase and have customer support as part of the package.
- **Batch or streaming?** If your analytical focus will be making real-time decisions, then streaming is your best practice. If your data is compiled, then analyzed for longer-term decisions, batch is the best choice.
- **What language should you use?** The architects and stewards of your data should be heavily involved in this decision, with their levels of experience and comfort playing a role alongside which language can most easily and completely meet your expectations. Three of the more well-used languages in data analytics are R, Python and Julia.
- **What style of architecture will you use?** Bleeding edge—generally seen as either untested or unreliable—can be of use in big data if you have a problem that the current solutions simply can't handle. The risks are self-explanatory, but there are software applications so new to the market that can help your goals, even if they haven't gained complete acceptance yet.

STAGE 3: OPTIMIZE AND DEMOCRATIZE

You've finished the final dress rehearsal, and you're ready for opening night. The architecture is built, your users are educated in the how and why, and your use cases are ready for deployment.

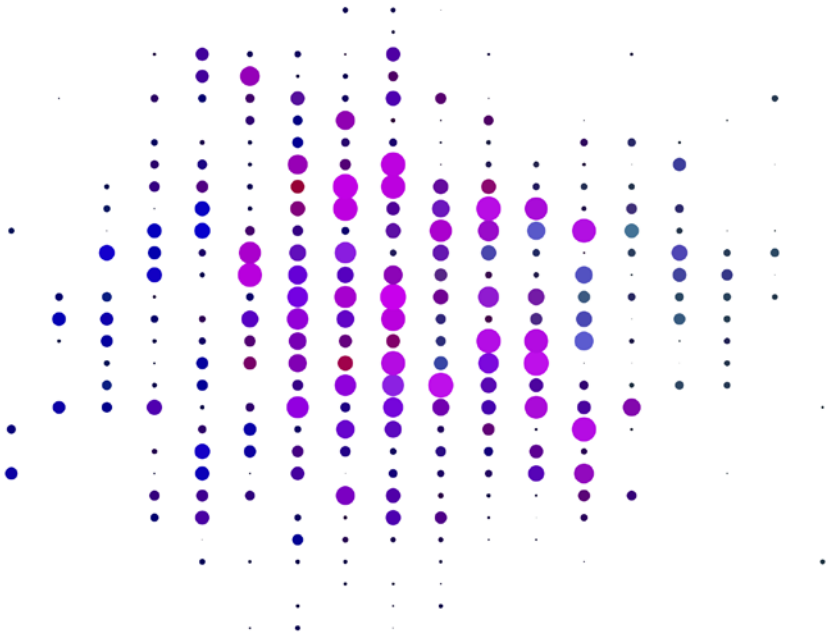
The next stages are making your strategies come to life and making your data available to consumers via self-service. Automation is a big priority here; you've built the machine, you don't want to hold everyone's hand every time they want to use it.

In Stage 2, you answered questions on what sort of platform you'll be using, what BI tools you'll use for visualization, what sort of storage vehicle you'll use, and more.

For stage 3, rolling out successful automation standards is essential to keeping up to date while allowing your data professionals to work on the important issues instead of turning into glorified maintenance workers. There are four aspects of big data that must be automated in order to keep it flowing in and out at the proper pace for collection and use.

- **Acquisition**: The challenge most businesses face is how to collect, collate and store information efficiently so it is accessible at a moment's notice. Automation ensures that data is collected from every available site, broken down and placed in the proper queue.
- **Distribution**: Once data is placed in your lake or warehouse, it still has to be sent out to the proper places for analysis, everywhere from the cloud to SaaS apps. Automation must be in place to direct this complex coordination to get data downstream.
- **Scalability**: Perhaps the toughest factor to anticipate. As your business grows, so to will your data. Automating your platform's ability to handle data flow performance is essential to keeping your analytics successful and secure.
- **Skills**: Automation is only as successful as the team behind it. Yours must be agile and knowledgeable from the get-go, with the understanding that turnover happens and must be accounted for.

STAGE 3: OPTIMIZE AND DEMOCRATIZE



Automation must extend to the onboarding of new user cases, setting up and maintaining user access and deploying analytics, with alerts and monitoring available for your data team, stewards and council members. This ensures that data is properly governed and can be audited at any time both for compliance checks and transparency.

Another component of the automation process is the importance of self-service tools to allow your end users the freedom to access data and analytics when they need them without the process of gaining access and needing supervision to run reports and generate analytics.

Self-service BI tools should be intuitive, easy to learn, and much less dependent on IT, which is often the biggest barrier for some employees. BI can be powerful tools that enhance your employees' production and should never be barriers that must be overcome to accomplish a task.

WINNING WITH ANALYTICS

Because a strategy only sticks when it becomes part of your company culture, evangelizing your work may be the most important part of the journey. Make sure to educate staff about the power of data-driven analytics, and it may turn out to be your biggest step towards success in the digital era.

With a solid foundation of big data, you'll need data science to analyze and extract meaning from all that data. More than just statistics, your approach to data science will determine how effectively your data gets put to use. Being able to predict what a customer will do next may seem like a long way off for many companies, but their competitive advantage is likely dependent on it.

As your analytics maturity evolves, you'll need cutting-edge tools, like AI and machine learning, to realize the benefits of prescriptive and cognitive analytics. Customers already expect a level of personalization that can only be achieved through well-timed nudges towards products and experiences that align with their past behavior. To unlock an Amazon-like customer experience, big data will be a prerequisite.

Don't wait on big data. It's an elemental building block that can be harnessed by the bold and innovative to accelerate past the competition.

Learn more about how PK is solving complex data challenges at pkglobal.com.

About the Author

Raj Nair is VP of Intelligence and Analytics at PK. His work includes researching creative solutions to challenging data problems, crafting elegant approaches to scaling data science algorithms and building plug-ins for the big data integration ecosystem.

About PK

PK is the experience engineering firm. Together with the world's most customer-obsessed companies, we combine great design and strong tech to build pioneering experiences that accelerate outcomes for your customers, partners, and employees. Through cutting-edge technology and a commitment to deep craftsmanship, we help our clients run the future.