COURSE PACK

FOR

Introduction to Big Data

e o

Course Code : 505-2-A

Course : BCA

Semester : V

Year : 2020-21

Course Leader: Mr. Mahesh Kumar Chaubey.

Course Instructor: Dr. Jitendra Singh

Forwarded by: HOD

Approved by: Director

Bharati Vidyapeeth (Deemed to be) University
Institute of management& research, New Delhi
An Iso 9001:2015 certified institute
"A+" grade accreditation by NAAC

INDEX

	MapReduce, R, WEKA.	160-167
MISC	Practice questions, internal Examination University Examination	

e o

Course Objective :

To introduce learner with Big Data Concept, decision making by doing analysis on the data and managing the data using Big Data Tools like Apache Hadoop, Pig and Hive. What are the problems of Big Data and how it can be solved by different tools? The concept of "big data" may have been around for a while, but the last few years have seen a sizeable swell in interest and media attention. Extremely recently, firms such as Cambridge Analytica and the way in which social media companies use the data collected about their users have hit the headlines and raised awareness even further of the enormous impact big data has on our lives every day. We looked into what it's actually like to work in big data and why students should specialize in the field.

Learning Outcome :

1. Good knowledge of Big Data Concepts
2. Knowledge of Decision making using analysis on the Big Data Introduction to
3. Big data Tools like Hadoop and Weka.

List of Modules

List of Modules	
Unit	Contents
1	Introduction: Big Data History, The Big Data Business Opportunity- Business Transformation Imperative, Big Data Business Model, Business Impact of Big Data
2	Big Data In Organization: Data Analytics Lifecycle, Data Scientist Roles and Responsibilities Discovery, Data Preparation, Model Planning, Model Building, Communicate Results, Operationalize, New Organizational Roles, Liberating Organizational Creativity.
3	Decision Theory And Strategy: Business Intelligence Challenge, Big Data User Interface Ramifications, Human Challenge of Decision Making, Strategy for Decision Making- Big Data Strategy Document, Case Study.
4	Value Creation Process: Understanding Big Data Value Creation, Value Creation Drivers, **Michael Porter's** Value Creation Models- Michael Porter's Five Forces Analysis, Michael Porter's Value Chain Analysis, Case Study.
5	Big Data User Experience: The Unintelligent User Experience, Understanding the Key Decisions to Build a Relevant User Experience, Using Big Data Analytics to Improve Customer Engagement, Uncovering and Leveraging Customer Insights, Big Data can Power a New Customer Experience.
6	Big Data Use Cases: The Big Data Envisioning Process 1. Research Business Intiatives, 2. Acquire and Analyze your Data, 3. Brainstorm New Ideas , 4. Prioritize Big Data Use Cases, 5. Document Next Steps, The Prioritization Process.
7	Big Data Architecture:
	New Big Data Architecture, Introducing Big Data Technologies _Apache Hadoop, MapReduce, R, WEKA etc.

Session Plan

Session	Topics	Learning Outcome
1	Introduction: Database ,data mining and Big Data History	To understand the importance of data science LO1
2	The Big Data Business Opportunity- Business Transformation	How bigdata transformed the process of BI LO1
3	Imperative, Big Data Business Models	BD business model and its use LO1
4	Business Impact of Big Data	Tech and business impact of BD LO1
5	Big Data in Organization:	BD Org structure LO1
6	Data Analytics Lifecycle	Stages and methodologies
7	Discovery, Data Preparation,	Data discovery activity LO1
8	Model Planning, Model Building	Modelling techs LO1
9	Communicate Results, Operationalize	Enhancing BD role in BI LO1
10	New Organizational Roles, Liberating Organizational Creativity	Enhancing BD role in BI
11	Decision Theory And Strategy	Intro to decision science LO2
12	Business Intelligence Challenge	BA and BI concepts LO2
13	Big Data User Interface Ramifications	BD UI LO2
14	Human Challenge of Decision Making	Spectrum OF BI LO2
15	Strategy for Decision Making- Big Data Strategy Document	Moreof decision science
16	Case Study	
17	Value Creation Process: Understanding Big.	Value addition to Big data outcomes LO2
18	Value Creation Drivers, Value Creation Models	same
19	Data Value Creation - Value Chain Analysis,	same
20	Case Study	
21	Big Data User Experience: The Unintelligent User Experience,	Role of end-user and deployment LO2
22	Understanding the Key Decisions to Build a Relevant User Experience	Creating user knowledge base
23	Using Big Data Analytics to Improve Customer Engagement,	CRM using bigdata
24	Uncovering and Leveraging Customer Insights	same

25	Big Data can Power a New Customer Experience.	Presentation mechanisms
26	Big Data Use Cases: The Big Data Envisioning Process Documents preparing Steps,	Documentation techniques LO2
27	Research Business Initiatives,	Tools and techniques used in R&D of BD
28	Acquire and Analyze your Data	Analysis and analytics
29	Acquire and Analyze your Data	Analysis and analytics
30	Brainstorm New Ideas, Prioritize Big Data Use Cases,	
31	Brainstorm New Ideas, Prioritize Big Data Use Cases,	Use cases of bigdata Big data Tools like Hadoop and Weka Big data Tools like Hadoop and Weka
32	The Prioritization Process.	Scheduling mechanisms
33	Big Data Architecture: New Big Data Architecture, ,	To understand the modern Big data Tools like Hadoop and Weka architecture of BD
34	Big Data Architecture: New Big Data Architecture, ,	To understand the modern Big data Tools like Hadoop and Weka architecture of BD
35	Introducing Big Data Technologies like Hadoop and MapReduce	practical
36	Introducing Big Data Technologies like Hadoop and MapReduce	practical
37	Concepts of R programming	practical
38	WEKA .	practical
39	Working on case studies	practical
40	Revision session	

1. Evaluation Criteria:

Component	Description	Weight age
First Internal Examination	First internal question paper will be based on first 3 unit of syllabus.	10marks
Second Internal Examination	Second internal question paper will be based on last 3 unit of syllabus.	10marks
CES Quiz	Class test will be conducted based on the different aspects Data analytics and business intelligence.	5 marks
CES Quiz	Class test will be conducted based on the different aspects of data analytics life cycle and decision science	5 marks
CES practical	Will have questions on analysis and analytics tools like R and Weka.	5 marks
Attendance	Above 75% - 10 marks Below 75% - 0 mark	10 marks
Note : All three CES will be mandatory. If any student misses anyone CES in that case the weightage of each CES would be 3.33 marks and if a student attempts all three CES then his/her best two CES will be considered, in that case the weightage would be 5 marks each.		

Unit1

Every now and then, new sources of data emerge that hold the potential to transform how organizations drive, or derive, business value. In the 1980s, we saw point-of-sale (POS) scanner data change the balance of power between consumer package goods (CPG) manufacturers like Procter & Gamble, Unilever, Frito Lay, and Kraft—and retailers like Walmart, Tesco, and Vons. The advent of detailed sources of data about product sales, soon coupled with customer loyalty data, provided retailers with unique insights about product sales, customer buying patterns, and overall market trends that previously were not available to any player in the CPG-to-retail value chain. The new data sources literally changed the business models of many companies.

Then in the late 1990s, web clicks became the new knowledge currency, enabling online merchants to gain significant competitive advantage over their brick-and-mortar counterparts. The detailed insights buried in the web logs gave online merchants new insights into product sales and customer purchase behaviors, and gave online retailers the ability to manipulate the user experience to influence (through capabilities like recommendation engines) customers' purchase choices and the contents of their electronic shopping carts. Again, companies had to change their business models to survive.

Today, we are in the midst of yet another data-driven business revolution. New sources of social media, mobile, and sensor or machine-generated data hold the potential to rewire an organization's value creation processes. Social media data provide insights into customer interests, passions, affiliations, and associations that can be used to optimize your customer engagement processes (from customer acquisition, activation, maturation, up-sell/cross-sell, retention, through advocacy development). Machine or sensor-generated data provide real-time data feeds at the most granular level of detail that enable predictive maintenance, product performance recommendations, and network optimization. In addition, mobile devices enable location-based insights and drive real-time customer engagement that allow

brick-and-mortar retailers to compete directly with online retailers in providing an improved, more engaging customer shopping experience.

The massive volumes (terabytes to petabytes), diversity, and complexity of the data are straining the capabilities of existing technology stacks. Traditional data warehouse and business intelligence architectures were not designed to handle petabytes of structured and unstructured data in real-time. This has resulted in the following challenges to both IT and business organizations:

- Rigid business intelligence, data warehouse, and data management architectures are impeding the business from identifying and exploiting fleeting, short-lived business opportunities.
- Retrospective reporting using aggregated data in batches can't leverage new analytic capabilities to develop predictive recommendations that guide business decisions.
- Social, mobile, or machine-generated data insights are not available in a timely manner in a world where the real-time customer experience is becoming the norm.
- Data aggregation and sampling destroys valuable nuances in the data that are key to uncovering new customer, product, operational, and market insights.

This blitz of new data has necessitated and driven technology innovation, much of it being powered by open source initiatives at digital media companies like Google (Big Table), Yahoo! (Hadoop), and Facebook (Hive and HBase), as well as universities (like Stanford, UC Irvine, and MIT). All of these big data developments hold the potential to paralyze businesses if they wait until the technology dust settles before moving forward. For those that wait, only bad things can happen:

- Competitors innovate more quickly and are able to realize compelling cost structure advantages.
- Profits and margins degenerate because competitors are able to identify, capture, and retain the most valuable customers.
- Market share declines result from not being able to get the right products to market at the right time for the right customers.
- Missed business opportunities occur because competitors have real-time listening devices rolling up real-time customer sentiment, product performance problems, and immediately-available monetization opportunities.

The time to move is now, because the risks of not moving can be devastating.

The Business Transformation Imperative

The big data movement is fueling a business transformation. Companies that are embracing big data as business transformational are moving from a retrospective, rearview mirror view of the business that uses partial slices of aggregated or sampled data in batch to monitor the business to a forward-looking, predictive view of operations that leverages all available data—including structured and unstructured data that may sit outside the four walls of the organization—in real-time to optimize business performance (see Table 1-1).

Table 1-1: Big data is about business transformation.

Today's Decision Making	Big Data Decision Making
"Rearview Mirror" hindsight	"Forward looking" recommendations
Less than 10% of available data	Exploit all data from diverse sources
Batch, incomplete, disjointed	Real-time, correlated, governed
Business Monitoring	Business Optimization

Think of this as the advent of the real-time, predictive enterprise!

In the end, it's all about the data. Insight-hungry organizations are liberating the data that is buried deep inside their transactional and operational systems, and integrating that data with data that resides outside the organization's four walls (such as social media, mobile, service providers, and publicly available data). These organizations are discovering that data—and the key insights buried inside the data—has the power to transform how organizations understand their customers, partners, suppliers, products, operations, and markets. In the process, leading organizations are transforming their thinking on data, transitioning from treating data as an operational cost to be minimized to a mentality that nurtures data as a strategic asset that needs to be acquired, cleansed, transformed, enriched, and analyzed to yield actionable insights. Bottom-line: companies are seeking ways to acquire even more data that they can leverage throughout the organization's value creation processes.

Walmart Case Study

Data can transform both companies and industries. Walmart is famous for their use of data to transform their business model.

> *The cornerstone of his [Sam Walton's] company's success ultimately lay in selling goods at the lowest possible price, something he was able to do by pushing aside the middlemen and directly haggling with manufacturers to bring costs down. The idea to "buy it low, stack it high, and sell it cheap" became a sustainable business model largely because Walton, at the behest of David Glass, his eventual successor,* **heavily invested in software that could track consumer behavior in real time from the bar codes read at Walmart's checkout counters.**
>
> *He shared the real-time data with suppliers to create partnerships that allowed Walmart to exert significant pressure on manufacturers to improve their productivity and become ever more efficient. As Walmart's influence grew, so did its power to nearly dictate the price, volume, delivery, packaging, and quality of many of its suppliers' products. The upshot: Walton flipped the supplier-retailer relationship upside down.*[1]

Walmart up-ended the balance of power in the CPG-to-retailer value chain. Before they had access to detailed POS scanner data, the CPG manufacturers (such as Procter & Gamble, Unilever, Kimberley Clark, and General Mills,) dictated to the retailers how much product they would be allowed to sell, at what prices, and using what promotions. But with access to customer insights that could be gleaned from POS data, the retailers were now in a position where they knew more about their customers' behaviors—what products they bought, what prices they were willing to pay, what promotions worked the most effectively, and what products they tended to buy in the same market basket. Add to this information the advent of the customer loyalty card, and the retailers knew in detail what products at what prices under what promotions appealed to which customers. Soon, the retailers were dictating terms to the CPG manufacturers—how much product they wanted to sell (demand-based forecasting), at what prices (yield and price optimization), and what promotions they wanted (promotional effectiveness). Some of these retailers even went one step further and figured out how to monetize their POS data by selling it back to the CPG manufacturers. For example, Walmart provides a data service to their CPG manufacturer partners, called Retail Link, which provides sales and inventory data on the manufacturer's products sold through Walmart.

Across almost all organizations, we are seeing multitudes of examples where data coupled with advanced analytics can transform key organizational business processes, such as:

[1] "The 12 greatest entrepreneurs of our time" Fortune/CNN Money (`http://money.cnn.com/galleries/2012/news/companies/1203/gallery.greatest-entrepreneurs.fortune/12.html`)

- **Procurement:** Identify which suppliers are most cost-effective in delivering products on-time and without damages.
- **Product Development:** Uncover product usage insights to speed product development processes and improve new product launch effectiveness.
- **Manufacturing:** Flag machinery and process variances that might be indicators of quality problems.
- **Distribution:** Quantify optimal inventory levels and optimize supply chain activities based on external factors such as weather, holidays, and economic conditions.
- **Marketing:** Identify which marketing promotions and campaigns are most effective in driving customer traffic, engagement, and sales, or use attribution analysis to optimize marketing mixes given marketing goals, customer behaviors, and channel behaviors.
- **Pricing and Yield Management:** Optimize prices for "perishable" goods such as groceries, airline seats, concert tickets and fashion merchandise.
- **Merchandising:** Optimize merchandise markdown based on current buying patterns, inventory levels, and product interest insights gleaned from social media data.
- **Sales:** Optimize sales resource assignments, product mix, commissions modeling, and account assignments.
- **Store Operations:** Optimize inventory levels given predicted buying patterns coupled with local demographic, weather, and events data.
- **Human Resources:** Identify the characteristics and behaviors of your most successful and effective employees.

The Big Data Business Model Maturity Index

Customers often ask me:

- How far can big data take us from a business perspective?
- What could the ultimate endpoint look like?
- How do I compare to others with respect to my organization's adoption of big data as a business enabler?
- How far can I push big data to power—or even transform—my value creation processes?

To help address these types of questions, I've created the Big Data Business Model Maturity Index. This index provides a benchmark against which organizations can

measure themselves as they look at what big data-enabled opportunities may lay ahead. Organizations can use this index to:

- Get an idea of where they stand with respect to exploiting big data and advanced analytics to power their value creation processes and business models (their current state).
- Identify where they want to be in the future (their desired state).

Organizations are moving at different paces with respect to how they are adopting big data and advanced analytics to create competitive advantages for themselves. Some organizations are moving very cautiously because they are unclear where and how to start, and which of the bevy of new technology innovations they need to deploy in order to start their big data journeys. Others are moving at a more aggressive pace to integrate big data and advanced analytics into their existing business processes in order to improve their organizational decision-making capabilities.

However, a select few are looking well beyond just improving their existing business processes with big data. These organizations are aggressively looking to identify and exploit new data monetization opportunities. That is, they are seeking out business opportunities where they can either sell their data (coupled with analytic insights) to others, integrate advanced analytics into their products to create "intelligent" products, or leverage the insights from big data to transform their customer relationships and customer experience.

Let's use the Big Data Business Model Maturity Index depicted in Figure 1-1 as a framework against which you can not only measure where your organization stands today, but also get some ideas on how far you can push the big data opportunity within your organization.

Figure 1-1: Big Data Business Model Maturity Index

Business Monitoring

In the *Business Monitoring* phase, you deploy Business Intelligence (BI) and traditional data warehouse capabilities to monitor, or report on, on-going business performance. Sometimes called *business performance management*, business monitoring uses basic analytics to flag under- or over-performing areas of the business, and automates sending alerts with pertinent information to concerned parties whenever such a situation occurs. The Business Monitoring phase leverages the following basic analytics to identify areas of the business requiring more investigation:

- Trending, such as time series, moving averages, or seasonality
- Comparisons to previous periods (weeks, months, etc.), events, or campaigns (for example, a back-to-school campaign)
- Benchmarks against previous periods, previous campaigns, and industry benchmarks
- Indices such as brand development, customer satisfaction, product performance, and financials
- Shares, such as market share, share of voice, and share of wallet

The Business Monitoring phase is a great starting point for your big data journey as you have already gone through the process—via your data warehousing and BI investments—of identifying your key business processes and capturing the KPIs, dimensions, metrics, reports, and dashboards that support those key business processes.

Business Insights

The *Business Insights* phase takes business monitoring to the next step by leveraging new unstructured data sources with advanced statistics, predictive analytics, and data mining, coupled with real-time data feeds, to identify material, significant, and actionable business insights that can be integrated into your key business processes. This phase looks to integrate those business insights back into the existing operational and management systems. Think of it as "intelligent" dashboards, where instead of just presenting tables of data and graphs, the application goes one step further to actually uncover material and relevant insights that are buried in the detailed data. The application can then make specific, actionable recommendations, calling out an observation on a particular area of the business where specific actions

can be taken to improve business performance. One client called this phase the "Tell me what I need to know" phase. Examples include:

- In marketing, uncovering observations that certain in-flight campaign activities or marketing treatments are more effective than others, coupled with specific recommendations as to how much marketing spend to shift to the more effective activities.
- In manufacturing, uncovering observations that certain production machines are operating outside of the bounds of their control charts (for example, upper limits or lower limits), coupled with a prioritized maintenance schedule with replacement part recommendations for each problem machine.
- In customer support, uncovering observations that certain gold card members' purchase and engagement activities have dropped below a certain threshold of normal activity, with a recommendation to e-mail them a discount coupon.

The following steps will transition your organization from the business monitoring to the business insights stage.

1. Invest the time to understand how users are using existing reports and dashboards to identify problems and opportunities. Check for situations where users are printing reports and making notes to the side of the reports. Find situations where users are downloading the reports into Excel or Access and capture what these users are doing with the data once they have it downloaded. Understanding what your users are doing with the existing reports and downloads is "gold" in identifying the areas where advanced analytics and real-time data can impact the business.
2. Understand how downstream constituents—those users that are the consumers of the analysis being done in Step 1—are using and making decisions based on the analysis. Ask, "What are these constituents doing with the results of the analysis? What actions are they trying to take? What decisions are they trying to make given the results of the analysis?"
3. Launch a prototype or pilot project that provides the opportunity to integrate detailed transactional data and new unstructured data sources with real-time data feeds and predictive analytics to automatically uncover potential problems and opportunities buried in the data (Insights), and generate actionable recommendations.

Business Optimization

The *Business Optimization* phase is the level of business maturity where organizations use embedded analytics to automatically optimize parts of their business operations. To many organizations, this is the Holy Grail where they can turn over certain parts of their business operations to analytic-powered applications that automatically optimize the selected business activities. Business optimization examples include:

- Marketing spend allocation based on in-flight campaign or promotion performance
- Resource scheduling based on purchase history, buying behaviors, and local weather and events
- Distribution and inventory optimization given current and predicted buying patterns, coupled with local demographic, weather, and events data
- Product pricing based on current buying patterns, inventory levels, and product interest insights gleaned from social media data
- Algorithmic trading in financial services

The following steps will transition your organization from the Business Insights phase to the Business Optimization phase:

1. The Business Insights phase resulted in a list of areas where you are already developing and delivering recommendations. Use this as the starting point in assembling the list of areas that are candidates for optimization. Evaluate these business insights recommendations based on the business or financial impact, feasibility of success, and their relative recommendation performance or effectiveness.
2. For each of the optimization candidates, identify the supporting business questions and decision-making process (the analytic process). You will also need to identify the required data sources and timing/latency of data feeds (depending on decision-making frequency and latency), the analytic modeling requirements, and the operational system and user experience requirements.
3. Finally, conduct "Proof of Value" or develop a prototype of your top optimization candidates to verify the business case, the financials (return on investment—ROI), and analytics performance.

You should also consider the creation of a formal analytics governance process that enables human subject matter experts to audit and evaluate the effectiveness and relevance of the resulting optimization models on a regular basis. As any good data scientist will tell you, the minute you build your analytic model it is obsolete due to changes in the real-world environment around it.

Data Monetization

The *Data Monetization* phase is where organizations are looking to leverage big data for net new revenue opportunities. While not an exhaustive list, this includes initiatives related to:

- Packaging customer, product, and marketing insights for sale to other organizations
- Integrating analytics directly into their products to create "intelligent" products
- Leveraging actionable insights and personalized recommendations based on customer behaviors and tendencies to upscale their customer relationships and dramatically rethink their "customer experience"

An example of the first type of initiative could be a smartphone app where data and insights about customer behaviors, product performance, and market trends are sold to marketers and manufacturers. For example, MapMyRun (`www.MapMyRun.com`) could package the customer usage insights from their smartphone application with audience and product insights for sale to sports apparel manufacturers, sporting goods retailers, insurance companies, and healthcare providers.

An example of the second type of initiative could be companies that leverage new big data sources (sensor data or user click/selection behaviors) with advanced analytics to create "intelligent" products, such as:

- Cars that learn your driving patterns and behaviors and use the data to adjust driver controls, seats, mirrors, brake pedals, dashboard displays, and other items to match your driving style.
- Televisions and DVRs that learn what types of shows and movies you like and use the data to search across the different cable channels to find and automatically record similar shows for you.
- Ovens that learn how you like certain foods cooked and uses the data to cook them in that manner automatically, and also include recommendations for other foods and cooking methods that others like you enjoy.

An example of the third type of initiative could be companies that leverage actionable insights and recommendations to "up-level" their customer relationships and dramatically rethink their customer's experience, such as:

- Small, medium business (SMB) merchant dashboards from online marketplaces that compare current and in-bound inventory levels with customer buying patterns to make merchandising and pricing recommendations
- Investor dashboards that assess investment goals, current income levels, and current financial portfolios to make specific asset allocation recommendations

The following steps will be useful in helping transition to the Data Monetization phase.

1. Identify your target customers and their desired solutions. Focus on identifying solutions that improve customers' business performance and help them make money. As part of that process, you will need to detail out the personas of the economic decision-makers. Invest time shadowing these decision-makers to understand what decisions they are trying to make, how frequently, and in what situations. Spend the time to gather details of what they are trying to accomplish, versus focusing on trying to understand what they are doing.
2. Inventory your current data assets. Capture what data you currently have. Also, identify what data you could have with a little more effort. This will require you to look at how the source data is being captured, to explore additional instrumentation strategies to capture even more data, and explore external sources of data that, when combined with your internal data, yields new insights on your customers, products, operations, and markets.
3. Determine the analytics, data enrichment, and data transformation processes necessary to transform your data assets into your targeted customers' desired solutions. This should include identifying:
    - The business questions and business decisions that your targeted personas are trying to ask and answer
    - The advanced analytics (algorithms, models), data augmentation, transformation, and enrichment processes necessary to create solutions that address your targeted persona's business questions and business decisions
    - Your targeted persona's user experience requirements, including their existing work environments and how you can leverage new mobile and data visualization capabilities to improve that user experience

Business Metamorphosis

The *Business Metamorphosis* phase is the ultimate goal for organizations that want to leverage the insights they are capturing about their customers' usage patterns, product performance behaviors, and overall market trends to transform their business models into new services in new markets. For example:

- Energy companies moving into the home energy optimization business by recommending when to replace appliances (based on predictive maintenance) and even recommending which brands to buy based on the performance of different appliances compared to customer usage patterns, local weather, and environmental conditions, such as local water conditions and energy costs.
- Farm equipment manufacturers transforming into farming optimization businesses by understanding crop performance given weather and soil conditions, and making seed, fertilizer, pesticide, and irrigation recommendations.
- Retailers moving into the shopping optimization business by recommending specific products given a customer's current buying patterns compared with others like them, including recommendations for products that may not even reside within their stores.
- Airlines moving into the "Travel Delight" business of not only offering discounts on air travel based on customers' travel behaviors and preferences, but also proactively finding and recommending deals on hotels, rental cars, limos, sporting or musical events, and local sites, shows, and shopping in the areas that they are visiting.

In order to make the move into the Business Metamorphosis phase, organizations need to think about moving away from a product-centric business model to a more platform- or ecosystem-centric business model.

Let's drill into this phase by starting with a history lesson. The North American video game console market was in a massive recession in 1985. Revenues that had peaked at $3.2 billion in 1983, fell to $100 million by 1985—a drop of almost 97 percent. The crash almost destroyed the then-fledgling industry and led to the bankruptcy of several companies, including Atari. Many business analysts doubted the long-term viability of the video game console industry.

There were several reasons for the crash. First, the hardware manufacturers had lost exclusive control of their platforms' supply of games, and consequently lost the ability to ensure that the toy stores were never overstocked with products. But the main culprit was the saturation of the market with low-quality games. Poor quality games, such as *Chase the Chuck Wagon* (about dogs eating food, bankrolled by the dog food company Purina), drove customers away from the industry.

The industry was revitalized in 1987 with the success of the Nintendo Entertainment System (NES). To ensure ecosystem success, Nintendo instituted strict measures to ensure high-quality games through licensing restrictions, maintained strict control of industry-wide game inventory, and implemented a security lockout system that only allowed certified games to work on the Nintendo platform. In the process, Nintendo ensured that third-party developers had a ready and profitable market.

As organizations contemplate the potential of big data to transform their business models, they need to start by understanding how they can leverage big data and the resulting analytic insights to transform the organization from a product-centric business model into a platform-centric business model. Much like the Nintendo lesson, you accomplish this by creating a marketplace that enables others—like app developers, partners, VARs, and third party solution providers—to make money off of your platform.

Let's build out the previous example of an energy company moving into the home energy optimization business. The company could capture home energy and appliance usage patterns that could be turned into insights and recommendations. For example, with the home energy usage information, the company could recommend when consumers should run their high energy appliances, like washers and dryers, to minimize energy costs. The energy company could go one step further and offer a service that automatically manages when the washer, dryer, and other high-energy appliances run—such as running the washer and dryer at 3:00 a.m. when energy prices are lower.

With all of the usage information, the company is also in a good position to predict when certain appliances might need maintenance (for example, monitoring their usage patterns using Six Sigma control charts to flag out-of-bounds performance problems). The energy company could make preventive maintenance recommendations to the homeowner, and even include the names of three to four local service dealers and their respective Yelp ratings.

But wait, there's more! With all of the product performance and maintenance data, the energy company is also in an ideal position to recommend which appliances are the best given the customer's usage patterns and local energy costs. They could become the *Consumer Reports* for appliances and other home and business equipment by recommending which brands to buy based on the performance of different appliances as compared to their customers' usage patterns, local weather, environmental conditions, and energy costs.

Finally, the energy company could package all of the product performance data and associated maintenance insights and sell the data and analytic insights back to the manufacturers who might want to know how their products perform within certain usage scenarios and versus key competitors.

In this scenario, there are more application and service opportunities than any single vendor can reasonably supply. That opens the door to transform to a platform-centric business model that creates a platform or ecosystem that enables third party developers to deliver products and services on that platform. And, of course, this puts the platform provider in a position to take a small piece of the "action" in the process, such as subscription fees, rental fees, transaction fees, and referral fees.

Much like the lessons of Nintendo with their third-party video games, and Apple and Google with their respective apps stores, creating such a platform not only benefits your customers who are getting access to a wider variety of high-value apps and services in a more timely manner, it also benefits the platform provider by creating a high level of customer dependency on your platform (for example, by increasing the switching costs).

Companies that try to do all of this on their own will eventually falter because they'll struggle to keep up with the speed and innovation of smaller, hungrier organizations that can spot and act on a market opportunity more quickly. Instead of trying to compete with the smaller, hungrier companies, enable such companies by giving them a platform on which they can quickly and profitability build, market, and support their apps and solutions.

So how does your company make the business metamorphosis from a product to a platform or ecosystem company? Three steps are typically involved:

1. Invest the time researching and shadowing your customers to understand their desired solutions. Focus on what the customer is trying to accomplish, not what they are doing. Think more broadly about their holistic needs, such as:
    - Feeding the family, not just cooking, buying groceries, and going to restaurants
    - Personal transportation, not just buying or leasing cars, scheduling maintenance, and filling the car with gas
    - Personal entertainment, not just going to the theater, buying DVDs, or downloading movies
2. Understand the potential ecosystem players (e.g., developers) and how they could make money off of your platform. Meet with potential ecosystem players to brainstorm and prioritize their different data monetization opportunities to:
    - Clarify, validate, and flush out the ecosystem players' business case
    - Identify the platform requirements that allow the ecosystem players to easily instrument, capture, analyze, and act on insights about their customers' usage patterns and product performance

3. As the platform provider, focus product development, marketing, and partnering efforts on ensuring that the platform:
    - Is easy to develop on and seamlessly supports app developer marketing, sales, service, and support (for example, app fixes, new product releases, addition of new services)
    - Is scalable and reliable with respect to availability, reliability, extensibility, data storage, and analytic processing power
    - Has all the tools, data processing, analytic capabilities (such as recommendation engines), and mobile capabilities to support modern application development
    - Simplifies how qualified third parties make money with respect to contracts, terms and conditions, and payments and collections
    - Enables developers to easily capture and analyze customer usage and product performance data in order to improve their customers' user experience and help the developers optimize their business operations (for example, pricing, promotion, and inventory management)

    This step includes creating user experience mockups and prototypes so that you can understand *exactly* how successfully and seamlessly customers are able to interact with the platform (for example, which interface processes cause users the most problems, or where do users spend an unusual amount of time). Mockups are ideal for web- or smartphone-based applications, but don't be afraid to experiment with different interfaces that have different sets of test customers to improve the user experience. Companies like Facebook have used live experimentation to iterate quickly in improving their user experience. Heavily instrument or tag every engagement point of the user experience so that you can see the usage patterns and potential bottlenecks and points of frustration that the users might have in interacting with the interface.

As your organization advances up the big data business model maturity index, you will see three key cultural or organizational changes:

- Data is becoming a corporate asset to exploit, not a cost of business to be minimized. Your organization starts to realize that data has value, and the more data you have at the most granular levels of detail, the more insights you will be able to tease out of the data.
- Analytics and the supporting analytic algorithms and analytic models are becoming organizational intellectual property that need to be managed, nurtured, and sometimes even protected legally. The models that profile, segment, and acquire your customers, the models that you measure campaign or healthcare treatment effectiveness, the models that you use to predict equipment

maintenance—all of these are potential differe
can be exploited for differentiated business va
protected.

- Your organization becomes more comfortab
the data and analytics. The business users and
more confident in the data and begin trustin
about their business. The need to rely solely
(Highest Paid Person's Opinion) gives way to
values making decisions based on what the dat

e o

Big Data Business Model Maturity Observations

The first observation is that the first three phases of the Big Data Business Model Maturity Index are internally focused—optimizing an organization's internal business processes, as highlighted in Figure 1-2. This part of the maturity index leverages an organization's data warehouse and business intelligence investments, especially the key performance indicators, data transformation algorithms, data models, and reports and dashboards around the organization's key business processes. There are four big data capabilities that organizations can leverage to enhance their existing internal business processes as part of the maturity process:

- Mine all the transactional data at the lowest levels of detail much of which is not being analyzed today due to data warehousing costs. We call this the organizational "dark" data.
- Integrate unstructured data with detailed structured (transactional) data to provide new metrics and new dimensions against which to monitor and optimize key business processes.
- Leverage real-time (or low-latency) data feeds to accelerate the organization's ability to identify and act upon business and market opportunities in a timely manner.
- Integrate predictive analytics into your key business processes to uncover insights buried in the massive volumes of detailed structured and unstructured data. (Note: having business users slice and dice the data to uncover insights worked fine when dealing with gigabytes of data, but doesn't work when dealing with terabytes and petabytes of data.)

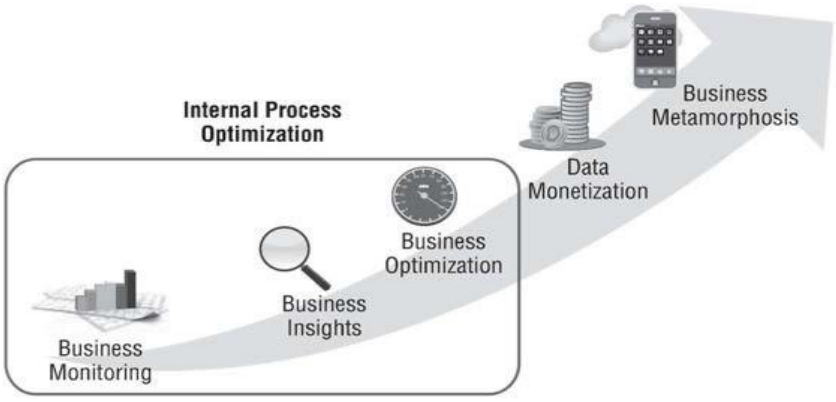


Figure 1-2: Big Data Business Model Maturity Index: Internal Process Optimization

The second observation is that the last two phases of the Big Data Business Model Maturity Index are externally focused—creating new monetization opportunities based upon the customer, product, and market insights gleaned from the first three phases of the maturity index, as highlighted in Figure 1-3. This is the part of the big data journey that catches most organizations' attention; the opportunity to leverage the insights gathered through the optimization of their internal business processes to create new monetization opportunities. We call this area of the Big Data Business Model Maturity Index the four Ms of big data: *Make Me More Money*!

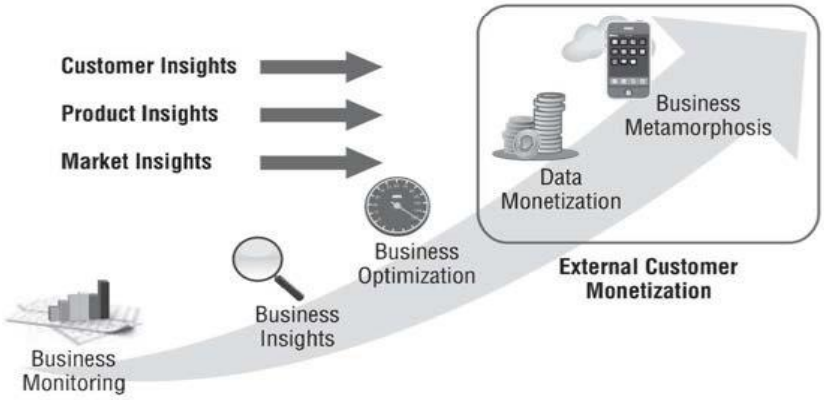


Figure 1-3: Big Data Business Model Maturity Index: External Customer Monetization

Summary

This chapter introduced you to the business drivers behind the big data movement. I talked about the bevy of new data sources available covering structured, semi-structured (for example, log files generated by sensors), and unstructured (e.g., text documents, social media postings, physician comments, service logs, consumer comments) data. I also discussed the growing sources of publicly available data that reside outside the four walls of an organization.

This chapter also briefly covered why traditional data warehousing and business intelligence technologies are struggling with the data volumes, the wide variety of new unstructured data sources and the high-velocity data that shrinks the latency between when a data event occurs and when that data is available for analysis and actions.

Probably most importantly, you learned how leading organizations are leveraging big data to transform their businesses—moving from a retrospective view of the business with partial chunks of data in batch to monitor their business performance, to an environment that integrates predictive analytics with real-time data feeds that leverage all available data in order to optimize the business.

Finally, you were introduced to the concept of the Big Data Business Model Maturity Index as a vehicle for helping your organization identify where they are today, and map out where they could be with respect to leveraging big data to uncover new monetization and business metamorphosis opportunities. Several "How To" guides were included in this chapter to help your organization move from one phase to the next in the maturity index.

Chapter 1 hinted at how retail point-of-sale (POS) scanner data caused a big data revolution that transformed the Consumer Package Goods (CPG) and retail industries in the late 1980s and 1990s. Let's spend a bit more time on that event, as there are some valuable lessons to be learned that apply to today's big data revolution.

Consumer Package Goods and Retail Industry Pre-1988

In the 1970s and early 1980s, CPG manufacturers such as Procter & Gamble, Unilever, Colgate-Palmolive, Kraft, and General Mills, to name but a few, and large grocery, drug, and mass merchandise retailers made their marketing decisions based on bi-monthly Nielson store audit data. That is, Nielsen would send people into a sample of stores (in only about 12 cities across the United States) to conduct physical audits—to count how much product was on the shelf, the price of the product, how much linear footage the product had across the front of the shelf, product sales within that store, and other data. Nielsen would aggregate this information by product category in order to calculate market share by volume and revenue, share of shelf space, etc. The results of the audits were then delivered every two months to the retailers and CPG manufacturers, usually in booklet format. CPG manufacturers could also request the data in tape format, but the data volumes were easily in the megabyte range.

So a company like Procter & Gamble would use this data for their Crest brand toothpaste, combined with their own internal orders and shipments data, to compare their sales to other toothpaste brands in the dentifrice product category. The Crest brand team would use this data to plan, execute, and measure their marketing strategies including promotional spending, new product introductions, and pricing decisions.

NOTE Not only was Crest's data and analysis only available bimonthly, but the audit books were delivered several weeks after the close of the audit period due to the data cleansing, alignment, and analytics that Nielsen had to do to ensure accuracy and consistency of the data. Needless to say, it could be two to three months after a marketing campaign ended before the manufacturer had any idea how successful their campaign had been in driving incremental revenue, unit sales, and market share increases.

Then in the late 1980s, Information Resources Inc. (IRI) introduced their Infoscan product, which combined retail POS scanner systems with barcodes (universal product codes—UPC) to revolutionize the CPG-to-retail value chain process. Data volumes jumped from megabytes to gigabytes and soon to terabytes of retail sales data. Existing mainframe-based executive information systems (EIS) broke under the data volumes, which necessitated a next generation of data processing capabilities as represented by client-server architectures, and data platforms such as Britton-Lee, Red Brick, Teradata, Sybase IQ, and Informix. This also saw the birth of the Business Intelligence (BI) software industry (e.g., Brio, Cognos, Microstrategy, Business Objects), as many early BI companies can trace their origins to the late 1980s Procter & Gamble-led "decision support" projects.

So data volumes jumped dramatically, breaking existing technology platforms and necessitating a next generation of data platforms and analytic capabilities. Sound familiar? But the most interesting thing wasn't the jump in data volumes that necessitated a new generation of data processing and analytic capabilities. The most interesting and relevant aspect of the scanner POS revolution was how companies like Procter & Gamble, Frito Lay, Tesco, and Walmart were able to leverage this new source of data and new technology innovations to create completely new business applications—business applications that previously were impossible to create. Much like what was discussed in Chapter 1 about moving to the Business Insights and Business Optimization phases of the Big Data Business Model Maturity Index, these new business applications leveraged the detailed POS data and new data management and analytic innovations to create new application categories such as:

- **Demand-based Forecasting**, where CPG manufacturers could create and update their product forecasts in near real-time based on what products were selling in the retail outlets for that current week. This was a major breakthrough for companies that sold products that were considered staples—products with relatively consistent consumption, such as toilet paper, toothpaste, soap, detergent, and most food products.

- **Supply Chain Optimization**, where detailed product sales data at the UPC level, combined with up-to-date inventory data (at each distribution center, at each store, and on order), allowed retailers and CPG manufacturers to drive excess inventory, holding, and distribution costs out of the supply chain. The savings in reduced capital required to maintain the supply chain was significant in itself, not to mention savings in other areas such as spoilage, shrinkage and unnecessary labor, distribution center and transportation costs.
- **Trade Promotion Effectiveness**, where CPG manufacturers could more quickly quantify what trade promotions were working most effectively with which retailers, and do this analysis in a more timely manner to actually impact current trade promotion programs.
- **Market Basket Analysis**, where retailers could gain intimate knowledge of what products sold together to what customers at what times of year. This insight could be used to not only change how retailers would lay out their stores, but also put the retailer in a superior position to inform the CPG manufacturers of the optimal cross-product category promotional opportunities.
- **Category Management** was an entirely new concept championed by leading CPG manufacturers. Much like how brand management had revolutionized the management and marketing of brands just a couple of decades earlier, Category Management allowed CPG manufacturers to re-apply many of the brand management concepts, but at a product category level (for example, heavy duty detergents, toilet paper, diapers, or dentifrice) in order to drive overall category demand, efficiencies, and profitability. This created a common language where retailers and CPG manufacturers could collaborate to drive overall category sales and profitability. The "category champion," which was the title or role given to the CPG manufacturer, was responsible for the management of the retailer's in-store product category including pricing, replenishment, promotions, and inventory.
- **Price and Yield Optimization**, where organizations are determining optimal product prices—at the individual store and seasonality levels—by combining real-time sales data with historical sales (demand), product sales seasonality trends and available inventory (on-hand and on-order). For example, retailers know that they can charge more for the same products in high tourist areas (e.g., Sanibel Island) than they can charge in normal residential areas due to the degree of price insensitivity of vacationing shoppers.
- **Markdown Management**, where retailers integrated POS historical sales data for seasonal or short-lived fashion products to intelligently reduce product prices based on current inventory data and product demand trends to optimize the product or merchandise markdown management process. For example, grocery, mass merchandiser, and drug chain retail outlets used the POS data

with advanced analytics to decide when and how much to mark down Easter, Christmas, Valentine's Day, and other holiday-specific items. And mass merchandisers and department stores used the POS data with advanced analytics to decide when and how much to mark down seasonal items such as swimsuits, parkas, winter boots, and fashion items.

- **Customer Loyalty Programs** were to me the biggest innovation. Retailers suddenly had the opportunity to introduce customer loyalty cards that could be scanned at the time of product purchase in exchange for product discounts and reward programs. Just check your billfold or purse to see how many of these programs you personally belong to. (For me that would include Starbucks, Safeway, Walgreens, Sports Authority, and Foot Locker, just to name a few.) This allowed retailers to tie specific product and market basket purchases to the demographics of their individual shoppers. The potential profiling, targeting, and segmentation possibilities were almost endless, and provided a potentially rich source of insights that retailers could use to better market, sell and service to their most important customers.

Figure 2-1 summarizes the key takeaways with respect to how point-of-sale scanner data drove the CPG-Retail industry transformation.

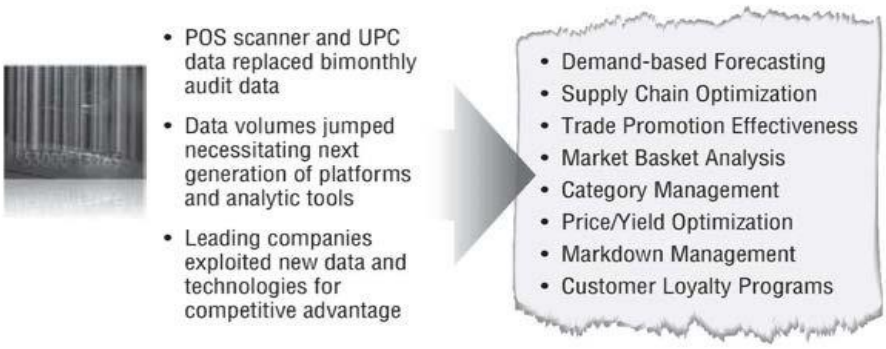


Figure 2-1: Big Data History Lesson: 1980s CPG and retail industries transitioned from bimonthly audit data to POS scanner data

The combination of new data sources and technology innovations also led to new data monetization opportunities (the Data Monetization phase of our Big Data Business Model Maturity Index), such as Walmart's Retail Link that provided detailed product sales information to Walmart's CPG manufacturing and distribution partners. The creation of a platform or ecosystem from which partners and other

value-added developers can deliver new services, capabilities and applications is the start of moving into the Business Metamorphosis phase discussed in Chapter 1.

Ultimately, this more detailed, high-velocity data changed the balance of power in the CPG-Retail industry. Prior to the advent of POS scanner data, CPG manufacturers leveraged their superior knowledge of their customers' buying behaviors (painstakingly gathered through countless focus groups, surveys, and primary research) to dictate sales and payment terms to the retailers. However, courtesy of the POS scanner data and resulting customer and product insights, the retailers suddenly knew more about their customers' buying behaviors, price and promotional sensitivities, and product and market basket preferences. Retailers were able to leverage these superior customer and product insights to dictate product pricing, promotional, and delivery terms to the CPG manufacturers.

Lessons Learned and Applicability to Today's Big Data Movement

The introduction of retail scanner POS systems created new sources of data that required new technologies to manage the data, and new analytic software to analyze the data. But the real competitive advantages came from organizations that exploited the new sources of data and new technology innovations to derive—or drive—new sources of business differentiation, competitive advantage, and monetization.

How does the POS scanner data history lesson apply to the big data movement today? First, new massive volumes of high-velocity structured and unstructured data—both inside and outside of the organization—are breaking traditional data management tools and platforms, and data and analytic modeling techniques. Data sources such as web logs, social media posts, doctor's notes, service comments, research papers, and machine and sensor-generated data are creating data volumes that have some leading organizations already working with petabytes of data, and planning for the inevitable introduction of zettabytes of data. Traditional data management and data warehousing platforms were never designed for the volume, velocity, or complexity of these types of data sources.

Next, new tools must be developed to exploit this tsunami of new data sources. Digital media companies such as Google, Yahoo!, and Facebook—companies whose primary value proposition is built around managing huge data volumes and consequently monetizing that data—have had to develop new technologies to manage and analyze this data, creating technologies such as Hadoop, MapReduce, Pig, Hive, and HBase.

Ultimately, though, the winners will be those organizations that exploit the new data sources, coupled with advancements in data management and advanced analytic technologies, to upgrade or enrich existing business processes or create new business applications that provide unique sources of competitive advantage and business differentiation. Much like how Procter & Gamble (with Category Management), Walmart (with Supply Chain Optimization), and Tesco (with Customer Loyalty Programs) gained competitive advantage from new data sources and new technology innovations, companies today should be focused on determining where data and technology innovation can rewire their existing value creation processes to create new value for their customers, and uncover new sources of revenue and profits for their organizations.

Summary

This chapter covered the history lesson from the late 1980s, where retail POS scanner data created an earlier "big data" revolution. POS scanner data volumes quickly jumped from megabytes to gigabytes and ultimately to terabytes of data, replacing the bimonthly store audit data that had previously been used to make marketing, promotional, product, pricing, and placement decisions.

You reviewed how the volume, diversity, and velocity of this POS data broke existing data management and analytical technologies. EIS analytic software that ran on mainframes could not handle the volume of data, which gave birth to new data processing technologies such as specialized data management platforms (Red Brick, Teradata, Britton Lee, Sybase IQ) and new analytic software packages (Brio, Cognos, Microstrategy, Business Objects).

Finally, the chapter covered how the ultimate winners were those companies who were able to create new analytics-driven business applications, such as category management and demand-based forecasting. Suddenly, retailers with immediate access to POS scanner data coupled with customer loyalty data knew more about customer shopping behaviors and product preferences that they used to change the industry balance of power and dictate terms to CPG manufacturers with respect to pricing, packaging, promotion, and in-store product placement.

3 Business Impact of Big Data

Organizations are starting to realize that big data is more about business transformation than IT transformation. Big data is allowing companies to answer questions they could not previously answer, and make more timely decisions at a finer level of fidelity than before, yielding new insights that can deliver business differentiation and new operational efficiencies. Let's take a look at an example of how big data is transforming how we look at business.

For decades, leading organizations have been exploiting new data sources, plus new technologies, for business differentiation and competitive advantage. And for the most part, the questions that the business users are trying to ask, and answer, with these data sources and new technologies really haven't changed:

- Who are my most valuable customers?
- What are my most important products?
- What are my most successful campaigns?
- What are my best performing channels?
- What are my most effective employees?

The more I thought about these "simple" questions, the more I realized just how "not simple" these questions really were. Because of the new insights available from new big data sources, companies are able to take these types of "simple" questions to the next level of sophistication and understanding.

Let's look at the most valuable customer question. When you ask who your most valuable customers are, do you mean largest by revenue (which is how many companies today still define their most valuable customers)? Or do you mean the most profitable customers, contemplating more aspects of the customer engagement including marketing and sales costs, cost to service, returns, and payment history (which is how some of the more advanced companies think today)? Or by adding

social media into the mix, do you now mean your most influential customers and the financial value associated with their circle of friends?

Companies are learning that their most profitable customers may not actually be their most valuable customers because of the net influencer or advocacy effect. Advocates can have significant influence and persuasive effect on a larger community of customers, and the profitability of the "baskets" associated with that community of customers. Same with the most important product question, which retailers have understood for quite a while (think loss leaders like milk that drive store traffic even though they don't drive much in the form of profits), and consumer goods manufacturers understand as well (think category strategies and the use of flanking products to protect their premium-priced core products).

Those nebulous and hard-to-define words, like valuable, important, and successful, allow the business users to move beyond just financial measures and to consider the entirety of the contributions those customers, products, and campaigns make to the business. It is the basis for a more engaging business discussion about what data sources could be critical in defining "valuable" and what analytic models could be used to quantify "valuable." It's the basis for a wonderful conversation that you can have with your business users about defining those valuable, important, and successful words in light of what big data and advanced analytics can bring to the table.

Big Data Impacts: The Questions Business Users Can Answer

Big data has changed the nuances for defining and quantifying terms such as valuable, important, and successful. It is these nuances that fuel the insights that are the source of competitive advantage and business differentiation. New big data sources, plus new advanced analytic capabilities, enable higher fidelity answers to these questions, and provide a more complete understanding of your customers, products, and operations that can drive business impact across various business functions, such as:

- Merchandising to identify which marketing promotions and campaigns are the most effective in driving store or site traffic and sales.
- Marketing to optimize prices for perishable goods such as groceries, airline seats, and fashion merchandise.
- Sales to optimize the allocation of scarce sales resources against the best sales opportunities and most important or highest potential accounts.
- Procurement to identify which suppliers are most cost-effective in delivering high-quality products in a predictable and timely manner.

- Manufacturing to flag machine performance a
  be indicators of manufacturing, processing, c
- Human Resources to identify the characterist
  successful and effective employees.

Managing Using the Right Metrics

Since baseball is one of my loves in life, and in honor of the enlightening book, *Moneyball: The Art of Winning an Unfair Game,* by Michael Lewis (Norton, 2004), I thought it was only appropriate to discuss how the pursuit and identification of the right metrics has not only changed how the game of baseball is managed, but has the same potential impact on how you manage your business.

In 2004, Lewis wrote the book *Moneyball*, which chronicled how the Oakland A's and Billy Beane, their general manager, were using new data and metrics in order to determine the value of any particular player. The A's were unique at that time in the use of *sabermetrics*, which is the application of statistical analysis to baseball data in order to evaluate and compare the performance of individual players. The results were that the A's had a demonstrable competitive advantage in determining how much to pay any particular player playing any specific position, especially in the costly era of free agency.

As a result, the A's enjoyed a significant cost advantage in what they were paying for wins versus a team like the Yankees. The comparison is shown in Figure 3-1.

2005
2004
2003
2002
2001
2000

Sa ar es ($M)		W ns		Cost per W n ($M)		
A's	Yankees	A's	Yankees	A's	Yankees	A's % of Yankees
$55.4	$208.3	88	95	$0.63	$2.19	29%
$59.4	$184.2	91	101	$0.65	$1.82	36%
$50.3	$152.7	96	101	$0.52	$1.51	35%
$40.0	$125.9	103	103	$0.39	$1.22	32%
$33.8	$112.3	102	95	$0.33	$1.18	28%
$32.1	$88.1	91	87	$0.35	$1.01	35%
$271.0	$871.5	571	582	$0.47	$1.50	32%

Yankee Batting KPIs:
- Batting average
- RBIs
- Fielding percentage
- Steals

A's Batting KPIs:
- On-base percentage
- Slugging percentage

Figure 3-1: Payroll cost per win: Athletics versus Yankees

Unfortunately for Billy Beane and the Oakland A's, other teams (most notably the Boston Red Sox) copied this model and reduced the competitive advantage that the A's briefly enjoyed. But that's the nature of a competitive business isn't it, whether it's in sports, retail, banking, entertainment, telecommunications, or healthcare.

So how does one survive in a world where competitive advantage via analytics can be so short-lived? By constantly innovating, thinking differently, and looking at new sources of data and analytic tools to bring to light those significant, material, and actionable insights that can differentiate your business from that of your competitors.

One of the challenges with metrics is that eventually folks learn how to game the metrics for their own advantage. Sticking with our baseball scenario, let's take the Fielding Percentage metric as an example. The Fielding Percentage metric is calculated as the total number of plays (chances minus errors) divided by the number of total chances. Some players have learned that one of the ways to improve their Fielding Percentage is to stop trying to field balls that are outside of their fielding comfort zone. If you don't try hard for the ball, there can't be an error assessed. While that might be good for the individual's performance numbers, it is obviously less than ideal for the team who wants all of their players trying to make plays in the field. Let's see how that works.

Let's say that an outfielder has 1,000 fielding chances, and makes 20 errors out of those 1,000 fielding chances for a Fielding Percentage of 98 percent (see Figure 3-2). Now, if the fielder doesn't try to field the 100 hardest opportunities (resulting in only 900 Fielding Chances), he will likely cut down significantly on the number of errors (let's say, eliminating 10 errors) resulting in an increased Fielding Percentage of 98.9 percent.

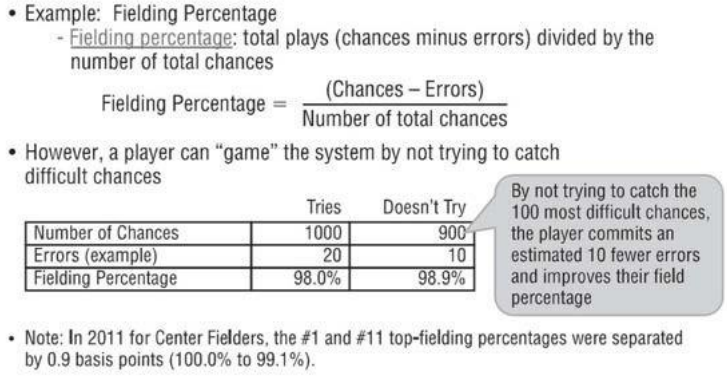


Figure 3-2: Picking the wrong metrics can incentivize the wrong behaviors

While the 0.9 basis-point difference (98.9 minus 98.0) between the two efforts may not seem significant, suffice it to say that the difference between the #1 center fielder in Major League Baseball in 2011 and the #11 center fielder was only 0.9 basis points. The difference probably means millions of dollars to their playing contract.

So the bottom line is that some players have figured out that they will perform better by only trying to field those opportunities within their comfort zone. Not the sort of behavior that leads to very many World Series appearances.

So how does the world of big data change this measure? Baseball stadiums have installed video cameras throughout the stadium to get a better idea as to actual game dynamics. One of the benefits of these cameras is a new set of metrics that are better predictors of players' performance.

For example, video cameras now can measure how much many feet a particular fielder can cover within a certain period of time in fielding their position. Ultimately, this will lead to the creation of an Effective Fielding Range metric which measures how much of the playing field the fielder can cover, and how effectively they cover the playing field (see Figure 3-3). This metric will allow baseball management to value players differently because Effective Fielding Range is a much better predictor of fielding performance than the traditional Fielding Percentage.

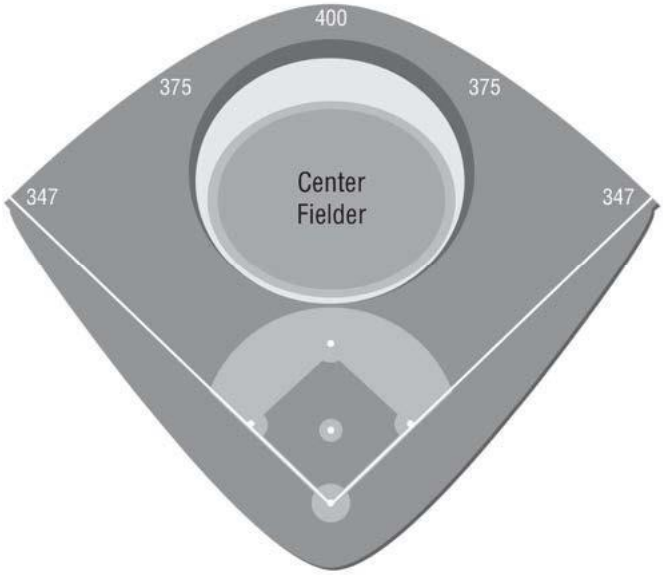


Figure 3-3: Big data hits baseball

As illustrated in the figure, the Center Fielder is very efficient in covering the outfield going left, right, or forward (indicated by the green coverage area), but is less efficient going backwards (indicated by the yellow and red coverage areas).

Much like the world of baseball, organizations must be constantly vigilant in search of metrics that are better predictors of business performance. The new data sources and analytic capabilities enabled by big data hold huge potential to be the first mover in uncovering those significant, measurable, and actionable insights that can lead to competitive advantage—on the baseball field or in the corporate battlefields.

Data Monetization Opportunities

Data monetization is certainly the holy grail of the big data discussion: How do I leverage my vast wealth of customer, product, and operational insights to provide new revenue-generating products and services, enhance product performance and the product experience, and create a more compelling and "sticky" customer relationship?

But how does one even start thinking about this data monetization discussion? Let me take a data monetization example from the digital media world and present a process that other industries can use to uncover and capitalize on potential data monetization opportunities.

Digital Media Data Monetization Example

Digital media companies like Yahoo!, Google, Facebook, and Twitter have worked to master the data monetization process. They must because their entire business model is built on monetizing data. These companies work with bytes to create services, unlike most other companies who work with atoms to build physical products like shoes, tractors, houses, and burrito bowls with double chicken and guacamole.

So what process do these digital media companies go through to identify how to monetize their data assets? The data monetization process starts with two key understandings:

1. Who are my target customers (targeted personas) and what business solutions do they need for which they are willing to pay?
2. What data assets do I have (or could I have)?

Once you have a solid understanding of these two questions, then you are in a position to start the data monetization process.

Digital Media Data Assets and Understanding Target Users

First, digital media companies need to identify and really (and I mean really!) understand their target customers—that is, who is making the million dollar marketing and campaign decisions, and what information and insights do they need to make those decisions? Digital media companies target the following three customers or personas: Media Planners and Buyers, Campaign Managers, and Digital Media Executives. These digital media decision-makers buy the following "solutions":

- Audiences, such as soccer moms, country squires, gray power, and weekend warriors
- Inventory (like sports, finance, news, and entertainment) available on certain days and times of days
- Results or measures, such as Cost per Thousands (CPM) of impressions, Cost Per Acquisition (CPA), product sales, or conversions (where conversions could include getting a visitor to share their e-mail address, request a quote, or schedule a reservation)

For each of these targeted personas, the digital media company needs to understand what questions they are trying to answer, what decisions they are trying to make, under what circumstances they are making these decisions, and within what sort of environment or user experience they are typically working when they have to answer their questions and make their decisions.

Next, digital media companies assess the breadth, depth, and quality of their data assets, including:

- Visitors and their associated demographic, psycho-demographic, and behavioral insights
- Properties and the type of content and advertising real estate (e.g., full banner, pop-under, skyscraper, leaderboard, half-page) that is provided on properties (like Yahoo! Finance, Yahoo! Sports, or Yahoo! Entertainment)
- Activities that visitors perform on those properties (for example, they viewed a display impression, moused over a display ad, clicked a display ad, entered a keyword search) including how often, how recent, and in what sequence

This data assessment process should also include what additional data could be captured through data acquisition, as well as through more robust instrumentation and experimentation techniques.

Data Monetization Transformations and Enrichments

The key challenge is then to transform, augment, enrich, and repackage the data assets into the solutions that the target digital media customers want to buy. For example, digital media companies instrument or set up their sites and tag their visitors (via cookies) to capture visitors' web site and search activities in order to determine or ascertain additional visitor insights, including:

- Geographic information such as ZIP code, city, state, and country
- Demographic information such as gender, age, income, social class, religion, race, and family lifecycle
- Psycho-demographic information such as lifestyle, personality, and values
- Behavioral attributes such as consumption behaviors, lifestyles, patterns of buying and using, patterns of spending money and time, and similar factors
- Product categories of interest (Schmarzo likes Chipotle, Starbucks, the Cubs and the Giants, and all things basketball)
- Social influences such as interests, passions, associations, and affiliations

With this information in hand, the digital media company needs the data processing capacity and advanced analytical skills to profile, segment, and package those visitors into the audiences that advertisers and advertising agencies want to buy.

This data transformation, augmentation, and enrichment process is then repeated in converting properties into inventory, visitor activities into digital treatments, and campaigns into results such as sales and conversions (see Table 3-1).

NOTE The table below has been organized with step 1 at the far right, as it represents the end solutions that we are trying to deliver. Step 2 is on the far left as it represents the key data assets, which will go through step 3 to be transformed and enriched into our targeted solutions.

Table 3-1: Data Monetization Example—Digital Media Company

Step 2: Assess Data Assets	Step 3: Identifying Transformation, Enrichment, and Analytic Requirements	Step 1: Define Digital Advertiser Solutions
Visitor	Demographics Insights Psycho-Demographics Insights Behavioral Insights Social and Mobile Insights	Audiences What audiences am I reaching? Who is my most engaged audience? What similar audiences could I target?

Step 2: Assess Data Assets	Step 3: Identifying Transformation, Enrichment, and Analytic Requirements	Step 1: Define Digital Advertiser Solutions
Properties (Sites)	Product categories (Sports, Finance) Audiences Premium vs. Remnant	Inventory What inventories are most effective? What product categories are most effective? What other product categories should I use?
Web Activities	Impressions Clicks Keyword Searches Social Posts and Activities Mobile Tracking	Marketing Treatments What marketing treatments are most effective? What are minimum frequency/recency levels? What is the optimal sequencing of treatments?
Campaigns	Instrumentation Analytics (Attribution, Audience Insights, Benchmarking) Optimizations and Predictions Recommendations User Experience	Sales/Conversions/CPM Will I achieve campaign objectives (predict)? What will be the impact if I re-allocate spending? What recommended changes will improve performance? How can I optimize inflight cross-media spending?

Based on this digital media example, here are the steps that your company needs to go through in order to better understand how to monetize your data assets.

1. Identify your target customers and their desired solutions (solution capabilities and required insights) in order to optimize their performance and simplify their jobs. Identify and profile the target business customers or personas for those solutions, and internalize how those customers will use that solution within their existing work environment. Quantify the business value of those solutions, and document the business questions the users need to answer and business decisions the business users need to make as part of the desired solution.
2. Inventory and assess your data assets; that is, identify the most important and valuable "nouns" of your business. Understand what additional data could be

gathered to enrich your data asset base via data acquisition and a more robust instrumentation and experimentation strategy.

3. Understand the aggregation, transformation, cleansing, alignment, data enrichment, and analytic processes necessary to transform your data assets into business solutions. Document what insights and analytics you can package that meets your customers' needs for a solution that optimizes business performance and simplifies their jobs. Identify the data enrichment and analytic processes necessary to transform data into actionable insights and understand how those insights manifest themselves within the customers' user experience.

There are numerous opportunities for organizations to improve product performance, enhance product design and development, preempt product failure, and enhance the overall user (shopper, driver, patient, subscriber, member) experience. More and more, the data and the resulting insights teased out of the data will become a key component, and potentially a differentiator, in the products and services that companies provide.

Summary

This chapter covered how asking the right questions is one of the key starting points in your big data journey. You learned how big data has changed the nuances for defining and quantifying terms, such as *valuable, important*, and *successful*, and saw some examples of how big data is helping various business functions ask the right questions at a finer level of fidelity.

Then I reviewed how big data is enabling organizations to identify new measures and metrics that are better predictors of business performance. I discussed the impact that the book *Moneyball* and the world of sabermetrics has had on helping baseball teams, particularly the Oakland A's, exploit a superior understanding of the "right" metrics to optimize baseball success on the baseball field. I also provided an example of how big data is taking the world of baseball analytics to the next level of predictive excellence with new insights about baseball player performance that are better predictors of in-game success.

The chapter concluded with a discussion on how you can monetize your data assets. I reviewed how your organization can leverage data assets to deliver new revenue opportunities and a more compelling, differentiated business relationship through superior customer, product, and market insights. I used the world of digital media marketing as an example and provided a "How To" framework to help your

Unit -2

Chapter 2
Data Analytics Lifecycle

Key Concepts

1. *Discovery*
2. *Data preparation*
3. *Model planning*
4. *Model execution*
5. *Communicate results*
6. *Operationalize*

Data science projects differ from most traditional Business Intelligence projects and many data analysis projects in that data science projects are more exploratory in nature. For this reason, it is critical to have a process to govern them and ensure that the participants are thorough and rigorous in their approach, yet not so rigid that the process impedes exploration.

Many problems that appear huge and daunting at first can be broken down into smaller pieces or actionable phases that can be more easily addressed. Having a good process ensures a comprehensive and repeatable method for conducting analysis. In addition, it helps focus time and energy early in the process to get a clear grasp of the business problem to be solved.

A common mistake made in data science projects is rushing into data collection and analysis, which precludes spending sufficient time to plan and scope the amount of work involved, understanding requirements, or even framing the business problem properly. Consequently, participants may discover mid-stream that the project sponsors are actually trying to achieve an objective that may not match the available data, or they are attempting to address an interest that differs from what has been explicitly communicated. When this happens, the project may need to revert to the initial phases of the process for a proper discovery phase, or the project may be canceled.

Creating and documenting a process helps demonstrate rigor, which provides additional credibility to the project when the data science team shares its findings. A well-defined process also offers a common framework for others to adopt, so the methods and analysis can be repeated in the future or as new members join a team.

2.1 Data Analytics Lifecycle Overview

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The lifecycle has six phases, and project work can occur in several phases at once. For most phases in the lifecycle, the movement can be either forward or backward. This iterative depiction of the lifecycle is intended to more closely portray a real project, in which aspects of the project move forward and may return to earlier stages as new information is uncovered and team members learn more about various stages of the project. This enables participants to move iteratively through the process and drive toward operationalizing the project work.

2.1.1 Key Roles for a Successful Analytics Project

In recent years, substantial attention has been placed on the emerging role of the data scientist. In October 2012, Harvard Business Review featured an article titled "Data Scientist: The Sexiest Job of the 21st Century" [1], in which experts DJ Patil and Tom Davenport described the new role and how to find and hire data scientists. More and more conferences are held annually focusing on innovation in the areas of Data Science and topics dealing with Big Data. Despite this strong focus on the emerging role of the data scientist specifically, there are actually seven key roles that need to be fulfilled for a high-functioning data science team to execute analytic projects successfully.

Figure 2.1 depicts the various roles and key stakeholders of an analytics project. Each plays a critical part in a successful analytics project. Although seven roles are listed, fewer or more people can accomplish the work depending on the scope of the project, the organizational structure, and the skills of the participants. For example, on a small, versatile team, these seven roles may be fulfilled by only 3 people, but a very large project may require 20 or more people. The seven roles follow.

- **Business User:** Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.
- **Project Sponsor:** Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.
- **Project Manager:** Ensures that key milestones and objectives are met on time and at the expected quality.
- **Business Intelligence Analyst:** Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective. Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

- **Database Administrator (DBA):** Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.
- **Data Engineer:** Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox, which was discussed in Chapter 1, "Introduction to Big Data Analytics." Whereas the DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.
- **Data Scientist:** Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project.

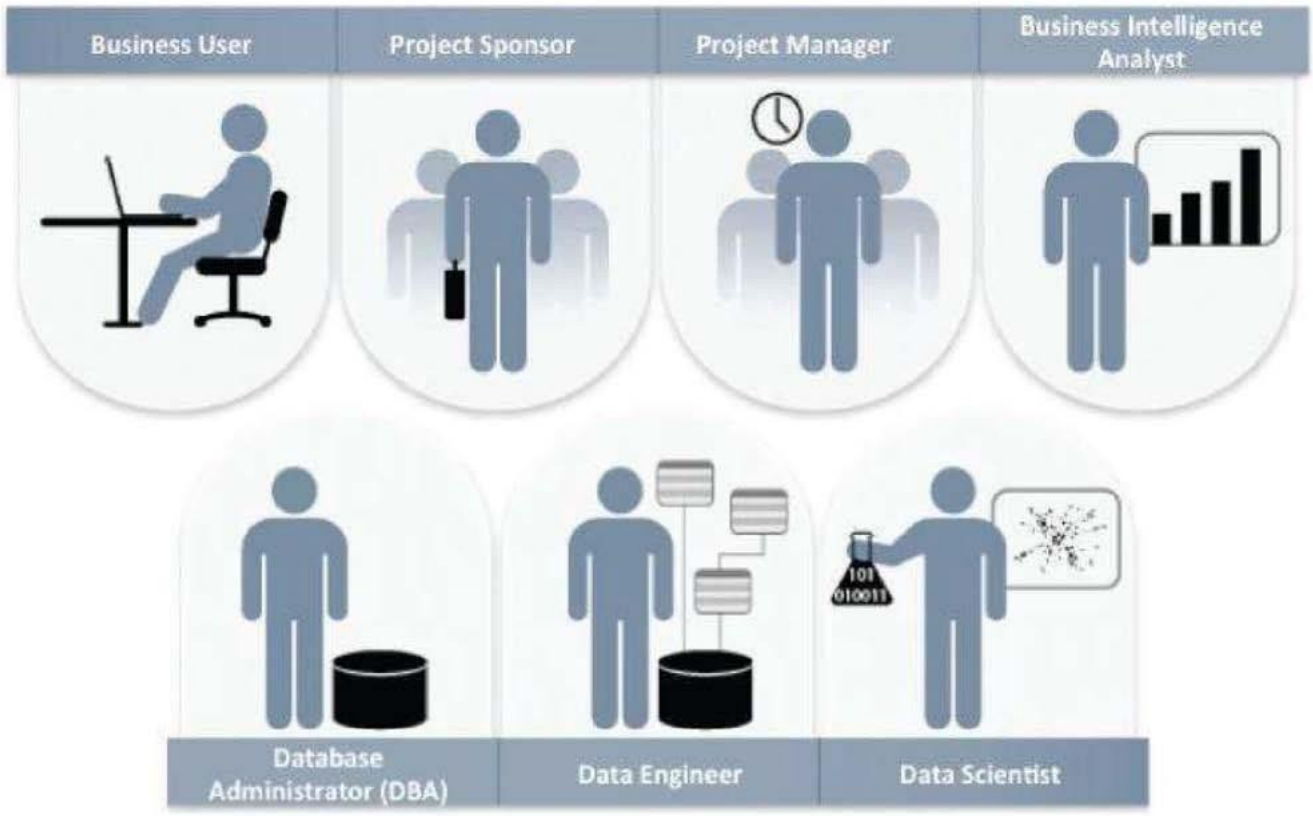


Figure 2.1 Key roles for a successful analytics project

Although most of these roles are not new, the last two roles—data engineer and data scientist—have become popular and in high demand [2] as interest in Big Data has grown.

2.1.2 Background and Overview of Data Analytics Lifecycle

The Data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion. The lifecycle draws from established methods in the realm of data analytics and decision science. This synthesis was developed after gathering input from data scientists and consulting established approaches that provided input on pieces of the

process. Several of the processes that were consulted include these:

- **Scientific method** [3], in use for centuries, still provides a solid framework for thinking about and deconstructing problems into their principal parts. One of the most valuable ideas of the scientific method relates to forming hypotheses and finding ways to test ideas.
- **CRISP-DM** [4] provides useful input on ways to frame analytics problems and is a popular approach for data mining.
- Tom Davenport's **DELTA** framework [5]: The DELTA framework offers an approach for data analytics projects, including the context of the organization's skills, datasets, and leadership engagement.
- Doug Hubbard's **Applied Information Economics (AIE)** approach [6]: AIE provides a framework for measuring intangibles and provides guidance on developing decision models, calibrating expert estimates, and deriving the expected value of information.
- "**MAD Skills**" by Cohen et al. [7] offers input for several of the techniques mentioned in Phases 2–4 that focus on model planning, execution, and key findings.

Figure 2.2 presents an overview of the Data Analytics Lifecycle that includes six phases. Teams commonly learn new things in a phase that cause them to go back and refine the work done in prior phases based on new insights and information that have been uncovered. For this reason, Figure 2.2 is shown as a cycle. The circular arrows convey iterative movement between phases until the team members have sufficient information to move to the next phase. The callouts include sample questions to ask to help guide whether each of the team members has enough information and has made enough progress to move to the next phase of the process. Note that these phases do not represent formal stage gates; rather, they serve as criteria to help test whether it makes sense to stay in the current phase or move to the next.

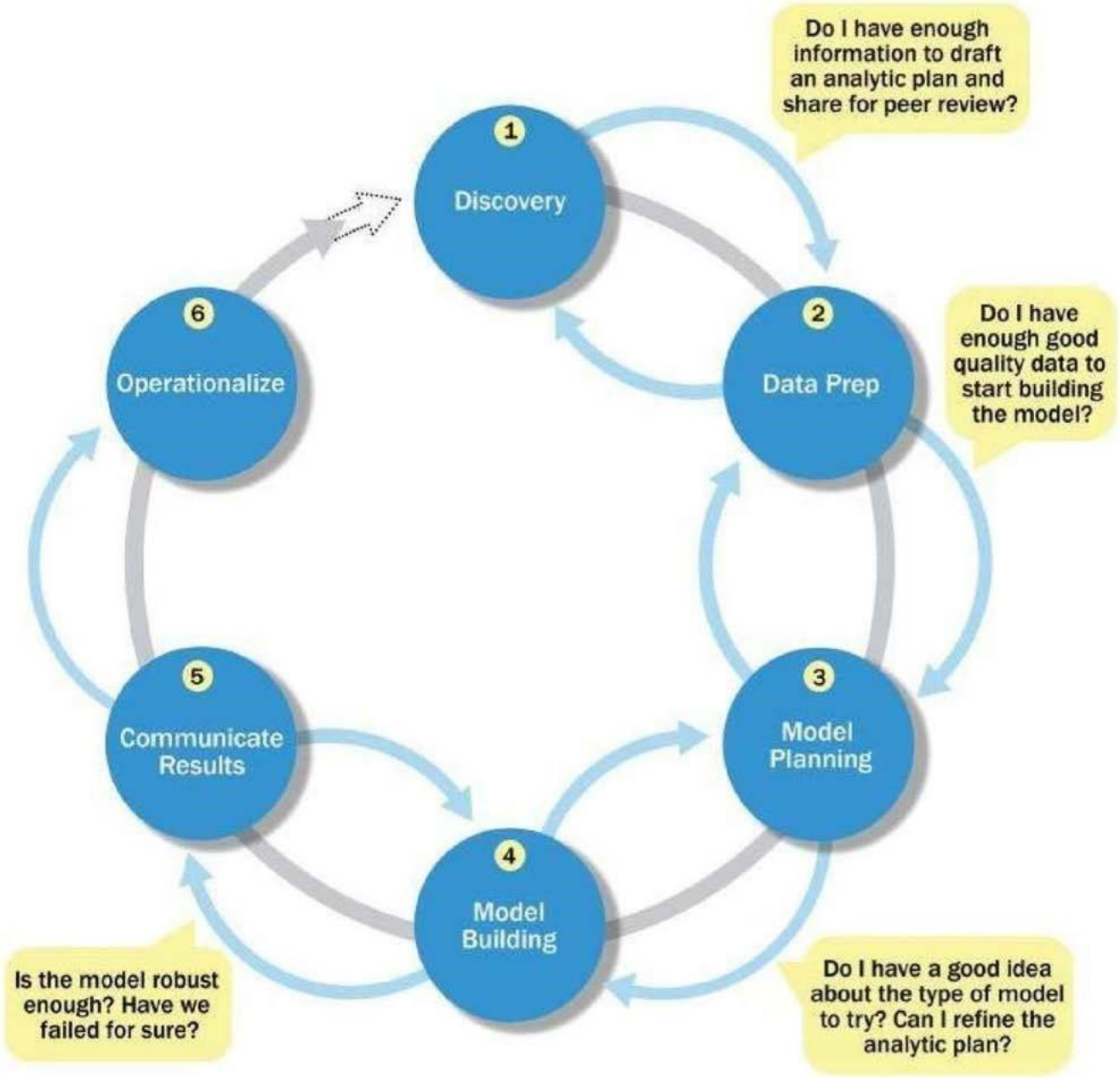


Figure 2.2 Overview of Data Analytics Lifecycle

Here is a brief overview of the main phases of the Data Analytics Lifecycle:

- **Phase 1—Discovery:** In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.
- **Phase 2—Data preparation:** Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are

n

sometimes abbreviated as ETLT. Data should be transformed i the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data (Section 2.3.4).

- **Phase 3—Model planning:** Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.
- **Phase 4—Model building:** In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).
- **Phase 5—Communicate results:** In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.
- **Phase 6—Operationalize:** In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

Once team members have run models and produced findings, it is critical to frame these results in a way that is tailored to the audience that engaged the team. Moreover, it is critical to frame the results of the work in a manner that demonstrates clear value. If the team performs a technically accurate analysis but fails to translate the results into a language that resonates with the audience, people will not see the value, and much of the time and effort on the project will have been wasted.

The rest of the chapter is organized as follows. Sections 2.2–2.7 discuss in detail how each of the six phases works, and Section 2.8 shows a case study of incorporating the Data Analytics Lifecycle in a real-world data science project.

ɜ n

2.2 Phase 1: Discovery

The first phase of the Data Analytics Lifecycle involves discovery (Figure 2.3). In this phase, the data science team must learn and investigate the problem, develop context and understanding, and learn about the data sources needed and available for the project. In addition, the team formulates initial hypotheses that can later be tested with data.

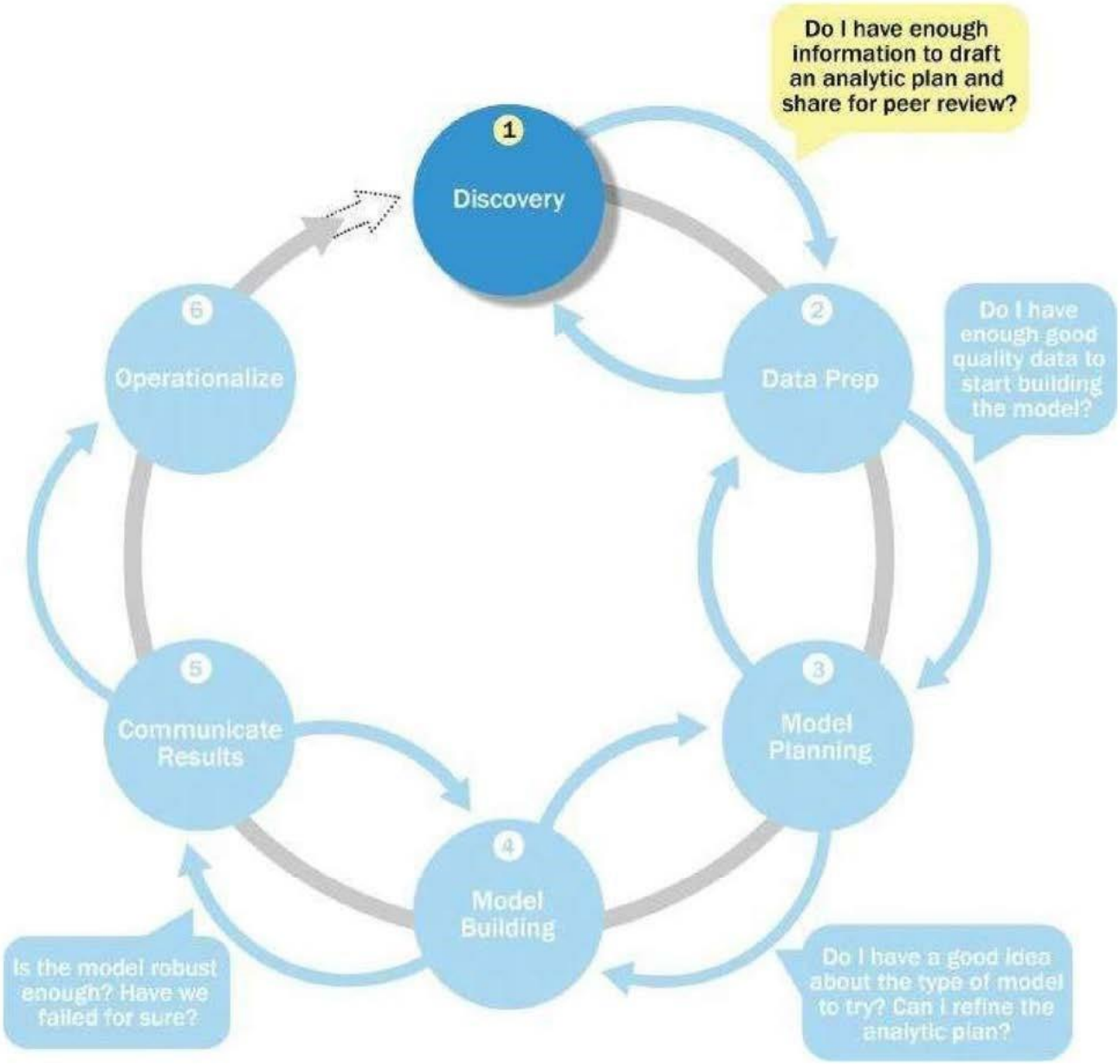


Figure 2.3 Discovery phase

2.2.1 Learning the Business Domain

Understanding the domain area of the problem is essential. In many cases, data scientists will have deep computational and quantitative knowledge that can be broadly applied across many disciplines. An example of this role would be someone with an advanced degree in applied mathematics or statistics.

These data scientists have deep knowledge of the methods, techniques, and ways for applying heuristics to a variety of business and conceptual problems. Others in this area

may have deep knowledge of a domain area, coupled with qua titative expertise. An example of this would be someone with a Ph.D. in life sciences. This person would have deep knowledge of a field of study, such as oceanography, biology, or genetics, with some depth of quantitative knowledge.

At this early stage in the process, the team needs to determine how much business or domain knowledge the data scientist needs to develop models in Phases 3 and 4. The earlier the team can make this assessment the better, because the decision helps dictate the resources needed for the project team and ensures the team has the right balance of domain knowledge and technical expertise.

2.2.2 Resources

As part of the discovery phase, the team needs to assess the resources available to support the project. In this context, resources include technology, tools, systems, data, and people.

During this scoping, consider the available tools and technology the team will be using and the types of systems needed for later phases to operationalize the models. In addition, try to evaluate the level of analytical sophistication within the organization and gaps that may exist related to tools, technology, and skills. For instance, for the model being developed to have longevity in an organization, consider what types of skills and roles will be required that may not exist today. For the project to have long-term success, what types of skills and roles will be needed for the recipients of the model being developed? Does the requisite level of expertise exist within the organization today, or will it need to be cultivated? Answering these questions will influence the techniques the team selects and the kind of implementation the team chooses to pursue in subsequent phases of the Data Analytics Lifecycle.

In addition to the skills and computing resources, it is advisable to take inventory of the types of data available to the team for the project. Consider if the data available is sufficient to support the project's goals. The team will need to determine whether it must collect additional data, purchase it from outside sources, or transform existing data. Often, projects are started looking only at the data available. When the data is less than hoped for, the size and scope of the project is reduced to work within the constraints of the existing data.

An alternative approach is to consider the long-term goals of this kind of project, without being constrained by the current data. The team can then consider what data is needed to reach the long-term goals and which pieces of this multistep journey can be achieved today with the existing data. Considering longer-term goals along with short-term goals enables teams to pursue more ambitious projects and treat a project as the first step of a more strategic initiative, rather than as a standalone initiative. It is critical to view projects as part of a longer-term journey, especially if executing projects in an organization that is new to Data Science and may not have embarked on the optimum datasets to support robust analyses up to this point.

Ensure the project team has the right mix of domain experts, customers, analytic talent, and project management to be effective. In addition, evaluate how much time is needed and if the team has the right breadth and depth of skills.

3

After taking inventory of the tools, technology, data, and people, consider if the team has sufficient resources to succeed on this project, or if additional resources are needed. Negotiating for resources at the outset of the project, while scoping the goals, objectives, and feasibility, is generally more useful than later in the process and ensures sufficient time to execute it properly. Project managers and key stakeholders have better success negotiating for the right resources at this stage rather than later once the project is underway.

2.2.3 Framing the Problem

Framing the problem well is critical to the success of the project. ***Framing*** is the process of stating the analytics problem to be solved. At this point, it is a best practice to write down the problem statement and share it with the key stakeholders. Each team member may hear slightly different things related to the needs and the problem and have somewhat different ideas of possible solutions. For these reasons, it is crucial to state the analytics problem, as well as why and to whom it is important. Essentially, the team needs to clearly articulate the current situation and its main challenges.

As part of this activity, it is important to identify the main objectives of the project, identify what needs to be achieved in business terms, and identify what needs to be done to meet the needs. Additionally, consider the objectives and the success criteria for the project. What is the team attempting to achieve by doing the project, and what will be considered "good enough" as an outcome of the project? This is critical to document and share with the project team and key stakeholders. It is best practice to share the statement of goals and success criteria with the team and confirm alignment with the project sponsor's expectations.

Perhaps equally important is to establish failure criteria. Most people doing projects prefer only to think of the success criteria and what the conditions will look like when the participants are successful. However, this is almost taking a best-case scenario approach, assuming that everything will proceed as planned and the project team will reach its goals. However, no matter how well planned, it is almost impossible to plan for everything that will emerge in a project. The failure criteria will guide the team in understanding when it is best to stop trying or settle for the results that have been gleaned from the data. Many times people will continue to perform analyses past the point when any meaningful insights can be drawn from the data. Establishing criteria for both success and failure helps the participants avoid unproductive effort and remain aligned with the project sponsors

2.2.4 Identifying Key Stakeholders

Another important step is to identify the key stakeholders and their interests in the project. During these discussions, the team can identify the success criteria, key risks, and stakeholders, which should include anyone who will benefit from the project or will be significantly impacted by the project. When interviewing stakeholders, learn about the domain area and any relevant history from similar analytics projects. For example, the team may identify the results each stakeholder wants from the project and the criteria it will use to judge the success of the project.

Keep in mind that the analytics project is being initiated for a reason. It is critical to

articulate the pain point as clearly as possible to address them and be aware of areas to pursue or avoid as the team gets further into the analytical process. Depending on the number of stakeholders and participants, the team may consider outlining the type of activity and participation expected from each stakeholder and participant. This will set clear expectations with the participants and avoid delays later when, for example, the team may feel it needs to wait for approval from someone who views himself as an adviser rather than an approver of the work product.

2.2.5 Interviewing the Analytics Sponsor

The team should plan to collaborate with the stakeholders to clarify and frame the analytics problem. At the outset, project sponsors may have a predetermined solution that may not necessarily realize the desired outcome. In these cases, the team must use its knowledge and expertise to identify the true underlying problem and appropriate solution.

For instance, suppose in the early phase of a project, the team is told to create a recommender system for the business and that the way to do this is by speaking with three people and integrating the product recommender into a legacy corporate system. Although this may be a valid approach, it is important to test the assumptions and develop a clear understanding of the problem. The data science team typically may have a more objective understanding of the problem set than the stakeholders, who may be suggesting solutions to a given problem. Therefore, the team can probe deeper into the context and domain to clearly define the problem and propose possible paths from the problem to a desired outcome. In essence, the data science team can take a more objective approach, as the stakeholders may have developed biases over time, based on their experience. Also, what may have been true in the past may no longer be a valid working assumption. One possible way to circumvent this issue is for the project sponsor to focus on clearly defining the requirements, while the other members of the data science team focus on the methods needed to achieve the goals.

When interviewing the main stakeholders, the team needs to take time to thoroughly interview the project sponsor, who tends to be the one funding the project or providing the high-level requirements. This person understands the problem and usually has an idea of a potential working solution. It is critical to thoroughly understand the sponsor's perspective to guide the team in getting started on the project. Here are some tips for interviewing project sponsors:

- Prepare for the interview; draft questions, and review with colleagues.
- Use open-ended questions; avoid asking leading questions.
- Probe for details and pose follow-up questions.
- Avoid filling every silence in the conversation; give the other person time to think.
- Let the sponsors express their ideas and ask clarifying questions, such as "Why? Is that correct? Is this idea on target? Is there anything else?"
- Use active listening techniques; repeat back what was heard to make sure the team heard it correctly, or reframe what was said.

a s n

- Try to avoid expres ing the team's opinions, which can introduce bias; instead, focus on listening.
- Be mindful of the body language of the interviewers and stakeholders; use eye contact where appropriate, and be attentive.
- Minimize distractions.
- Document what the team heard, and review it with the sponsors.

Following is a brief list of common questions that are helpful to ask during the discovery phase when interviewing the project sponsor. The responses will begin to shape the scope of the project and give the team an idea of the goals and objectives of the project.

- What business problem is the team trying to solve?
- What is the desired outcome of the project?
- What data sources are available?
- What industry issues may impact the analysis?
- What timelines need to be considered?
- Who could provide insight into the project?
- Who has final decision-making authority on the project?
- How will the focus and scope of the problem change if the following dimensions change:
  - **Time:** Analyzing 1 year or 10 years' worth of data?
  - **People:** Assess impact of changes in resources on project timeline.
  - **Risk:** Conservative to aggressive
  - **Resources:** None to unlimited (tools, technology, systems)
  - **Size and attributes of data:** Including internal and external data sources

2.2.6 Developing Initial Hypotheses

Developing a set of IHs is a key facet of the discovery phase. This step involves forming ideas that the team can test with data. Generally, it is best to come up with a few primary hypotheses to test and then be creative about developing several more. These IHs form the basis of the analytical tests the team will use in later phases and serve as the foundation for the findings in Phase 5. Hypothesis testing from a statistical perspective is covered in greater detail in Chapter 3, "Review of Basic Data Analytic Methods Using R."

In this way, the team can compare its answers with the outcome of an experiment or test to generate additional possible solutions to problems. As a result, the team will have a much richer set of observations to choose from and more choices for agreeing upon the most impactful conclusions from a project.

Another part of this process involves gathering and assessing hypotheses from stakeholders and domain experts who may have their own perspective on what the

problem is, what the olution should be, and how to arrive at a solution. These stakeholders would know the domain area well and can offer suggestions on ideas to test as the team formulates hypotheses during this phase. The team will likely collect many ideas that may illuminate the operating assumptions of the stakeholders. These ideas will also give the team opportunities to expand the project scope into adjacent spaces where it makes sense or design experiments in a meaningful way to address the most important interests of the stakeholders. As part of this exercise, it can be useful to obtain and explore some initial data to inform discussions with stakeholders during the hypothesis-forming stage.

2.2.7 Identifying Potential Data Sources

As part of the discovery phase, identify the kinds of data the team will need to solve the problem. Consider the volume, type, and time span of the data needed to test the hypotheses. Ensure that the team can access more than simply aggregated data. In most cases, the team will need the raw data to avoid introducing bias for the downstream analysis. Recalling the characteristics of Big Data from Chapter 1, assess the main characteristics of the data, with regard to its volume, variety, and velocity of change. A thorough diagnosis of the data situation will influence the kinds of tools and techniques to use in Phases 2-4 of the Data Analytics Lifecycle. In addition, performing data exploration in this phase will help the team determine the amount of data needed, such as the amount of historical data to pull from existing systems and the data structure. Develop an idea of the scope of the data needed, and validate that idea with the domain experts on the project.

The team should perform five main activities during this step of the discovery phase:

- **Identify data sources:** Make a list of candidate data sources the team may need to test the initial hypotheses outlined in this phase. Make an inventory of the datasets currently available and those that can be purchased or otherwise acquired for the tests the team wants to perform.
- **Capture aggregate data sources:** This is for previewing the data and providing high-level understanding. It enables the team to gain a quick overview of the data and perform further exploration on specific areas. It also points the team to possible areas of interest within the data.
- **Review the raw data:** Obtain preliminary data from initial data feeds. Begin understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.
- **Evaluate the data structures and tools needed:** The data type and structure dictate which tools the team can use to analyze the data. This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.
- **Scope the sort of data infrastructure needed for this type of problem:** In addition to the tools needed, the data influences the kind of infrastructure that's required, such as disk storage and network capacity.

Unlike many traditional stage-gate processes, in which the team can advance only when

specific criteria are met, the Data Analytics Lifecycle is intended to accommodate more ambiguity. This more closely reflects how data science projects work in real-life situations. For each phase of the process, it is recommended to pass certain checkpoints as a way of gauging whether the team is ready to move to the next phase of the Data Analytics Lifecycle.

The team can move to the next phase when it has enough information to draft an analytics plan and share it for peer review. Although a peer review of the plan may not actually be required by the project, creating the plan is a good test of the team's grasp of the business problem and the team's approach to addressing it. Creating the analytic plan also requires a clear understanding of the domain area, the problem to be solved, and scoping of the data sources to be used. Developing success criteria early in the project clarifies the problem definition and helps the team when it comes time to make choices about the analytical methods being used in later phases.

2.3 Phase 2: Data Preparation

The second phase of the Data Analytics Lifecycle involves data preparation, which includes the steps to explore, preprocess, and condition data prior to modeling and analysis. In this phase, the team needs to create a robust environment in which it can explore the data that is separate from a production environment. Usually, this is done by preparing an analytics sandbox. To get the data into the sandbox, the team needs to perform ETLT, by a combination of extracting, transforming, and loading data into the sandbox. Once the data is in the sandbox, the team needs to learn about the data and become familiar with it. Understanding the data in detail is critical to the success of the project. The team also must decide how to condition and transform data to get it into a format to facilitate subsequent analysis. The team may perform data visualizations to help team members understand the data, including its trends, outliers, and relationships among data variables. Each of these steps of the data preparation phase is discussed throughout this section.

Data preparation tends to be the most labor-intensive step in the analytics lifecycle. In fact, it is common for teams to spend at least 50% of a data science project's time in this critical phase. If the team cannot obtain enough data of sufficient quality, it may be unable to perform the subsequent steps in the lifecycle process.

Figure 2.4 shows an overview of the Data Analytics Lifecycle for Phase 2. The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often. This is because most teams and leaders are anxious to begin analyzing the data, testing hypotheses, and getting answers to some of the questions posed in Phase 1. Many tend to jump into Phase 3 or Phase 4 to begin rapidly developing models and algorithms without spending the time to prepare the data for modeling. Consequently, teams come to realize the data they are working with does not allow them to execute the models they want, and they end up back in Phase 2 anyway.

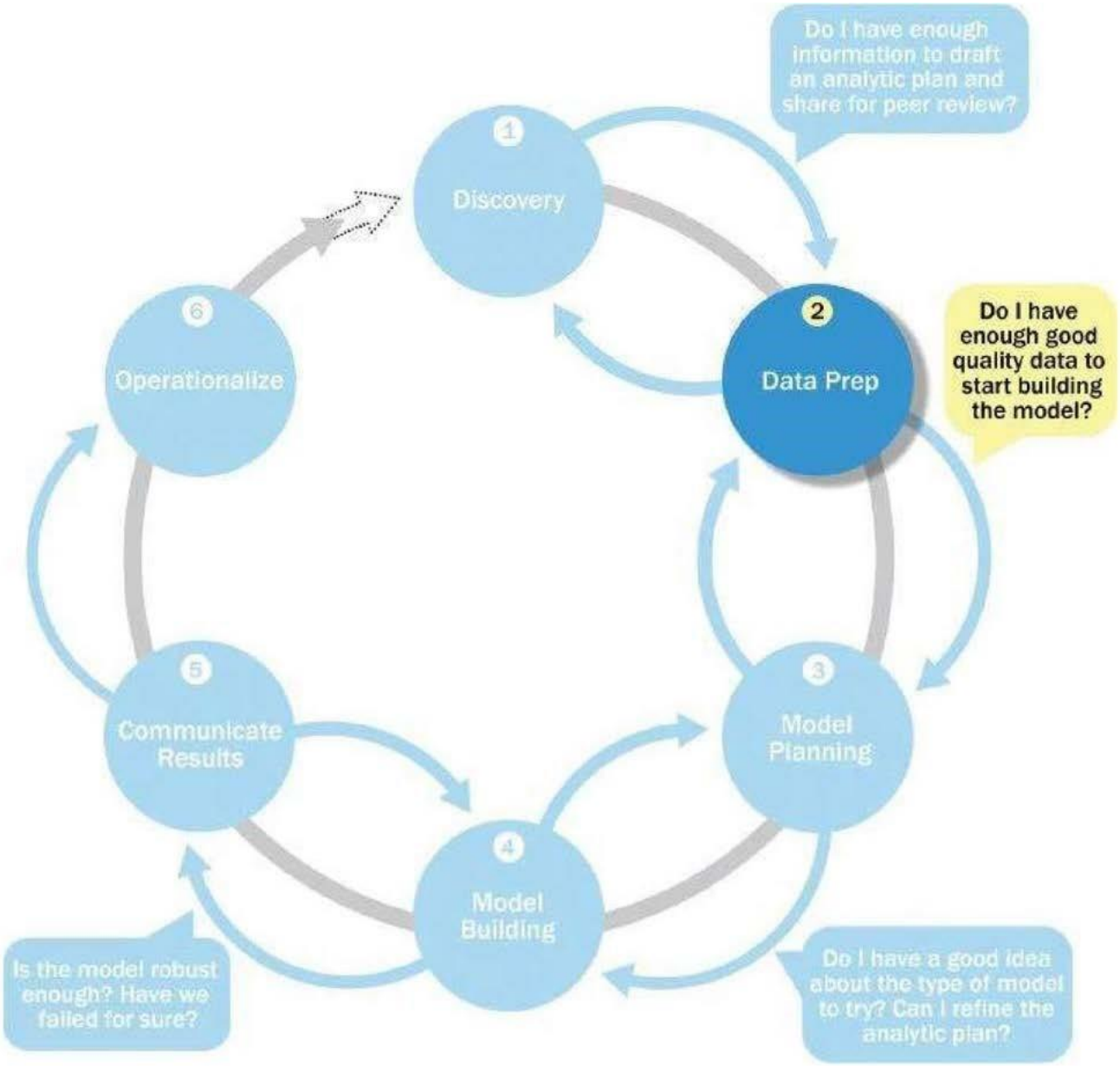


Figure 2.4 Data preparation phase

2.3.1 Preparing the Analytic Sandbox

The first subphase of data preparation requires the team to obtain an analytic sandbox (also commonly referred to as a ***workspace***), in which the team can explore the data without interfering with live production databases. Consider an example in which the team needs to work with a company's financial data. The team should access a copy of the financial data from the analytic sandbox rather than interacting with the production version of the organization's main database, because that will be tightly controlled and needed for financial reporting.

When developing the analytic sandbox, it is a best practice to collect all kinds of data there, as team members need access to high volumes and varieties of data for a Big Data analytics project. This can include everything from summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or web logs, depending on the kind of analysis the team plans to undertake.

This expansive approach for attracting data of all kind differs considerably from he approach advocated by many information technology (IT) organizations. Many IT groups provide access to only a particular subsegment of the data for a specific purpose. Often, the mindset of the IT group is to provide the minimum amount of data required to allow the team to achieve its objectives. Conversely, the data science team wants access to everything. From its perspective, more data is better, as oftentimes data science projects are a mixture of purpose-driven analyses and experimental approaches to test a variety of ideas. In this context, it can be challenging for a data science team if it has to request access to each and every dataset and attribute one at a time. Because of these differing views on data access and use, it is critical for the data science team to collaborate with IT, make clear what it is trying to accomplish, and align goals.

During these discussions, the data science team needs to give IT a justification to develop an analytics sandbox, which is separate from the traditional IT-governed data warehouses within an organization. Successfully and amicably balancing the needs of both the data science team and IT requires a positive working relationship between multiple groups and data owners. The payoff is great. The analytic sandbox enables organizations to undertake more ambitious data science projects and move beyond doing traditional data analysis and Business Intelligence to perform more robust and advanced predictive analytics.

Expect the sandbox to be large. It may contain raw data, aggregated data, and other data types that are less commonly used in organizations. Sandbox size can vary greatly depending on the project. A good rule is to plan for the sandbox to be at least 5–10 times the size of the original datasets, partly because copies of the data may be created that serve as specific tables or data stores for specific kinds of analysis in the project.

Although the concept of an analytics sandbox is relatively new, companies are making progress in this area and are finding ways to offer sandboxes and workspaces where teams can access datasets and work in a way that is acceptable to both the data science teams and the IT groups.

2.3.2 Performing ETLT

As the team looks to begin data transformations, make sure the analytics sandbox has ample bandwidth and reliable network connections to the underlying data sources to enable uninterrupted read and write. In ETL, users perform extract, transform, load processes to extract data from a datastore, perform data transformations, and load the data back into the datastore. However, the analytic sandbox approach differs slightly; it advocates extract, load, and then transform. In this case, the data is extracted in its raw form and loaded into the datastore, where analysts can choose to transform the data into a new state or leave it in its original, raw condition. The reason for this approach is that there is significant value in preserving the raw data and including it in the sandbox before any transformations take place.

For instance, consider an analysis for fraud detection on credit card usage. Many times, outliers in this data population can represent higher-risk transactions that may be indicative of fraudulent credit card activity. Using ETL, these outliers may be inadvertently filtered out or transformed and cleaned before being loaded into the datastore. In this case, the very data that would be needed to evaluate instances of

t

fraudulent activity would be inadvertent y cleansed, preventing the kind of analysis tha a team would want to do.

Following the ELT approach gives the team access to clean data to analyze after the data has been loaded into the database and gives access to the data in its original form for finding hidden nuances in the data. This approach is part of the reason that the analytic sandbox can quickly grow large. The team may want clean data and aggregated data and may need to keep a copy of the original data to compare against or look for hidden patterns that may have existed in the data before the cleaning stage. This process can be summarized as ETLT to reflect the fact that a team may choose to perform ETL in one case and ELT in another.

Depending on the size and number of the data sources, the team may need to consider how to parallelize the movement of the datasets into the sandbox. For this purpose, moving large amounts of data is sometimes referred to as Big ETL. The data movement can be parallelized by technologies such as Hadoop or MapReduce, which will be explained in greater detail in Chapter 10, "Advanced Analytics—Technology and Tools: MapReduce and Hadoop." At this point, keep in mind that these technologies can be used to perform parallel data ingest and introduce a huge number of files or datasets in parallel in a very short period of time. Hadoop can be useful for data loading as well as for data analysis in subsequent phases.

Prior to moving the data into the analytic sandbox, determine the transformations that need to be performed on the data. Part of this phase involves assessing data quality and structuring the datasets properly so they can be used for robust analysis in subsequent phases. In addition, it is important to consider which data the team will have access to and which new data attributes will need to be derived in the data to enable analysis.

As part of the ETLT step, it is advisable to make an inventory of the data and compare the data currently available with datasets the team needs. Performing this sort of gap analysis provides a framework for understanding which datasets the team can take advantage of today and where the team needs to initiate projects for data collection or access to new datasets currently unavailable. A component of this subphase involves extracting data from the available sources and determining data connections for raw data, online transaction processing (OLTP) databases, online analytical processing (OLAP) cubes, or other data feeds.

Application programming interface (API) is an increasingly popular way to access a data source [8]. Many websites and social network applications now provide APIs that offer access to data to support a project or supplement the datasets with which a team is working. For example, connecting to the Twitter API can enable a team to download millions of tweets to perform a project for sentiment analysis on a product, a company, or an idea. Much of the Twitter data is publicly available and can augment other datasets used on the project.

2.3.3 Learning About the Data

A critical aspect of a data science project is to become familiar with the data itself. Spending time to learn the nuances of the datasets provides context to understand what constitutes a reasonable value and expected output versus what is a surprising finding. In

addition, it is important to catalog the data sources that the team has access to and iden ify additional data sources that the team can leverage but perhaps does not have access to today. Some of the activities in this step may overlap with the initial investigation of the datasets that occur in the discovery phase. Doing this activity accomplishes several goals.

- Clarifies the data that the data science team has access to at the start of the project
- Highlights gaps by identifying datasets within an organization that the team may find useful but may not be accessible to the team today. As a consequence, this activity can trigger a project to begin building relationships with the data owners and finding ways to share data in appropriate ways. In addition, this activity may provide an impetus to begin collecting new data that benefits the organization or a specific long-term project.
- Identifies datasets outside the organization that may be useful to obtain, through open APIs, data sharing, or purchasing data to supplement already existing datasets

Table 2.1 demonstrates one way to organize this type of data inventory.

Table 2.1 Sample Dataset Inventory

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	•			
Product Financials		•		
Product Call Center Data		•		
Live Product Feedback Surveys			•	
Product Sentiment from Social Media				•

2.3.4 Data Conditioning

Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations on the data. A critical step within the Data Analytics Lifecycle, data conditioning can involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables analysis in further phases. Data conditioning is often viewed as a preprocessing step for the data analysis because it involves many operations on the dataset before developing models to process or analyze the data. This implies that the data-conditioning step is performed only by IT, the data owners, a DBA, or a data engineer. However, it is also important to involve the data scientist in this step because many decisions are made in the data conditioning phase that affect subsequent analysis. Part of this phase involves deciding which aspects of particular datasets will be useful to analyze in later steps. Because teams begin forming ideas in this

phase about which data to keep and which data to transform or discard, it is importan to involve multiple team members in these decisions. Leaving such decisions to a single person may cause teams to return to this phase to retrieve data that may have been discarded.

As with the previous example of deciding which data to keep as it relates to fraud detection on credit card usage, it is critical to be thoughtful about which data the team chooses to keep and which data will be discarded. This can have far-reaching consequences that will cause the team to retrace previous steps if the team discards too much of the data at too early a point in this process. Typically, data science teams would rather keep more data than too little data for the analysis. Additional questions and considerations for the data conditioning step include these.

- What are the data sources? What are the target fields (for example, columns of the tables)?
- How clean is the data?
- How consistent are the contents and files? Determine to what degree the data contains missing or inconsistent values and if the data contains values deviating from normal.
- Assess the consistency of the data types. For instance, if the team expects certain data to be numeric, confirm it is numeric or if it is a mixture of alphanumeric strings and text.
- Review the content of data columns or other inputs, and check to ensure they make sense. For instance, if the project involves analyzing income levels, preview the data to confirm that the income values are positive or if it is acceptable to have zeros or negative values.
- Look for any evidence of systematic error. Examples include data feeds from sensors or other data sources breaking without anyone noticing, which causes invalid, incorrect, or missing data values. In addition, review the data to gauge if the definition of the data is the same over all measurements. In some cases, a data column is repurposed, or the column stops being populated, without this change being annotated or without others being notified.

2.3.5 Survey and Visualize

After the team has collected and obtained at least some of the datasets needed for the subsequent analysis, a useful step is to leverage data visualization tools to gain an overview of the data. Seeing high-level patterns in the data enables one to understand characteristics about the data very quickly. One example is using data visualization to examine data quality, such as whether the data contains many unexpected values or other indicators of dirty data. (Dirty data will be discussed further in Chapter 3.) Another example is skewness, such as if the majority of the data is heavily shifted toward one value or end of a continuum.

Shneiderman [9] is well known for his mantra for visual data analysis of “overview first,

zoom and filter, then details-on-demand." This is a pragmatic approach to visual data analysis. It enables the user to find areas of interest, zoom and filter to find more detailed information about a particular area of the data, and then find the detailed data behind a particular area. This approach provides a high-level view of the data and a great deal of information about a given dataset in a relatively short period of time.

When pursuing this approach with a data visualization tool or statistical package, the following guidelines and considerations are recommended.

- Review data to ensure that calculations remained consistent within columns or across tables for a given data field. For instance, did customer lifetime value change at some point in the middle of data collection? Or if working with financials, did the interest calculation change from simple to compound at the end of the year?
- Does the data distribution stay consistent over all the data? If not, what kinds of actions should be taken to address this problem?
- Assess the granularity of the data, the range of values, and the level of aggregation of the data.
- Does the data represent the population of interest? For marketing data, if the project is focused on targeting customers of child-rearing age, does the data represent that, or is it full of senior citizens and teenagers?
- For time-related variables, are the measurements daily, weekly, monthly? Is that good enough? Is time measured in seconds everywhere? Or is it in milliseconds in some places? Determine the level of granularity of the data needed for the analysis, and assess whether the current level of timestamps on the data meets that need.
- Is the data standardized/normalized? Are the scales consistent? If not, how consistent or irregular is the data?
- For geospatial datasets, are state or country abbreviations consistent across the data? Are personal names normalized? English units? Metric units?

These are typical considerations that should be part of the thought process as the team evaluates the datasets that are obtained for the project. Becoming deeply knowledgeable about the data will be critical when it comes time to construct and run models later in the process.

2.3.6 Common Tools for the Data Preparation Phase

Several tools are commonly used for this phase:

- **Hadoop** [10] can perform massively parallel ingest and custom analysis for web traffic parsing, GPS location analytics, genomic analysis, and combining of massive unstructured data feeds from multiple sources.
- **Alpine Miner** [11] provides a graphical user interface (GUI) for creating analytic workflows, including data manipulations and a series of analytic events such as staged data-mining techniques (for example, first select the top 100 customers, and

then run descriptive statistics and c ustering) on Postgres SQL and other Big Data sources.

- **OpenRefine** (formerly called Google Refine) [12] is "a free, open source, powerful tool for working with messy data." It is a popular GUI-based tool for performing data transformations, and it's one of the most robust free tools currently available.
- Similar to OpenRefine, **Data Wrangler** [13] is an interactive tool for data cleaning and transformation. Wrangler was developed at Stanford University and can be used to perform many transformations on a given dataset. In addition, data transformation outputs can be put into Java or Python. The advantage of this feature is that a subset of the data can be manipulated in Wrangler via its GUI, and then the same operations can be written out as Java or Python code to be executed against the full, larger dataset offline in a local analytic sandbox.

For Phase 2, the team needs assistance from IT, DBAs, or whoever controls the Enterprise Data Warehouse (EDW) for data sources the data science team would like to use.

2.4 Phase 3: Model Planning

In Phase 3, the data science team identifies candidate models to apply to the data for clustering, classifying, or finding relationships in the data depending on the goal of the project, as shown in Figure 2.5. It is during this phase that the team refers to the hypotheses developed in Phase 1, when they first became acquainted with the data and understanding the business problems or domain area. These hypotheses help the team frame the analytics to execute in Phase 4 and select the right methods to achieve its objectives.

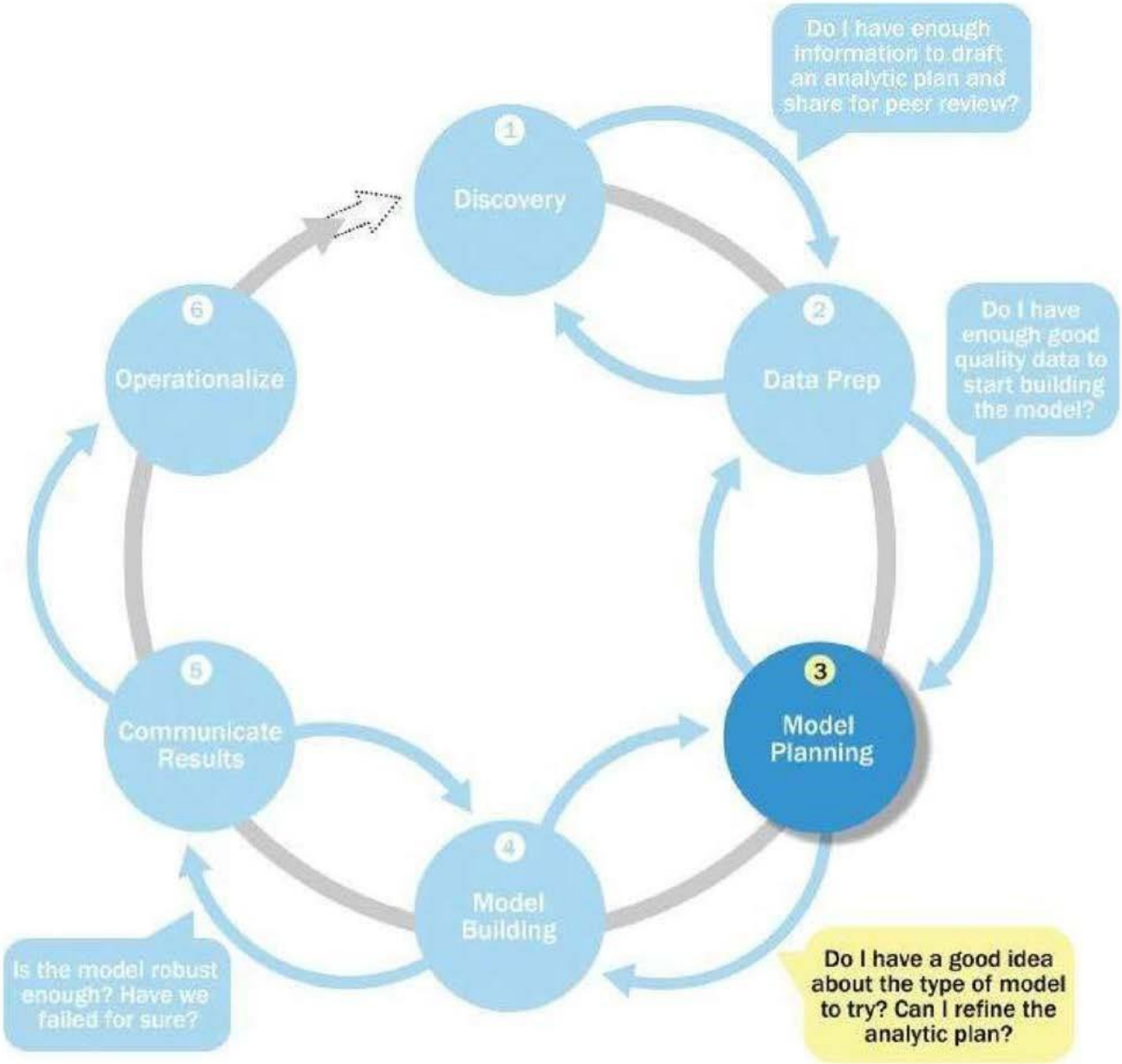


Figure 2.5 Model planning phase

Some of the activities to consider in this phase include the following:

- Assess the structure of the datasets. The structure of the datasets is one factor that dictates the tools and analytical techniques for the next phase. Depending on whether the team plans to analyze textual data or transactional data, for example, different

tools and approaches are required.

- Ensure that the analytical techniques enable the team to meet the business objectives and accept or reject the working hypotheses.
- Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow. A few example models include association rules (Chapter 5, “Advanced Analytical Theory and Methods: Association Rules”) and logistic regression (Chapter 6, “Advanced Analytical Theory and Methods: Regression”). Other tools, such as Alpine Miner, enable users to set up a series of steps and analyses and can serve as a front-end user interface (UI) for manipulating Big Data sources in PostgreSQL.

In addition to the considerations just listed, it is useful to research and understand how other analysts generally approach a specific kind of problem. Given the kind of data and resources that are available, evaluate whether similar, existing approaches will work or if the team will need to create something new. Many times teams can get ideas from analogous problems that other people have solved in different industry verticals or domain areas. Table 2.2 summarizes the results of an exercise of this type, involving several domain areas and the types of models previously used in a classification type of problem after conducting research on churn models in multiple industry verticals. Performing this sort of diligence gives the team ideas of how others have solved similar problems and presents the team with a list of candidate models to try as part of the model planning phase.

Table 2.2 Research on Model Planning in Industry Verticals

Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression

2.4.1 Data Exploration and Variable Selection

Although some data exploration takes place in the data preparation phase, those activities focus mainly on data hygiene and on assessing the quality of the data itself. In Phase 3, the objective of the data exploration is to understand the relationships among the variables to inform selection of the variables and methods and to understand the problem domain. As with earlier phases of the Data Analytics Lifecycle, it is important to spend time and focus attention on this preparatory work to make the subsequent phases of model selection and execution easier and more efficient. A common way to conduct this step involves using tools to perform data visualizations. Approaching the data exploration in this way aids the team in previewing the data and assessing relationships between variables at a high level.

In many cases, stakeholders and subject matter experts have instincts and hunches about

what the data science team should be considering and analyzing. Likely, this group had some hypothesis that led to the genesis of the project. Often, stakeholders have a good grasp of the problem and domain, although they may not be aware of the subtleties within the data or the model needed to accept or reject a hypothesis. Other times, stakeholders may be correct, but for the wrong reasons (for instance, they may be correct about a correlation that exists but infer an incorrect reason for the correlation). Meanwhile, data scientists have to approach problems with an unbiased mind-set and be ready to question all assumptions.

As the team begins to question the incoming assumptions and test initial ideas of the project sponsors and stakeholders, it needs to consider the inputs and data that will be needed, and then it must examine whether these inputs are actually correlated with the outcomes that the team plans to predict or analyze. Some methods and types of models will handle correlated variables better than others. Depending on what the team is attempting to solve, it may need to consider an alternate method, reduce the number of data inputs, or transform the inputs to allow the team to use the best method for a given business problem. Some of these techniques will be explored further in Chapter 3 and Chapter 6.

The key to this approach is to aim for capturing the most essential predictors and variables rather than considering every possible variable that people think may influence the outcome. Approaching the problem in this manner requires iterations and testing to identify the most essential variables for the intended analyses. The team should plan to test a range of variables to include in the model and then focus on the most important and influential variables.

If the team plans to run regression analyses, identify the candidate predictors and outcome variables of the model. Plan to create variables that determine outcomes but demonstrate a strong relationship to the outcome rather than to the other input variables. This includes remaining vigilant for problems such as serial correlation, multicollinearity, and other typical data modeling challenges that interfere with the validity of these models. Sometimes these issues can be avoided simply by looking at ways to reframe a given problem. In addition, sometimes determining correlation is all that is needed (“black box prediction”), and in other cases, the objective of the project is to understand the causal relationship better. In the latter case, the team wants the model to have explanatory power and needs to forecast or stress test the model under a variety of situations and with different datasets.

2.4.2 Model Selection

In the model selection subphase, the team’s main goal is to choose an analytical technique, or a short list of candidate techniques, based on the end goal of the project. For the context of this book, a *model* is discussed in general terms. In this case, a model simply refers to an abstraction from reality. One observes events happening in a real-world situation or with live data and attempts to construct models that emulate this behavior with a set of rules and conditions. In the case of machine learning and data mining, these rules and conditions are grouped into several general sets of techniques, such as classification, association rules, and clustering. When reviewing this list of types of potential models, the team can winnow down the list to several viable models to try to address a given problem.

More details on matching the right models to common types of business problems are provided in Chapter 3 and Chapter 4, “Advanced Analytical Theory and Methods: Clustering.”

An additional consideration in this area for dealing with Big Data involves determining if the team will be using techniques that are best suited for structured data, unstructured data, or a hybrid approach. For instance, the team can leverage MapReduce to analyze unstructured data, as highlighted in Chapter 10. Lastly, the team should take care to identify and document the modeling assumptions it is making as it chooses and constructs preliminary models.

Typically, teams create the initial models using a statistical software package such as R, SAS, or Matlab. Although these tools are designed for data mining and machine learning algorithms, they may have limitations when applying the models to very large datasets, as is common with Big Data. As such, the team may consider redesigning these algorithms to run in the database itself during the pilot phase mentioned in Phase 6.

The team can move to the model building phase once it has a good idea about the type of model to try and the team has gained enough knowledge to refine the analytics plan. Advancing from this phase requires a general methodology for the analytical model, a solid understanding of the variables and techniques to use, and a description or diagram of the analytic workflow.

2.4.3 Common Tools for the Model Planning Phase

Many tools are available to assist in this phase. Here are several of the more common ones:

- **R** [14] has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code. In addition, it has the ability to interface with databases via an ODBC connection and execute statistical tests and analyses against Big Data via an open source connection. These two factors make R well suited to performing statistical tests and analytics on Big Data. As of this writing, R contains nearly 5,000 packages for data analysis and graphical representation. New packages are posted frequently, and many companies are providing value-add services for R (such as training, instruction, and best practices), as well as packaging it in ways to make it easier to use and more robust. This phenomenon is similar to what happened with Linux in the late 1980s and early 1990s, when companies appeared to package and make Linux easier for companies to consume and deploy. Use R with file extracts for offline analysis and optimal performance, and use RODBC connections for dynamic queries and faster development.
- **SQL Analysis services** [15] can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
- **SAS/ACCESS** [16] provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, and OLE DB. SAS itself is generally used on file extracts, but with SAS/ACCESS, users can connect to relational databases (such as Oracle or Teradata) and data warehouse appliances (such as Greenplum or Aster), files, and enterprise applications (such as SAP and `Salesforce.com`).

2.5 Phase 4: Model Building

In Phase 4, the data science team needs to develop datasets for training, testing, and production purposes. These datasets enable the data scientist to develop the analytical model and train it ("training data"), while holding aside some of the data ("hold-out data" or "test data") for testing the model. (These topics are addressed in more detail in Chapter 3.) During this process, it is critical to ensure that the training and test datasets are sufficiently robust for the model and analytical techniques. A simple way to think of these datasets is to view the training dataset for conducting the initial experiments and the test sets for validating an approach once the initial experiments and models have been run.

In the model building phase, shown in Figure 2.6, an analytical model is developed and fit on the training data and evaluated (scored) against the test data. The phases of model planning and model building can overlap quite a bit, and in practice one can iterate back and forth between the two phases for a while before settling on a final model.

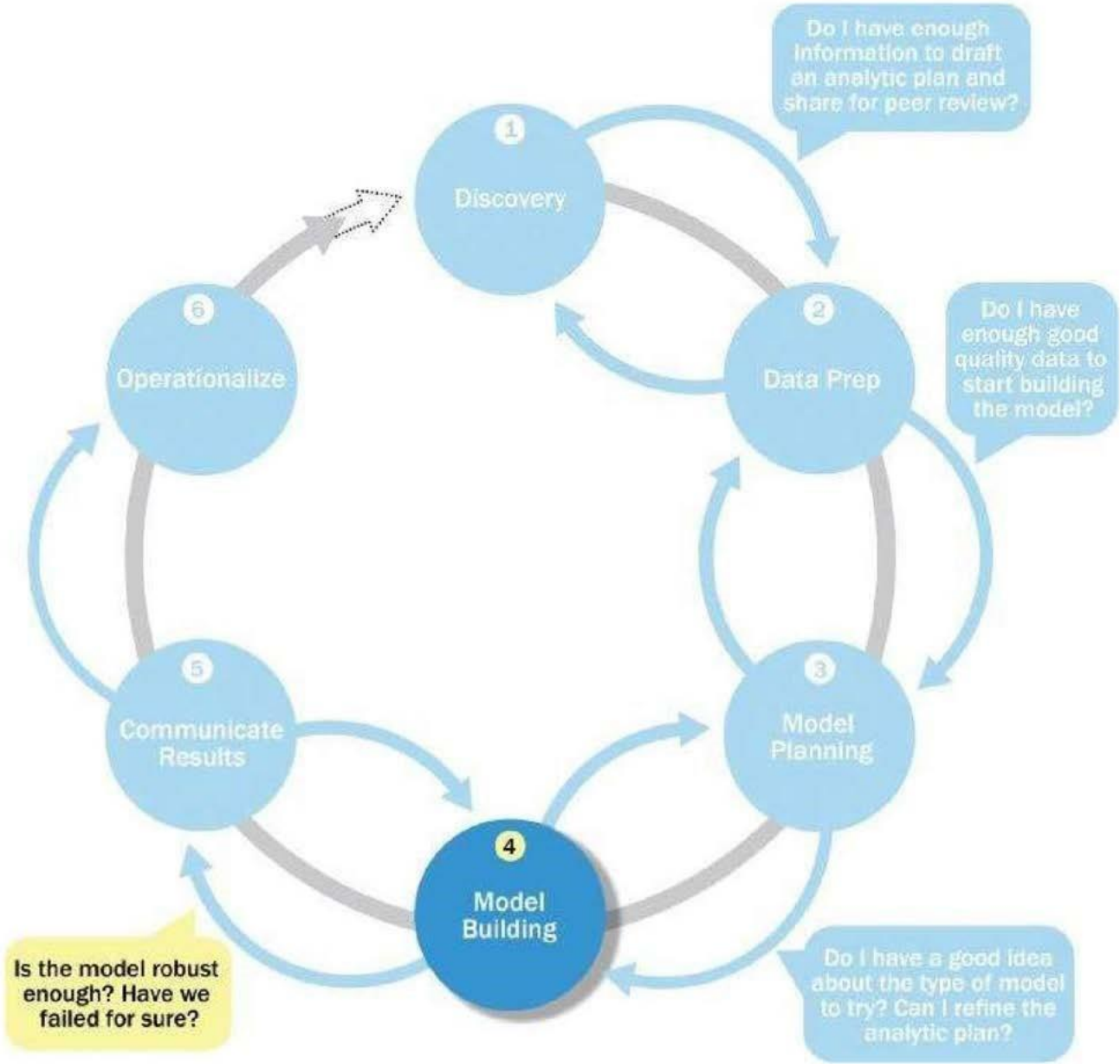


Figure 2.6 Model building phase

Although the modeling techniques and logic required to develop models can be highly complex, the actual duration of this phase can be short compared to the time spent preparing the data and defining the approaches. In general, plan to spend more time preparing and learning the data (Phases 1–2) and crafting a presentation of the findings (Phase 5). Phases 3 and 4 tend to move more quickly, although they are more complex from a conceptual standpoint.

As part of this phase, the data science team needs to execute the models defined in Phase 3.

During this phase, users run models from analytical software packages, such as R or SAS, on file extracts and small datasets for testing purposes. On a small scale, assess the validity of the model and its results. For instance, determine if the model accounts for most of the data and has robust predictive power. At this point, refine the models to optimize the results, such as by modifying variable inputs or reducing correlated variables where appropriate. In Phase 3, the team may have had some knowledge of correlated variables or problematic data attributes, which will be confirmed or denied once the models are actually executed. When immersed in the details of constructing models and transforming data, many small decisions are often made about the data and the approach for the modeling. These details can be easily forgotten once the project is completed. Therefore, it is vital to record the results and logic of the model during this phase. In addition, one must take care to record any operating assumptions that were made in the modeling process regarding the data or the context.

Creating robust models that are suitable to a specific situation requires thoughtful consideration to ensure the models being developed ultimately meet the objectives outlined in Phase 1. Questions to consider include these:

- Does the model appear valid and accurate on the test data?
- Does the model output/behavior make sense to the domain experts? That is, does it appear as if the model is giving answers that make sense in this context?
- Do the parameter values of the fitted model make sense in the context of the domain?
- Is the model sufficiently accurate to meet the goal?
- Does the model avoid intolerable mistakes? Depending on context, false positives may be more serious or less serious than false negatives, for instance. (False positives and false negatives are discussed further in Chapter 3 and Chapter 7, “Advanced Analytical Theory and Methods: Classification.”)
- Are more data or more inputs needed? Do any of the inputs need to be transformed or eliminated?
- Will the kind of model chosen support the runtime requirements?
- Is a different form of the model required to address the business problem? If so, go back to the model planning phase and revise the modeling approach.

Once the data science team can evaluate either if the model is sufficiently robust to solve the problem or if the team has failed, it can move to the next phase in the Data Analytics

Lifecycle.

2.5.1 Common Tools for the Model Building Phase

There are many tools available to assist in this phase, focused primarily on statistical analysis or data mining software. Common tools in this space include, but are not limited to, the following:

- Commercial Tools:
  - **SAS Enterprise Miner** [17] allows users to run predictive and descriptive models based on large volumes of data from across the enterprise. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.
  - **SPSS Modeler** [18] (provided by IBM and now called IBM SPSS Modeler) offers methods to explore and analyze data through a GUI.
  - **Matlab** [19] provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.
  - **Alpine Miner** [11] provides a GUI front end for users to develop analytic workflows and interact with Big Data tools and platforms on the back end.
  - **STATISTICA** [20] and **Mathematica** [21] are also popular and well-regarded data mining and analytics tools.
- Free or Open Source tools:
  - **R and PL/R** [14] R was described earlier in the model planning phase, and PL/R is a procedural language for PostgreSQL with R. Using this approach means that R commands can be executed in database. This technique provides higher performance and is more scalable than running R in memory.
  - **Octave** [22], a free software programming language for computational modeling, has some of the functionality of Matlab. Because it is freely available, Octave is used in major universities when teaching machine learning.
  - **WEKA** [23] is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
  - **Python** is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib.
  - **SQL** in-database implementations, such as **MADlib** [24], provide an alterative to in-memory desktop analytical tools. MADlib provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

2.6 Phase 5: Communicate Results

After executing the model, the team needs to compare the outcomes of the modeling to the criteria established for success and failure. In Phase 5, shown in Figure 2.7, the team considers how best to articulate the findings and outcomes to the various team members and stakeholders, taking into account caveats, assumptions, and any limitations of the results. Because the presentation is often circulated within an organization, it is critical to articulate the results properly and position the findings in a way that is appropriate for the audience.

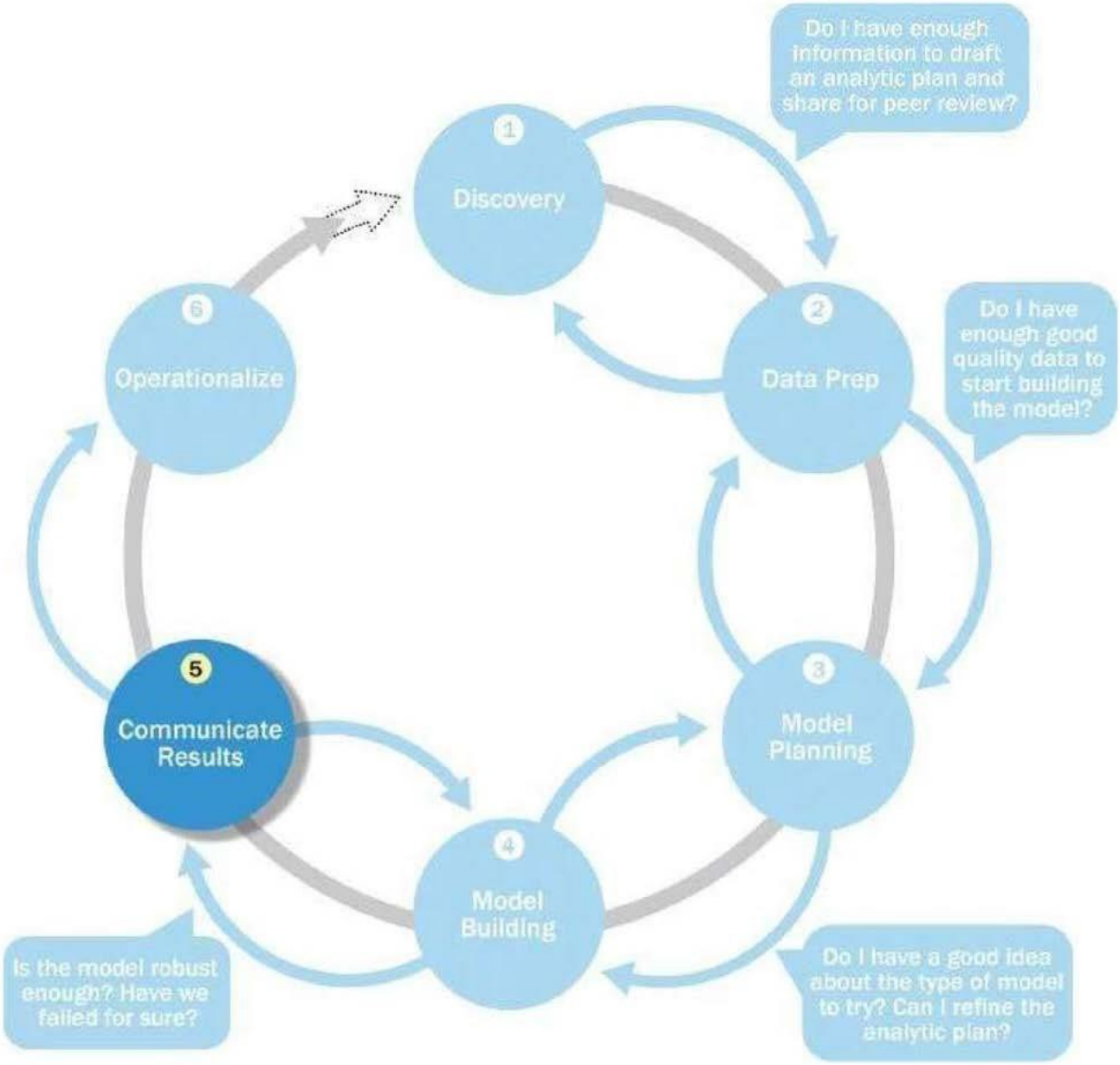


Figure 2.7 Communicate results phase

As part of Phase 5, the team needs to determine if it succeeded or failed in its objectives. Many times people do not want to admit to failing, but in this instance failure should not be considered as a true failure, but rather as a failure of the data to accept or reject a given hypothesis adequately. This concept can be counterintuitive for those who have been told their whole careers not to fail. However, the key is to remember that the team must be

rigorous enough with the data to determine whether it will prove or disprove the hypotheses outlined in Phase 1 (discovery). Sometimes teams have only done a superficial analysis, which is not robust enough to accept or reject a hypothesis. Other times, teams perform very robust analysis and are searching for ways to show results, even when results may not be there. It is important to strike a balance between these two extremes when it comes to analyzing data and being pragmatic in terms of showing real-world results.

When conducting this assessment, determine if the results are statistically significant and valid. If they are, identify the aspects of the results that stand out and may provide salient findings when it comes time to communicate them. If the results are not valid, think about adjustments that can be made to refine and iterate on the model to make it valid. During this step, assess the results and identify which data points may have been surprising and which were in line with the hypotheses that were developed in Phase 1. Comparing the actual results to the ideas formulated early on produces additional ideas and insights that would have been missed if the team had not taken time to formulate initial hypotheses early in the process.

By this time, the team should have determined which model or models address the analytical challenge in the most appropriate way. In addition, the team should have ideas of some of the findings as a result of the project. The best practice in this phase is to record all the findings and then select the three most significant ones that can be shared with the stakeholders. In addition, the team needs to reflect on the implications of these findings and measure the business value. Depending on what emerged as a result of the model, the team may need to spend time quantifying the business impact of the results to help prepare for the presentation and demonstrate the value of the findings. Doug Hubbard's work [6] offers insights on how to assess intangibles in business and quantify the value of seemingly unmeasurable things.

Now that the team has run the model, completed a thorough discovery phase, and learned a great deal about the datasets, reflect on the project and consider what obstacles were in the project and what can be improved in the future. Make recommendations for future work or improvements to existing processes, and consider what each of the team members and stakeholders needs to fulfill her responsibilities. For instance, sponsors must champion the project. Stakeholders must understand how the model affects their processes. (For example, if the team has created a model to predict customer churn, the Marketing team must understand how to use the churn model predictions in planning their interventions.) Production engineers need to operationalize the work that has been done. In addition, this is the phase to underscore the business benefits of the work and begin making the case to implement the logic into a live production environment.

As a result of this phase, the team will have documented the key findings and major insights derived from the analysis. The deliverable of this phase will be the most visible portion of the process to the outside stakeholders and sponsors, so take care to clearly articulate the results, methodology, and business value of the findings. More details will be provided about data visualization tools and references in Chapter 12, "The Endgame, or Putting It All Together."

2.7 Phase 6: Operationalize

In the final phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users. In Phase 4, the team scored the model in the analytics sandbox. Phase 6, shown in Figure 2.8, represents the first time that most analytics teams approach deploying the new analytical methods or models in a production environment. Rather than deploying these models immediately on a wide-scale basis, the risk can be managed more effectively and the team can learn by undertaking a small scope, pilot deployment before a wide-scale rollout. This approach enables the team to learn about the performance and related constraints of the model in a production environment on a small scale and make adjustments before a full deployment. During the pilot project, the team may need to consider executing the algorithm in the database rather than with in-memory tools such as R because the run time is significantly faster and more efficient than running in-memory, especially on larger datasets.

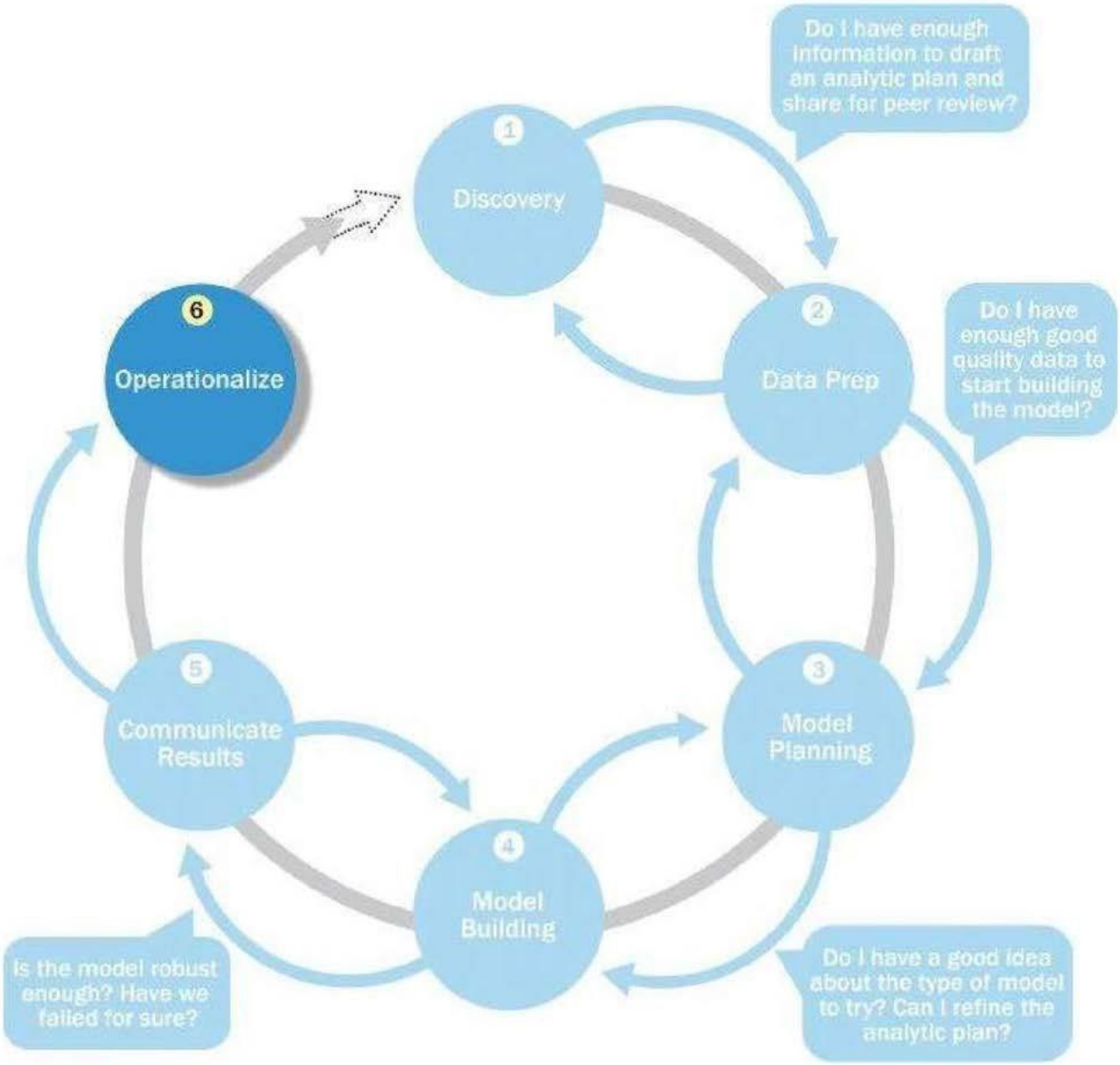


Figure 2.8 Model operationalize phase

While scoping the effort involved in conducting a pilot project, consider running the model in a production environment for a discrete set of products or a single line of business, which tests the model in a live setting. This allows the team to learn from the deployment and make any needed adjustments before launching the model across the enterprise. Be aware that this phase can bring in a new set of team members—usually the engineers responsible for the production environment who have a new set of issues and concerns beyond those of the core project team. This technical group needs to ensure that running the model fits smoothly into the production environment and that the model can be integrated into related business processes.

Part of the operationalizing phase includes creating a mechanism for performing ongoing monitoring of model accuracy and, if accuracy degrades, finding ways to retrain the model. If feasible, design alerts for when the model is operating "out-of-bounds." This includes situations when the inputs are beyond the range that the model was trained on, which may cause the outputs of the model to be inaccurate or invalid. If this begins to happen regularly, the model needs to be retrained on new data.

Often, analytical projects yield new insights about a business, a problem, or an idea that people may have taken at face value or thought was impossible to explore. Four main deliverables can be created to meet the needs of most stakeholders. This approach for developing the four deliverables is discussed in greater detail in Chapter 12.

Figure 2.9 portrays the key outputs for each of the main stakeholders of an analytics project and what they usually expect at the conclusion of a project.

- **Business User** typically tries to determine the benefits and implications of the findings to the business.
- **Project Sponsor** typically asks questions related to the business impact of the project, the risks and return on investment (ROI), and the way the project can be evangelized within the organization (and beyond).
- **Project Manager** needs to determine if the project was completed on time and within budget and how well the goals were met.
- **Business Intelligence Analyst** needs to know if the reports and dashboards he manages will be impacted and need to change.
- **Data Engineer** and **Database Administrator** (DBA) typically need to share their code from the analytics project and create a technical document on how to implement it.
- **Data Scientist** needs to share the code and explain the model to her peers, managers, and other stakeholders.

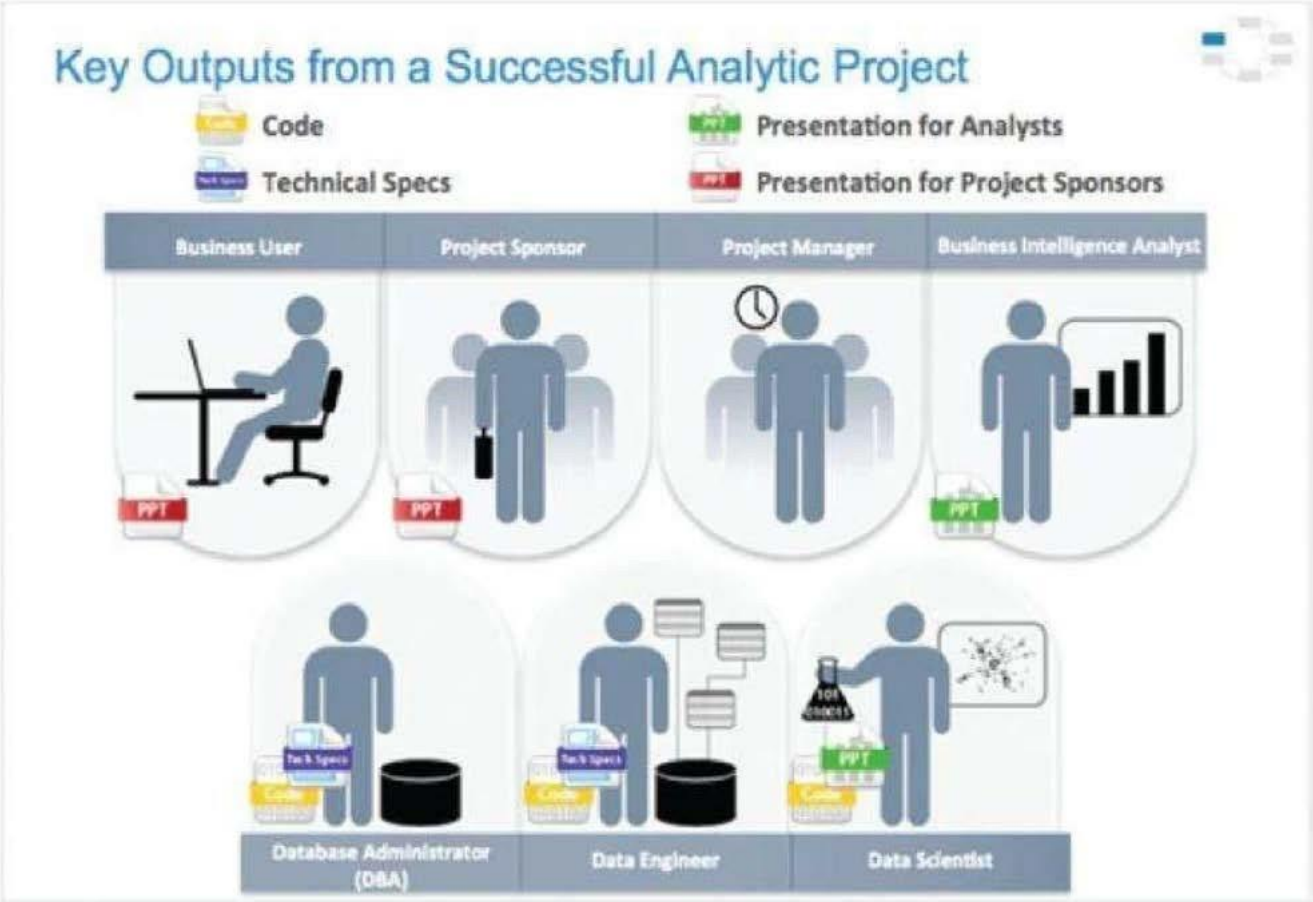


Figure 2.9 Key outputs from a successful analytics project

Although these seven roles represent many interests within a project, these interests usually overlap, and most of them can be met with four main deliverables.

- Presentation for project sponsors: This contains high-level takeaways for executive level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- Presentation for analysts, which describes business process changes and reporting changes. Fellow data scientists will want the details and are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC] curves, density plots, and histograms shown in Chapter 3 and Chapter 7).
- Code for technical people.
- Technical specifications of implementing the code.

As a general rule, the more executive the audience, the more succinct the presentation needs to be. Most executive sponsors attend many briefings in the course of a day or a week. Ensure that the presentation gets to the point quickly and frames the results in terms of value to the sponsor's organization. For instance, if the team is working with a bank to analyze cases of credit card fraud, highlight the frequency of fraud, the number of cases in the past month or year, and the cost or revenue impact to the bank (or focus on the reverse —how much more revenue the bank could gain if it addresses the fraud problem). This demonstrates the business impact better than deep dives on the methodology. The

presentation needs to include supporting information about analytical methodology and data sources, but generally only as supporting detail or to ensure the audience has confidence in the approach that was taken to analyze the data.

When presenting to other audiences with more quantitative backgrounds, focus more time on the methodology and findings. In these instances, the team can be more expansive in describing the outcomes, methodology, and analytical experiment with a peer group. This audience will be more interested in the techniques, especially if the team developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems. In addition, use imagery or data visualization when possible. Although it may take more time to develop imagery, people tend to remember mental pictures to demonstrate a point more than long lists of bullets [25]. Data visualization and presentations are discussed further in Chapter 12.

2.8 Case Study: Global Innovation Network and Analysis (GINA)

EMC's Global Innovation Network and Analytics (GINA) team is a group of senior technologists located in centers of excellence (COEs) around the world. This team's charter is to engage employees across global COEs to drive innovation, research, and university partnerships. In 2012, a newly hired director wanted to improve these activities and provide a mechanism to track and analyze the related information. In addition, this team wanted to create more robust mechanisms for capturing the results of its informal conversations with other thought leaders within EMC, in academia, or in other organizations, which could later be mined for insights.

The GINA team thought its approach would provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separated geographically. It planned to create a data repository containing both structured and unstructured data to accomplish three main goals.

- Store formal and informal data.
- Track research from global technologists.
- Mine the data for patterns and insights to improve the team's operations and strategy.

The GINA case study provides an example of how a team applied the Data Analytics Lifecycle to analyze innovation data at EMC. Innovation is typically a difficult concept to measure, and this team wanted to look for ways to use advanced analytical methods to identify key innovators within the company.

2.8.1 Phase 1: Discovery

In the GINA project's discovery phase, the team began identifying data sources. Although GINA was a group of technologists skilled in many different aspects of engineering, it had some data and ideas about what it wanted to explore but lacked a formal team that could perform these analytics. After consulting with various experts including Tom Davenport, a noted expert in analytics at Babson College, and Peter Gloor, an expert in collective intelligence and creator of CoIN (Collaborative Innovation Networks) at MIT, the team decided to crowdsource the work by seeking volunteers within EMC.

Here is a list of how the various roles on the working team were fulfilled.

- **Business User, Project Sponsor, Project Manager:** Vice President from Office of the CTO
- **Business Intelligence Analyst:** Representatives from IT
- **Data Engineer** and **Database Administrator (DBA):** Representatives from IT
- **Data Scientist:** Distinguished Engineer, who also developed the social graphs shown in the GINA case study

The project sponsor's approach was to leverage social media and blogging [26] to accelerate the collection of innovation and research data worldwide and to motivate teams of "volunteer" data scientists at worldwide locations. Given that he lacked a formal team, he needed to be resourceful about finding people who were both capable and willing to volunteer their time to work on interesting problems. Data scientists tend to be passionate about data, and the project sponsor was able to tap into this passion of highly talented people to accomplish challenging work in a creative way.

The data for the project fell into two main categories. The first category represented five years of idea submissions from EMC's internal innovation contests, known as the Innovation Roadmap (formerly called the Innovation Showcase). The Innovation Roadmap is a formal, organic innovation process whereby employees from around the globe submit ideas that are then vetted and judged. The best ideas are selected for further incubation. As a result, the data is a mix of structured data, such as idea counts, submission dates, inventor names, and unstructured content, such as the textual descriptions of the ideas themselves.

The second category of data encompassed minutes and notes representing innovation and research activity from around the world. This also represented a mix of structured and unstructured data. The structured data included attributes such as dates, names, and geographic locations. The unstructured documents contained the "who, what, when, and where" information that represents rich data about knowledge growth and transfer within the company. This type of information is often stored in business silos that have little to no visibility across disparate research teams.

The 10 main IHs that the GINA team developed were as follows:

- **IH1:** Innovation activity in different geographic regions can be mapped to corporate strategic directions.
- **IH2:** The length of time it takes to deliver ideas decreases when global knowledge transfer occurs as part of the idea delivery process.
- **IH3:** Innovators who participate in global knowledge transfer deliver ideas more quickly than those who do not.
- **IH4:** An idea submission can be analyzed and evaluated for the likelihood of receiving funding.
- **IH5:** Knowledge discovery and growth for a particular topic can be measured and compared across geographic regions.
- **IH6:** Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.
- **IH7:** Strategic corporate themes can be mapped to geographic regions.
- **IH8:** Frequent knowledge expansion and transfer events reduce the time it takes to generate a corporate asset from an idea.
- **IH9:** Lineage maps can reveal when knowledge expansion and transfer did not (or has not) resulted in a corporate asset.
- **IH10:** Emerging research topics can be classified and mapped to specific ideators,

innovators, boundary spanners, and assets.

The GINA (IHs) can be grouped into two categories:

- Descriptive analytics of what is currently happening to spark further creativity, collaboration, and asset generation
- Predictive analytics to advise executive management of where it should be investing in the future

2.8.2 Phase 2: Data Preparation

The team partnered with its IT department to set up a new analytics sandbox to store and experiment on the data. During the data exploration exercise, the data scientists and data engineers began to notice that certain data needed conditioning and normalization. In addition, the team realized that several missing datasets were critical to testing some of the analytic hypotheses.

As the team explored the data, it quickly realized that if it did not have data of sufficient quality or could not get good quality data, it would not be able to perform the subsequent steps in the lifecycle process. As a result, it was important to determine what level of data quality and cleanliness was sufficient for the project being undertaken. In the case of the GINA, the team discovered that many of the names of the researchers and people interacting with the universities were misspelled or had leading and trailing spaces in the datastore. Seemingly small problems such as these in the data had to be addressed in this phase to enable better analysis and data aggregation in subsequent phases.

2.8.3 Phase 3: Model Planning

In the GINA project, for much of the dataset, it seemed feasible to use social network analysis techniques to look at the networks of innovators within EMC. In other cases, it was difficult to come up with appropriate ways to test hypotheses due to the lack of data. In one case (IH9), the team made a decision to initiate a longitudinal study to begin tracking data points over time regarding people developing new intellectual property. This data collection would enable the team to test the following two ideas in the future:

- **IH8:** Frequent knowledge expansion and transfer events reduce the amount of time it takes to generate a corporate asset from an idea.
- **IH9:** Lineage maps can reveal when knowledge expansion and transfer did not (or has not) result(ed) in a corporate asset.

For the longitudinal study being proposed, the team needed to establish goal criteria for the study. Specifically, it needed to determine the end goal of a successful idea that had traversed the entire journey. The parameters related to the scope of the study included the following considerations:

- Identify the right milestones to achieve this goal.
- Trace how people move ideas from each milestone toward the goal.
- Once this is done, trace ideas that die, and trace others that reach the goal. Compare the journeys of ideas that make it and those that do not.
- Compare the times and the outcomes using a few different methods (depending on how the data is collected and assembled). These could be as simple as t-tests or perhaps involve different types of classification algorithms.

2.8.4 Phase 4: Model Building

In Phase 4, the GINA team employed several analytical methods. This included work by the data scientist using Natural Language Processing (NLP) techniques on the textual descriptions of the Innovation Roadmap ideas. In addition, he conducted social network analysis using R and RStudio, and then he developed social graphs and visualizations of the network of communications related to innovation using R's `ggplot2` package. Examples of this work are shown in Figures 2.10 and 2.11.

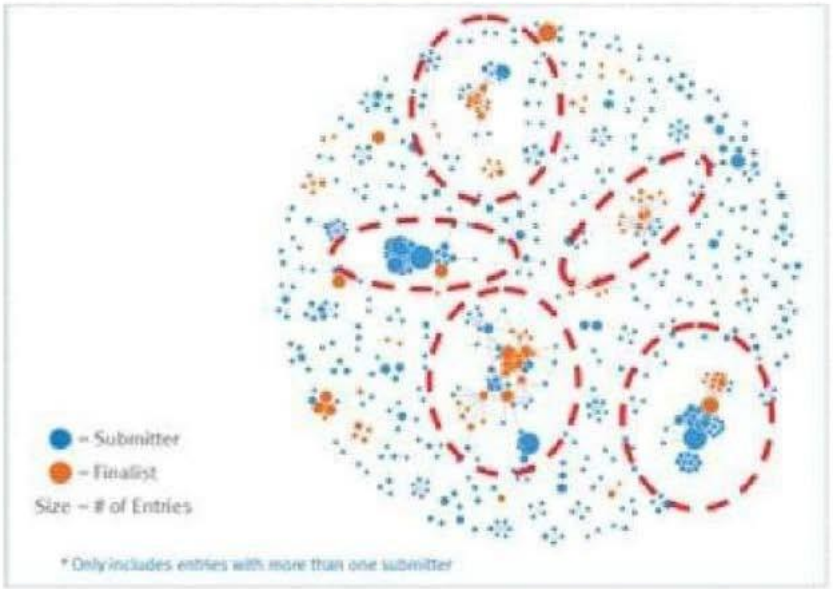


Figure 2.10 Social graph [27] visualization of idea submitters and finalists

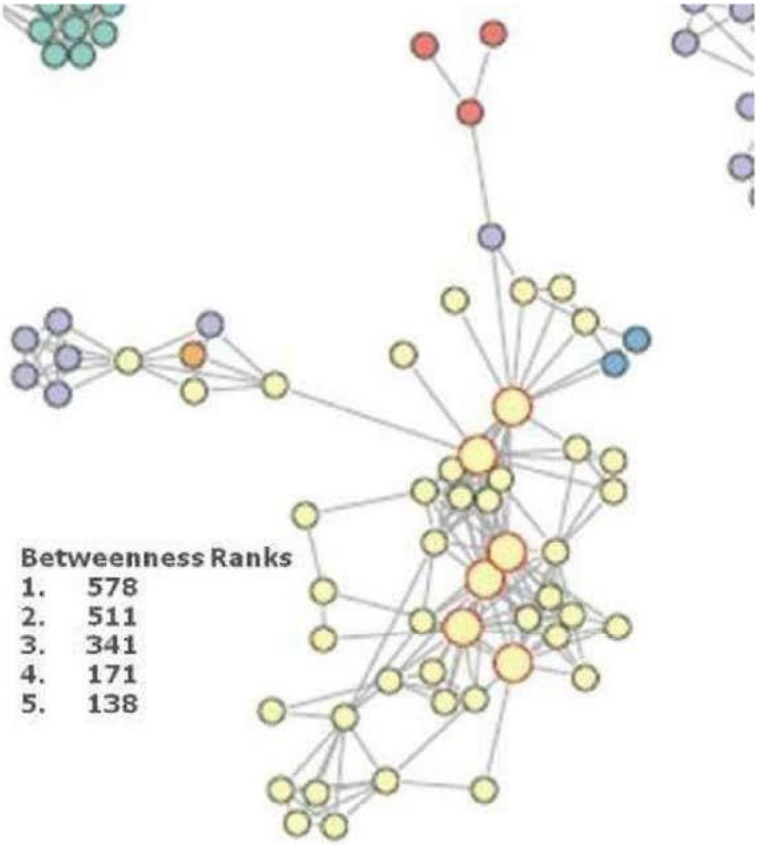


Figure 2.11 Social graph visualization of top innovation influencers

Figure 2.10 shows social graphs that portray the relationships between idea submitters within GINA. Each color represents an innovator from a different country. The large dots with red circles around them represent hubs. A ***hub*** represents a person with high connectivity and a high "betweenness" score. The cluster in Figure 2.11 contains geographic variety, which is critical to prove the hypothesis about geographic boundary spanners. One person in this graph has an unusually high score when compared to the rest of the nodes in the graph. The data scientist identified this person and ran a query against his name within the analytic sandbox. These actions yielded the following information about this research scientist (from the social graph), which illustrated how influential he was within his business unit and across many other areas of the company worldwide:

- In 2011, he attended the ACM SIGMOD conference, which is a top-tier conference on large-scale data management problems and databases.
- He visited employees in France who are part of the business unit for EMC's content management teams within Documentum (now part of the Information Intelligence Group, or IIG).
- He presented his thoughts on the SIGMOD conference at a virtual brownbag session attended by three employees in Russia, one employee in Cairo, one employee in Ireland, one employee in India, three employees in the United States, and one employee in Israel.
- In 2012, he attended the SDM 2012 conference in California.

- On the same trip he visited innovators and researchers at EMC federated companies, Pivotal and VMware.
- Later on that trip he stood before an internal council of technology leaders and introduced two of his researchers to dozens of corporate innovators and researchers.

This finding suggests that at least part of the initial hypothesis is correct; the data can identify innovators who span different geographies and business units. The team used Tableau software for data visualization and exploration and used the Pivotal Greenplum database as the main data repository and analytics engine.

2.8.5 Phase 5: Communicate Results

In Phase 5, the team found several ways to cull results of the analysis and identify the most impactful and relevant findings. This project was considered successful in identifying boundary spanners and hidden innovators. As a result, the CTO office launched longitudinal studies to begin data collection efforts and track innovation results over longer periods of time. The GINA project promoted knowledge sharing related to innovation and researchers spanning multiple areas within the company and outside of it. GINA also enabled EMC to cultivate additional intellectual property that led to additional research topics and provided opportunities to forge relationships with universities for joint academic research in the fields of Data Science and Big Data. In addition, the project was accomplished with a limited budget, leveraging a volunteer force of highly skilled and distinguished engineers and data scientists.

One of the key findings from the project is that there was a disproportionately high density of innovators in Cork, Ireland. Each year, EMC hosts an innovation contest, open to employees to submit innovation ideas that would drive new value for the company. When looking at the data in 2011, 15% of the finalists and 15% of the winners were from Ireland. These are unusually high numbers, given the relative size of the Cork COE compared to other larger centers in other parts of the world. After further research, it was learned that the COE in Cork, Ireland had received focused training in innovation from an external consultant, which was proving effective. The Cork COE came up with more innovation ideas, and better ones, than it had in the past, and it was making larger contributions to innovation at EMC. It would have been difficult, if not impossible, to identify this cluster of innovators through traditional methods or even anecdotal, word-of-mouth feedback. Applying social network analysis enabled the team to find a pocket of people within EMC who were making disproportionately strong contributions. These findings were shared internally through presentations and conferences and promoted through social media and blogs.

2.8.6 Phase 6: Operationalize

Running analytics against a sandbox filled with notes, minutes, and presentations from innovation activities yielded great insights into EMC's innovation culture. Key findings from the project include these:

innovation/research activities.

- Some of the data is sensitive, and the team needs to consider security and privacy related to the data, such as who can run the models and see the results.
- In addition to running models, a parallel initiative needs to be created to improve basic Business Intelligence activities, such as dashboards, reporting, and queries on research activities worldwide.
- A mechanism is needed to continually reevaluate the model after deployment. Assessing the benefits is one of the main goals of this stage, as is defining a process to retrain the model as needed.

In addition to the actions and findings listed, the team demonstrated how analytics can drive new insights in projects that are traditionally difficult to measure and quantify. This project informed investment decisions in university research projects by the CTO office and identified hidden, high-value innovators. In addition, the CTO office developed tools to help submitters improve ideas using topic modeling as part of new recommender systems to help idea submitters find similar ideas and refine their proposals for new intellectual property.

Table 2.3 outlines an analytics plan for the GINA case study example. Although this project shows only three findings, there were many more. For instance, perhaps the biggest overarching result from this project is that it demonstrated, in a concrete way, that analytics can drive new insights in projects that deal with topics that may seem difficult to measure, such as innovation.

Table 2.3 Analytic Plan from the EMC GINA Project

Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis

13 Chapter 9

Key Findings	3. Created tools to help submitters improve ideas with idea recommender systems

Innovation is an idea that every company wants to promote, but it can be measure innovation or identify ways to increase innovation. This project ex issue from the standpoint of evaluating informal social networks to identif spanners and influential people within innovation subnetworks. In essence, took a seemingly nebulous problem and applied advanced analytical methods answers using an objective, fact-based approach.

Another outcome from the project included the need to supplement analy separate datastore for Business Intelligence reporting, accessible innovation/research initiatives. Aside from supporting decision making, this w mechanism to be informed on discussions and research happening worldwide a members in disparate locations. Finally, it highlighted the value that can through data and subsequent analysis. Therefore, the need was identified to marketing programs to convince people to submit (or inform) the global co their innovation/research activities. The knowledge sharing was critical. Witho would not have been able to perform the analysis and identify the hidden within the company.

Summary

This chapter described the Data Analytics Lifecycle, which is an approach to managing and executing analytical projects. This approach describes the process in six phases.

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

Through these steps, data science teams can identify problems and perform rigorous investigation of the datasets needed for in-depth analysis. As stated in the chapter, although much is written about the analytical methods, the bulk of the time spent on these kinds of projects is spent in preparation—namely, in Phases 1 and 2 (discovery and data preparation). In addition, this chapter discussed the seven roles needed for a data science team. It is critical that organizations recognize that Data Science is a team effort, and a balance of skills is needed to be successful in tackling Big Data projects and other complex projects involving data analytics.

Unit 3

TOPIC 3: STRATEGIC DECISION MAKING

13 Chapter 9

A STRATEGIC DECISION

A strategic decision is one that deals with the long-run future of an entire organization.

Strategic decision making, or strategic planning, describes the process of creating a company's mission and objectives and deciding upon the courses of action a company should pursue to achieve those goals.

WHAT MAKES A DECISION STRATEGIC?

A strategic decision is;

1. RARE-Strategic decisions are unusual and typically have no precedent to follow
2. CONSEQUENTIAL-Commit substantial resources and demand a great deal of commitment from people at all levels
3. DIRECTIVE-Set precedents for lesser decisions and future actions throughout an organization

MODES OF STRATEGIC DECISION MAKING.

According to Mintzberg, there are three typical approaches or modes of strategic decision making;

1. Entrepreneurial Mode-Strategy is made by one powerful individual.

The major focus is on opportunities.

Strategy is guided by the founder’s own vision

The dominant goal is growth of the business

Example…..Amazon.com, founded by Jeff Bezos

2. Adaptive Mode(muddling through)

Characterized by reactive solutions to existing problems

There is no proactive search for new opportunities

There is much bargaining on priorities

Strategy is fragmented and developed to move a corporation forward incrementally.

Common in universities, large hospitals, government agencies and large corporations.

3. Planning Mode

Involves systematic gathering of appropriate information for;

- situation analysis,
- generation of feasible alternative strategies
- Rational selection of the most appropriate strategy

Includes both proactive search for new opportunities and the reactive solution of existing problems

4. Logical Incrementalism

(Added later by Quinn)

Can be viewed as a synthesis of the planning, adaptive, and to a lesser extent the entrepreneurial mode.

Top management has a clear idea of the firm's mission and objectives but chooses to use an interactive process.

They probe the future, experiment and learn from a series of partial (incremental) commitments.

Strategy is allowed to emerge out of debate, discussion and experimentation.

This mode is useful in a rapidly changing environment and where it is crucial to build consensus and develop needed resources before committing an entire organization to a specific strategy.

THE STRATEGIC DECISION MAKING PROCESS.

N.B .The planning mode in most situations (which includes the basic elements of the strategic management process) is more rational and thus a better way to make strategic decisions.

It is more analytical and less political and more appropriate for dealing with complex, changing environments.

THE 8-STEP STRATEGIC DECISION MAKING PROCESS

1. Evaluate current performance results in terms of;

    - Return on investment, profitability, etc.
    - Current mission, objectives, strategies and policies

2. Review corporate governance

    - I.e. performance of the firm's board of directors and top management.

3. Scan and assess the external environment

    - To determine the strategic factors that pose opportunities and threats

4. Scan and assess the internal corporate environment

- To determine the strategic factors that are strengths (esp. core competences) and weaknesses

5. Analyze strategic (SWOT) factors so as to;

- Pinpoint problem areas
- Review and revise the corporate mission and objectives, as necessary

6. Generate, evaluate and select the best alternative strategy

- In light of the analysis conducted in step 5.

7. Implement selected strategies

- Via programs, budgets and procedures

8. Evaluate implemented strategies

- Via feedback systems, and the control of activities to ensure their minimum deviation from plans

THE STRATEGY MAKERS

The ideal strategic management team includes decision makers from all three organizational levels;

- The corporate
- The Business
- The Functional

They are;

1. The Chief executive office (CEO)
2. The product managers (heads of various businesses)
3. The heads of functional areas such as;

    - Finance
    - Marketing
    - Procurement
    - Production
    - Human resource
    - etc

In addition the team obtains input from the operatives or lower-level managers and supervisors.

Planning departments, often headed by a corporate vice president for planning, in large organizations

Medium – sized firms often employ at least one full-time staff member to spearhead strategic data- collection efforts.

In small firms or less progressive larger firms, strategic planning is often spearheaded by an officer or a group of officers designated as a planning committee.

RESPONSIBILITIES OF TOP MANAGEMENT IN THE STRATEGIC PLANNING PROCESS

- They Shoulder broad responsibilities for all the major elements of strategic planning and management.
- They develop the major portions of the strategic plan and reviews, and evaluate and counsel on all other portions.
- They develop environmental analysis and forecasting
- They establish business objectives
- They develop business plans prepared by staff groups

Responsibilities of the chief executive officer (CEO)

- Plays a dominant role in the role in the strategic planning process
- Giving long-term direction to the firm
- The CEO is ultimately responsible for the firm's success and the success of its strategy
- For them to achieve this, they must be strong- willed, company-oriented individuals with self-esteem.
- They often resist delegating authority to formulate or approve strategic decisions
- The CEO is highly involved in the implementation process and together with the managers, they work as a team to drive the process forward.
- The CEO provides the required leadership to influence the employees towards the vision and the mission of the organization.

RISKS OF STRATEGIC MANAGEMENT

- The time that managers spend on the strategic management process may have a negative impact on operational responsibilities

- Managers must be trained to minimize that impact by scheduling their duties to allow the necessary time for strategic activities
- If the formulators of strategy are not intimately involved in its implementation, then the plan may not work.

To minimize the effects of these risks strategy managers must be trained to anticipate and respond to the disappointment of participating subordinates over unattained goals and objectives.

Sensitizing managers to these possible negative consequences and preparing them with effective means of minimizing such consequences will greatly enhance the potential of strategic planning.

In this example, something happened on May 25 and 26 that warrants additional investigation. Getting the participants to start brainstorming about what might have happened before or on those days (perhaps company news, competitive activities, market news, or economic news) is a good starting point for the envisioning work that will occur in the ideation workshop.

There are also many external data sources that can be coupled with the organization's data to provide new perspectives on that same old client data. For example, `www.data.gov` is a valuable source of data covering a wide range of information sources which could be used to help the business users start envisioning what's possible. Figure 9-5 shows an example of integrating government-provided Consumer Price Index (CPI) data with the organization's customer sales data to ascertain if there are customer segments in which the organization's marketing spend is overcommitted or under-committed given the market segment potential.

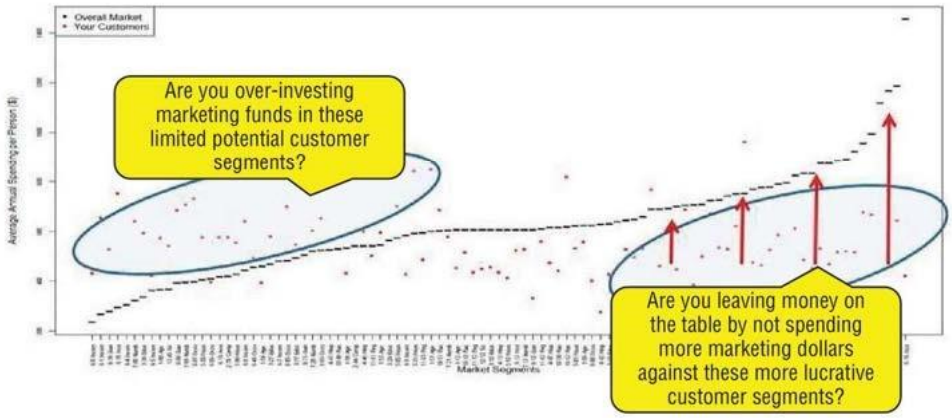


Figure 9-5: Comparison of Markets vs. Your Customers' Spending by Market Segments

Again, the objective of analyzing a small chunk of the organization's data is to personalize the linkage between the business stakeholders and the big data opportunity. You want to fuel the creative thought process to help the business stakeholders explore the realm of what might be possible if the business users had access to new customer, product, and operational insights that can be leveraged as part of their day-to-day business processes.

Step 3: Ideation Workshop: Brainstorm New Ideas

Now you're ready for the one-day ideation workshop. The goal of the ideation workshop is to employ the various business valuation techniques discussed in Chapter 7,

coupled with the client-specific envisioning exercise that you just developed using the client's data, to help the business stakeholders to brainstorm how these new sources of big data (both internal and external data sources) coupled with advanced analytics can provide unique insights for use with their targeted business initiative. You'll want to inspire the business stakeholders to envision how they might leverage internal and external data sources to help them:

- Answer the business questions they need to answer in support of the targeted business initiative. You'll want to challenge them to rethink the questions they ask of the business, and to contemplate the potential business impact of answering those questions at a lower level of granularity, with new metrics (gleaned from structured and unstructured data sources, both internal and external to the organization), and across more dimensions of the business.
- Make the decisions that are necessary to support the targeted business initiative. You'll want to challenge the business users to explore more detailed, timelier, and more robust decisions enabled by access to new sources of data, coupled with advanced analytics to uncover the drivers for each of the key decisions.

The ideation workshop will cover three key envisioning steps: brainstorming, prioritization, and documentation. A sample ideation workshop agenda is shown in Table 9-1.

Table 9-1: Ideation Workshop Agenda

Minutes	Workshop Section
15	Welcome and Introductions
30	Strategic Business Initiative Discussion Goal: Discuss targeted business initiatives including objectives, business drivers, key performance indicators, critical success factors, and timeline
30	Share Interview Findings Goal: Share interview findings and some initial insights and observations
45	Data Science/Advanced Analytics Envisioning Goal: Stimulate creative thinking regarding how advanced analytics could energize the targeted business initiative

Continues

Table 9-1: *(continued)*

Minutes	Workshop Section
60	Big data Opportunities Brainstorming Goal: Use envisioning techniques to brainstorm the use cases where big data could impact targeted business initiative
60	Big data Opportunities Prioritization Goal: Use Prioritization Matrix to drive group consensus on identified use cases
30	Summarize Workshop Findings and Define Next Steps Goal: Review the list of top priority use cases and gain consensus on next steps

Brainstorming

You will start the ideation workshop by brainstorming where and how to leverage big data—new sources of customer, product, and operational data coupled with advanced and predictive analytics—to power your targeted business initiative. You will review the client-specific envisioning exercise just developed to help the business stakeholders visualize what is possible with respect to new data sources and advanced analytics tools. You will demonstrate to the business and IT stakeholders how applying advanced analytics to their internal data, coupled with third-party data as appropriate, can provide new business insights and new monetization opportunities.

You will leverage the envisioning techniques outlined in Chapter 7 (such as big data envisioning worksheet, and Michael Porter's Five Forces and Value Chain Analysis methodologies) to brainstorm the business questions, ideas, and business decisions that can supercharge the targeted business initiative. You will need to track the ideas—for example, by recording them on individual Post-it notes—in the form of business questions or statements, such as "How do I identify our most engaged customer segments?" or "I want to see what baskets of products my gold card customers typically buy."

You will want to leverage the client-specific example, as well as examples from similar and other industries, to fuel the creative thought process with respect to how other organizations and other industries are leveraging big data to drive business value. Take time to review several scenarios that will help the workshop participants

envision where and how new sources of big data and advanced analytics could deliver financial and competitive value to the targeted business initiative.

The key is to challenge the group's current thinking processes and assumptions in an open, facilitated conversation. Ignite the creative processes by asking the participants to explore "what if" and "how might" thinking such as:

- *What if* I can get new insights into my customer shopping behaviors and product preferences, and *how might* that change my customer engagement opportunities?
- *What if* I had insights into my patients' current and historical lifestyles and diet patterns, and *how might* that impact my ability to diagnose their current health problems and prescribe more specific health changes?
- *What if* I knew which of my products were operating at the edges of acceptable performance, and *how might* those insights be used to improve maintenance scheduling, crew training, and inventory management?
- *What if* I knew the characteristics of my safest, most successful drivers, and *how might* those insights be used to change how I hire, train, and pay my most valuable drivers?

All of these business questions, statements, and ideas should be captured on individual Post-it notes. Capturing each of these questions, statements and ideas on a separate note is key to the grouping step that takes place next. The questions could look like the following for a client targeting a "churn reduction" business initiative:

- What customer segments are experiencing the most churn?
- Are there similarities in product usage patterns and propensities across my churning customers?
- What are the social characteristics of my highest churning customer segments?
- What are the common characteristics or usage patterns of customers who churn?
- Are there any customer segments that have experienced reduced churn?
- What marketing offers have we tested with high-probability churn customers?
- Who are our most profitable customers?
- Who are our most valuable customers?

It's not uncommon in the brainstorming session to capture 60, 80, or 120 different questions, statements, and ideas. Capture them all, and you'll sort and group them later.

Here is a list of some useful facilitation tips and techniques for managing the creative process during the facilitated, brainstorming session

- Hold the brainstorming session in a room that has an open feel to encourage open discussions and the open sharing of ideas. Explore options outside of the client's office, such as a hotel conference room or a partner's conference room. We once conducted a brainstorming session on a wind turbine farm, just to get the participants out of their comfort zone. So be creative.
- Minimize clutter by getting rid of tables and setting up chairs in a horseshoe style (avoid lecture hall or classroom set ups).
- Tape multiple flip charts on the walls around the room to capture ideas.
- Place a "parking lot" flip chart on the wall that can be used to capture discussions that may be interesting but threaten to derail the brainstorming process. It's a polite way of saying that you need to move on.
- Randomly place the Post-it notes on the multiple flip charts. Don't worry about grouping the Post-it notes as you place them on the flip charts. You'll use a grouping process in the next step in the Ideation Workshop process.
- Ensure that everyone works individually. When participants work in groups, it's not unusual that one person dominates the conversation and many good ideas from other group members never get recognized or recorded.
- Capture one idea per Post-it note. If you get a Post-it note with multiple questions, divide it into multiple Post-it notes.
- Read out loud what others have written as you place the Post-it notes on the flip charts. Reading the notes as you post them helps to fuel the creative thinking.
- Run the brainstorming session as long as anyone is still generating ideas. In fact, let silence work to your advantage by continuing to encourage folks to think of new questions, statements, and ideas.
- Give participants a heads-up that you'll stop capturing Post-it notes at 5-, 3-, and 1-minute intervals. Don't feel obligated to stick to those particular timeframes. Again, let the process run as long as it's productive.

Aggregation or Grouping

The goal of the aggregation or grouping step is to group the questions, statements, and ideas captured on the Post-it notes into common themes. Have the participants

huddle around the flip charts and look for common themes amongst the Post-it notes. Move the business questions and statements into common "themes" (use cases), for example, revenue analysis, customer up-sell, customer churn, and branch performance analysis. It is not unusual to have multiple Post-it notes that are very similar because many of the business stakeholders are asking the same questions, although they may use different metrics or dimensions. For example, the Sales department might want to see sales performance by sales reps and sales territories, while the Marketing department might want to see sales performance by campaign and promotion, and the Product Development department might want to see sales performance by products and product lines. Every group is interested in sales performance, just by different dimensions of the business.

Once you have established a "theme" and have grouped the common Post-it notes together around that theme, use a marker to draw a large circle around that group of Post-its and give it a label, such as customer acquisition, customer churn, or up-sell. Keep the title or description short (three- to four-word descriptions). Later in the documentation phase, you'll flush out the themes with a more descriptive title and more details gathered from the Post-it notes associated with that theme.

Typically, the targeted business initiative will break down into multiple (6 to 12) use cases. For example, "leverage customer behavioral insights to optimize the customer lifecycle engagement processes" might break down into the following "themes" or use cases:

- Reduce churn
- Most important customer segments
- Competitive churn benchmarks
- Product usage characteristics
- Network performance trends
- Customer acquisition
- Customer profiling and segmentation
- Package audience segments
- Location-based services

The end result of the brainstorming process will be several flip charts covered with Post-it notes with the common themes or use cases grouped together (see Figure 9-6).

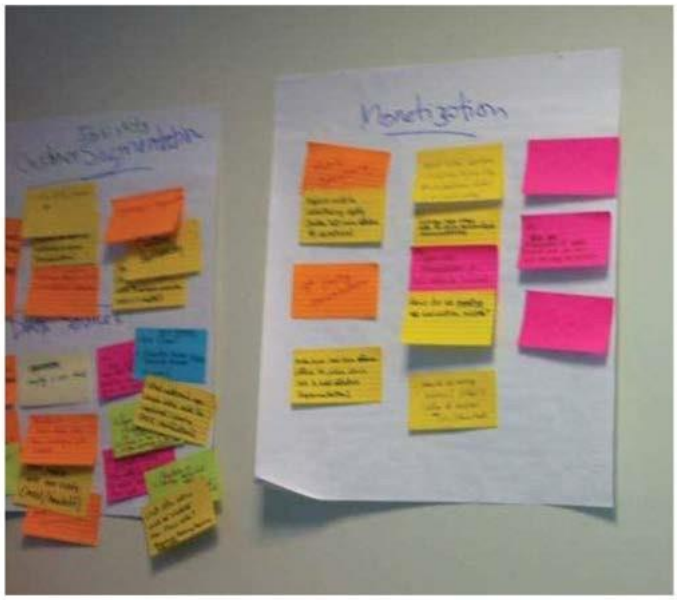


Figure 9-6: Using Post-it Notes for the brainstorming process

Finally, create a separate Post-it note for each identified theme or use case. These Post-it notes will be used in the prioritization exercise.

Step 4: Ideation Workshop: Prioritize Big Data Use Cases

Finally, you will guide the workshop participants through a prioritization process where each use case is judged based on its relative business value vis-à-vis its implementation feasibility. During this process, you will capture details regarding the business value drivers (for instance, why one business opportunity was valued more highly than another) and the reasons behind the feasibility determination (such as why one business opportunity is more difficult to implement than another). The end result of the prioritization process is a matrix like that shown in Figure 9-7.

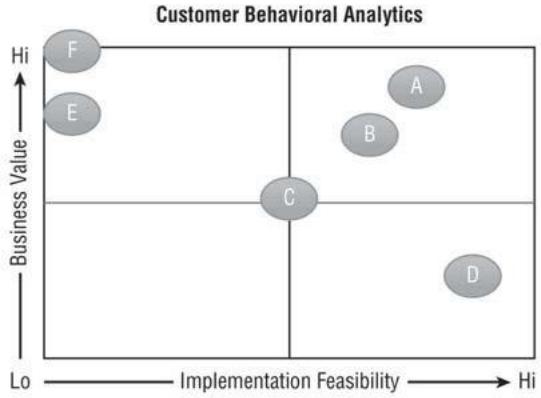


Figure 9-7: Sample prioritization results

Use Cases:

A. **Churn**: Couple smartphone app usage data with customer financial and demographic data to improve Churn Predictive Model Effectiveness
B. **Product Performance**: Drive changes to network bandwidth based upon customer's usage and customer profitability
C. **Network Optimization**: Optimize Network investments to reduce congestion based upon customers' app usage patterns
D. **Standardization**: Standardize tools, processes, analytic models, and hiring profiles across analytics teams
E. **Recommendations**: Create customer-specific product and service recommendations based upon their smartphone app usage patterns
F. **Monetization**: Leverage smartphone app usage data to drive new location-based services business opportunities

I will cover how to facilitate the prioritization process later in this chapter, as it is the key capstone activity of the ideation workshop, and turns all the prior research and brainstorming into an executable action plan.

Step 5: Document Next Steps

As the last step, you will summarize the identified and prioritized business opportunities, and recommend steps for deploying advanced analytics in support of the targeted business initiatives. You will document the results of the envisioning process which include:

- Key interview findings as related to the targeted business initiative including key business questions, business decisions, and required data sources

- Analytic use cases that came out of the brainstorming step
- The Prioritization Matrix results including details on the placement of each use case, business value drivers, and implementation risk items
- Recommended next steps

The final stage of the vision process workshop is a presentation of the findings and recommendations, as well, as the detailed insights from the envisioning exercise, to executive management. The findings and recommendations will confirm the relevance of big data to help drive the targeted business initiative and determine next steps for implementation.

The Prioritization Process

One key challenge to a successful big data journey is gaining consensus and alignment between the business and IT stakeholders in identifying the initial big data business use cases that deliver sufficient value to the business, while possessing a high probability of success. One can find multiple business use cases where big data and advanced analytics can deliver compelling business value. However, many of these use cases have a low probability of execution success due to:

- Unavailability of timely, accurate data
- Lack of experience with new data sources like social media, mobile, logs, and telemetry data
- Limited data science or advanced analytics resources or skills
- Lack of experience with new technologies like Hadoop, MapReduce, and text mining
- Architectural and technology limitations with managing and analyzing unstructured data, and ingesting and analyzing real-time data feeds
- Weak working relationship between the business and IT teams
- Lack of management fortitude and support

I have found one tool for driving business and IT collaboration and agreement around identifying the right initial use cases for your big data journey—those with sufficient business value and a high probability of success. This tool is the *Prioritization Matrix*. Let me share how the Prioritization Matrix works to not only prioritize

the initial big data use cases, but how to use it to foster an atmosphere of collaboration between the business and IT stakeholders.

The prioritization process is the single most important step in the envisioning process. While I expect that most readers would think the brainstorming process is the most important, the truth is that many use cases are probably already known ahead of the brainstorming session. The brainstorming session is useful in validating and expanding on those known use cases and helping to fuel the identification of additional use cases.

But if you cannot gain group consensus on the right use cases on which to start your big data initiative, then the big data initiative has a greatly diminished chance of success. To be successful, the big data initiative needs the initial support *and on-going leadership* of both business and IT stakeholders in order to drive the potential business transformation. Let's start the prioritization process lesson by first understanding the mechanics of the Prioritization Matrix.

The Prioritization Matrix is a 2×2 grid that facilitates the interactive process and debate between the business and IT stakeholders to determine where on the matrix to place each use case in relation to the other use cases. The use cases are placed on the matrix based on:

- **Business value**: the vertical axis of the matrix. The business stakeholders are typically responsible for the relative positioning of each business use case on the Business Value axis. The Business Value axis reads from low business value at the bottom to high business value at the top as shown in Figure 9-8.
- **Implementation feasibility**: the probability of a successful implementation considering availability, granularity and timeliness of data, skills, tools, organizational readiness, and needed experience. Implementation feasibility is the horizontal axis of the matrix. The IT stakeholders are typically responsible for the relative positioning of each business use case on the Implementation Feasibility axis. The Implementation Feasibility axis reads from low implementation feasibility on the left (higher probability of failure) to high implementation feasibility on the right (higher probability of success).

As a reminder, you are not looking for the exact valuation of each use case from a Business Value perspective. Instead, you want to know the relative business value of each use case and some level of justification from the business stakeholders as to the reasoning behind the placement of the use case.

Figure 9-8: The Prioritization Matrix

The Prioritization Matrix Process

Focusing the Prioritization Matrix process on a key business initiative—such as reducing churn, increasing same store sales, minimizing financial risk, optimizing market spend, or reducing hospital readmissions—is critical as it provides the foundation upon which the business value and implementation feasibility discussion can occur.

The Prioritization Matrix process starts by placing each use case identified in the brainstorming and aggregation stages on a Post-it note (one use case per Post-it). The group, which must include both business and IT stakeholders, decides the placement of each use case on the Prioritization Matrix by weighing business value and implementation feasibility, vis-à-vis the relative placement of the other use cases on the matrix.

The business stakeholders are responsible for the relative positioning of each business case on the Business Value axis, while the IT stakeholders are primarily responsible for the relative positioning of each business case on the Implementation Feasibility axis (considering data, technology, skills, and organizational readiness).

The heart of the prioritization process is the discussion that ensues about the relative placement of each of the use cases (see Figure 9-10), such as:

- Why is use case [B] more or less valuable than use case [A]? What are the specific business drivers or variables that make use case [B] more or less valuable than use case [A]? (See Figure 9-9.)
- Why is use case [B] less or more feasible from an implementation perspective than use case [A]? What are the specific implementation risks that make use case [B] less or more feasible than use case [A]?

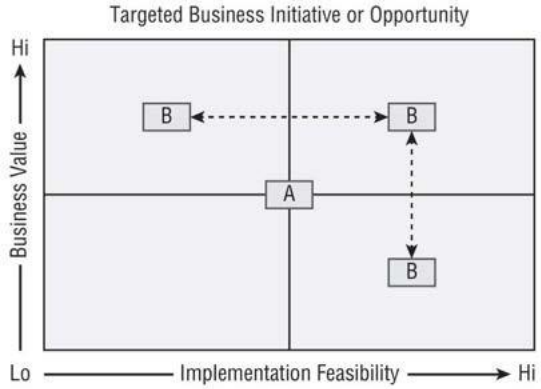


Figure 9-9: Prioritization process drives group alignment

It is critical to the prioritization process to capture the reasons for the relative positioning of each use case, in order to identify the critical business value drivers and potential implementation risks.

Prioritization Matrix Traps

One of the keys to affectively using the Prioritization Matrix is to understand the potential discussion traps and to guide the workshop participants around those traps. In particular, you want to avoid use cases that fall into the following matrix zones (see Figure 9-10):

- "Zone of Mismanaged Expectations" are those use cases with huge business value but little chance of successful execution (for example, solve world

1 Chapter 9

hunger). It is not uncommon for a senior executive to have a pet project that is grand in vision and scale. The Prioritization Matrix will highlight the specific reasons why that might be a poor use case against which to start your big data journey. The Prioritization Matrix process will also highlight what steps need to be taken to move the use case into a more highly feasible situation.

- "Zone of User Disillusionment" are those use cases which are easy to execute but provide little business value. These types of use cases tend to be technological science experiments, where the IT group has developed some skills in a new technology or has gained access to some new data sources and are desperately trying to find a use case against which to apply their new capabilities. Don't go there. While there is always room within IT for experiments in order to develop more knowledge and experience, don't make your business stakeholders guinea pigs in those experiments.
- "Zone of Career-limiting Moves" are those use cases that have little business value and have a low probability of success. These sorts of use cases should be self-evident and no one on either the business or IT sides of the room should want to target one of these use cases.

Figure 9-10: Prioritization Matrix traps

Use cases that fall into one of these zones should be avoided because they either don't provide enough business value to be meaningful to the business stakeholders, or are too risky to IT from an implementation perspective.

It is important to note that understanding where each use case falls, and the open discussion between the business and IT stakeholders about why each use case is positioned where it is, is key to understanding the implementation and business risks and avoiding surprises once the project is implemented—Eyes wide open!

Finally, the end result of the Prioritization Matrix process will look something like that shown in Figure 9-11. All the use cases have been placed on the Prioritization Matrix and justification for both the business value and implementation feasibility discussed and agreed upon. The use cases in the upper-right quadrant of Figure 9-11 end up being the "low-hanging fruit" for your initial big data engagement.

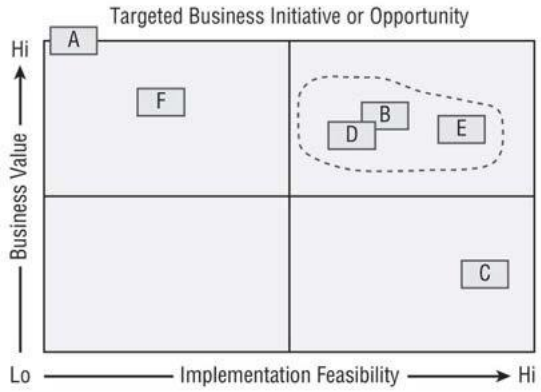


Figure 9-11: Prioritization Matrix end result

The prioritization matrix is a marvelous tool for facilitating a conversation between the business and IT stakeholders about where and how to start the big data journey. It provides a framework for identifying the relative business value of each business use case (with respect to the targeted business initiative) and for identifying and understanding the implementation risks. Out of this prioritization process, both the business and IT stakeholders should know what use cases they are targeting and the potential business value of each use case. Participants also have their eyes wide open to the implementation risks that the project needs to avoid or manage.

Using User Experience Mockups to Fuel the Envisioning Process

Developing simple user experience mockups is a powerful way to help the business users "envision the realm of what's possible." Organizations can combine big data concepts with user experience mockups to help break out of their current mental boxes—to think differently—and identify new ways that big data can power the organization's value creation processes. The new customer, product, and operational

insights gathered from these mockups can also help identify new revenue or monetization opportunities. Let's review a few examples of how a simple mockup can help drive the envisioning process.

The following example takes an organization's website or mobile apps and poses some challenging questions about how the organization could improve the website or app to drive a more engaging customer experience. The mockup shows a credit union that released a smartphone app to support their new "MyBranch" customer engagement initiative (see Figure 9-12). (Note: All of the information used to create this mockup was retrieved from the credit union's public-facing website.) The new smartphone app supports the following customer transactions:

- View current and available balances across all the customers' accounts
- Transfer funds between accounts or make loan payments
- View transaction history and access details on specific transactions
- Electronically pay bills anywhere and anytime
- Get directions to the nearest branch or ATM
- Set alerts on account balances, debit card transactions, and withdrawals

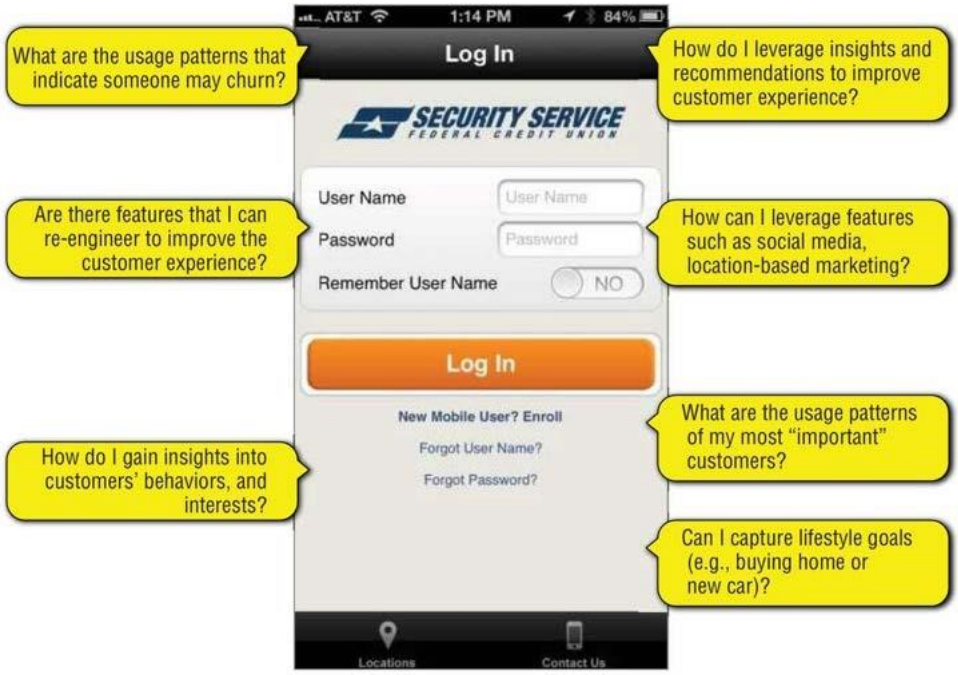


Figure 9-12: Mobile app functionality mockup

These customer transactions are a ripe source of customer insights and product preferences that can be mined to provide a more compelling and relevant user experience. That same user experience can also yield new customer and product insights that can be converted into new monetization opportunities such as new services and products. With this mockup in hand, the business users can now be taken through a series of envisioning exercises to explore and brainstorm the following types of questions (see Figure 9-13):

- What are the usage patterns of my most valuable customers?
- What are the usage patterns that indicate someone may be churning?
- How do we leverage personalized insights and previous activities to improve the customer experience?
- How can we provide additional features, such as social media, to capture more information about our customers' interests, passions, associations, and affiliations?
- How can we leverage these insights, coupled with the GPS features of our smartphone apps, to offer location-based customer services?
- How can we leverage recommendations to enhance the customer experience?
- How can we capture lifestyle goals, such as saving to buy a home or new car?
- Are there instrumentation opportunities we can use to gain insights into our customers' behaviors, preferences, and interests?
- Are their combinations of features that we can re-engineer to improve the customer experience?

I hope you can realize how powerful even simple mockups can be in helping the vision workshop business and IT stakeholders identify how big data can power an improved customer experience and uncover new monetization opportunities. A simple customer experience mockup can bring to life the potential of big data to:

- Identify additional opportunities to capture customer usage and product preference data through additional instrumentation of the website and smartphone apps
- Leverage advanced analytics to uncover customer-specific insights, recommendations, and benchmarks to power a more relevant and compelling user experience

- Leverage experimentation techniques to tease out more customer and product insights by presenting different recommendations to see which audiences respond to which offers and recommendations

The mockup in Figure 9-13 is a little more advanced and explores how a cellular provider could leverage their subscribers' app usage data to improve the subscriber's user experience—make the experience more relevant and actionable—in order to improve customer engagement processes and uncover new monetization opportunities.

This example evaluates how a cellular phone company could leverage a subscriber's app usage data, and the app usage behaviors of similar customers, to develop personalized e-mail recommendations that might be beneficial to that subscriber. In the process, the cellular provider will learn more about their subscribers' preferences—what they like and what they don't like—that can yield even more subscriber and product insights. This is a counter-example to the "unintelligent" user experience presented in Chapter 8.

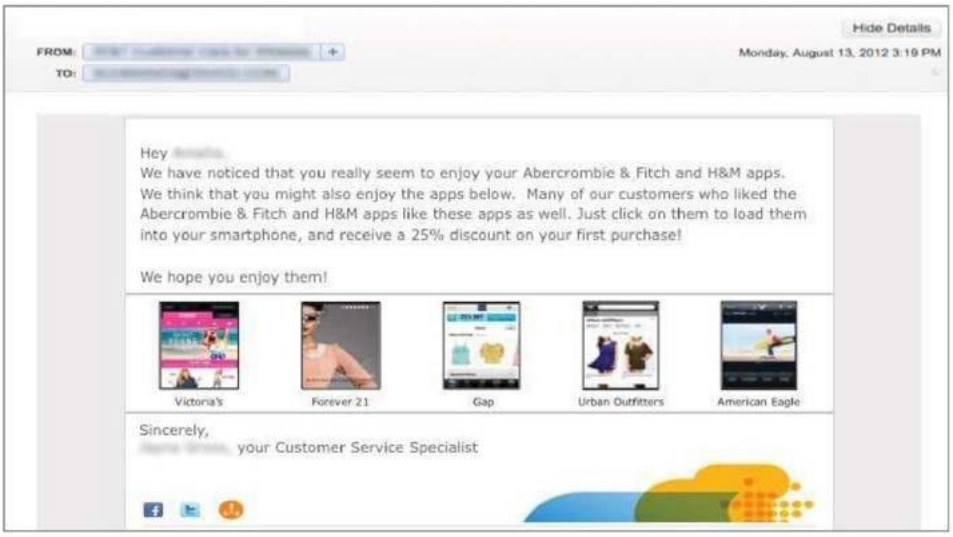


Figure 9-13: Leverage apps usage data to improve a subscriber's user experience

There are numerous "what if" questions that could fuel the brainstorming process for this mockup, such as:

- What if we could leverage our subscribers' app usage patterns to recommend apps that move the user into a more profitable, high-retention usage category

(for example, from "Moderate Female Teenage Browser" to "Female Teenage Shopaholic")?

- What if we could score the customers' app usage patterns to identify and act on potential churn situations more quickly?
- What if we could integrate app performance data across our subscriber base to recommend apps that provide a superior customer experience and help create more loyalty to the cellular provider?
- What if we could aggregate subscriber app usage insights across the entire network to create new monetization opportunities, such as app developer referral fees and co-marketing fees?
- What if we could integrate customers' app usage insights with real-time GPS location information to offer personalized location-based services?

Creating mockups is an effective technique for fueling the creative thinking process during the ideation workshop. Don't be concerned about the professional level of mockup (my mockups look like I drew them with a crayon). It's more important that the mockups challenge the current conventional thinking of the business stakeholders. The mockups can push the business stakeholders out of their current thinking ruts to contemplate the realm of what might be possible by leveraging all their customer and product insights to optimize existing customer engagement processes and uncover new monetization opportunities.

Summary

This chapter reviewed in detail the vision workshop or envisioning process. I described each of the five steps in the vision workshop methodology and provided details on each step using real-world examples.

You spent quite a bit of time on the data preparation and analysis work required to transform business initiative-specific data into an envisioning exercise that can be used as part of the ideation workshop. This is an important part of the vision workshop methodology because it helps the envisioning process *come to life* for the workshop participants. I provided several examples of creating customer-specific envisioning exercises.

You learned about the brainstorming and aggregation process of the ideation workshop. You also reviewed how to use the Michael Porter value creation processes—Value Chain and Five Forces Analysis—as well as the business initiative-specific envisioning exercise to tease out new business opportunities as part of the envisioning process.

Unit 5

In this example, something happened on May 25 and 26 that warrants additional investigation. Getting the participants to start brainstorming about what might have happened before or on those days (perhaps company news, competitive activities, market news, or economic news) is a good starting point for the envisioning work that will occur in the ideation workshop.

There are also many external data sources that can be coupled with the organization's data to provide new perspectives on that same old client data. For example, `www.data.gov` is a valuable source of data covering a wide range of information sources which could be used to help the business users start envisioning what's possible. Figure 9-5 shows an example of integrating government-provided Consumer Price Index (CPI) data with the organization's customer sales data to ascertain if there are customer segments in which the organization's marketing spend is overcommitted or under-committed given the market segment potential.

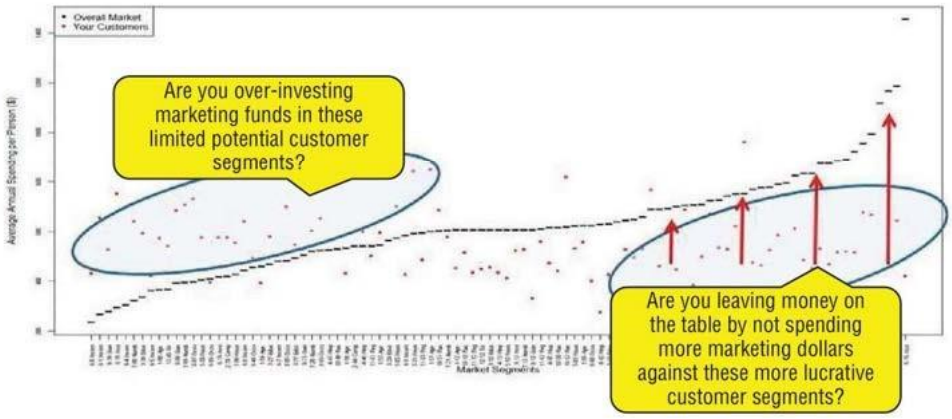


Figure 9-5: Comparison of Markets vs. Your Customers' Spending by Market Segments

Again, the objective of analyzing a small chunk of the organization's data is to personalize the linkage between the business stakeholders and the big data opportunity. You want to fuel the creative thought process to help the business stakeholders explore the realm of what might be possible if the business users had access to new customer, product, and operational insights that can be leveraged as part of their day-to-day business processes.

Step 3: Ideation Workshop: Brainstorm New Ideas

Now you're ready for the one-day ideation workshop. The goal of the ideation workshop is to employ the various business valuation techniques discussed in Chapter 7,

coupled with the client-specific envisioning exercise that you just developed using the client's data, to help the business stakeholders to brainstorm how these new sources of big data (both internal and external data sources) coupled with advanced analytics can provide unique insights for use with their targeted business initiative. You'll want to inspire the business stakeholders to envision how they might leverage internal and external data sources to help them:

- Answer the business questions they need to answer in support of the targeted business initiative. You'll want to challenge them to rethink the questions they ask of the business, and to contemplate the potential business impact of answering those questions at a lower level of granularity, with new metrics (gleaned from structured and unstructured data sources, both internal and external to the organization), and across more dimensions of the business.
- Make the decisions that are necessary to support the targeted business initiative. You'll want to challenge the business users to explore more detailed, timelier, and more robust decisions enabled by access to new sources of data, coupled with advanced analytics to uncover the drivers for each of the key decisions.

The ideation workshop will cover three key envisioning steps: brainstorming, prioritization, and documentation. A sample ideation workshop agenda is shown in Table 9-1.

Table 9-1: Ideation Workshop Agenda

Minutes	Workshop Section
15	Welcome and Introductions
30	Strategic Business Initiative Discussion Goal: Discuss targeted business initiatives including objectives, business drivers, key performance indicators, critical success factors, and timeline
30	Share Interview Findings Goal: Share interview findings and some initial insights and observations
45	Data Science/Advanced Analytics Envisioning Goal: Stimulate creative thinking regarding how advanced analytics could energize the targeted business initiative

Continues

Table 9-1: *(continued)*

Minutes	Workshop Section
60	Big data Opportunities Brainstorming Goal: Use envisioning techniques to brainstorm the use cases where big data could impact targeted business initiative
60	Big data Opportunities Prioritization Goal: Use Prioritization Matrix to drive group consensus on identified use cases
30	Summarize Workshop Findings and Define Next Steps Goal: Review the list of top priority use cases and gain consensus on next steps

Brainstorming

You will start the ideation workshop by brainstorming where and how to leverage big data—new sources of customer, product, and operational data coupled with advanced and predictive analytics—to power your targeted business initiative. You will review the client-specific envisioning exercise just developed to help the business stakeholders visualize what is possible with respect to new data sources and advanced analytics tools. You will demonstrate to the business and IT stakeholders how applying advanced analytics to their internal data, coupled with third-party data as appropriate, can provide new business insights and new monetization opportunities.

You will leverage the envisioning techniques outlined in Chapter 7 (such as big data envisioning worksheet, and Michael Porter's Five Forces and Value Chain Analysis methodologies) to brainstorm the business questions, ideas, and business decisions that can supercharge the targeted business initiative. You will need to track the ideas—for example, by recording them on individual Post-it notes—in the form of business questions or statements, such as "How do I identify our most engaged customer segments?" or "I want to see what baskets of products my gold card customers typically buy."

You will want to leverage the client-specific example, as well as examples from similar and other industries, to fuel the creative thought process with respect to how other organizations and other industries are leveraging big data to drive business value. Take time to review several scenarios that will help the workshop participants

envision where and how new sources of big data and advanced analytics could deliver financial and competitive value to the targeted business initiative.

The key is to challenge the group's current thinking processes and assumptions in an open, facilitated conversation. Ignite the creative processes by asking the participants to explore "what if" and "how might" thinking such as:

- *What if* I can get new insights into my customer shopping behaviors and product preferences, and *how might* that change my customer engagement opportunities?
- *What if* I had insights into my patients' current and historical lifestyles and diet patterns, and *how might* that impact my ability to diagnose their current health problems and prescribe more specific health changes?
- *What if* I knew which of my products were operating at the edges of acceptable performance, and *how might* those insights be used to improve maintenance scheduling, crew training, and inventory management?
- *What if* I knew the characteristics of my safest, most successful drivers, and *how might* those insights be used to change how I hire, train, and pay my most valuable drivers?

All of these business questions, statements, and ideas should be captured on individual Post-it notes. Capturing each of these questions, statements and ideas on a separate note is key to the grouping step that takes place next. The questions could look like the following for a client targeting a "churn reduction" business initiative:

- What customer segments are experiencing the most churn?
- Are there similarities in product usage patterns and propensities across my churning customers?
- What are the social characteristics of my highest churning customer segments?
- What are the common characteristics or usage patterns of customers who churn?
- Are there any customer segments that have experienced reduced churn?
- What marketing offers have we tested with high-probability churn customers?
- Who are our most profitable customers?
- Who are our most valuable customers?

It's not uncommon in the brainstorming session to capture 60, 80, or 120 different questions, statements, and ideas. Capture them all, and you'll sort and group them later.

t

Here is a list of some useful facilitation tips and techniques for managing the creative process during the facilitated, brainstorming session

- Hold the brainstorming session in a room that has an open feel to encourage open discussions and the open sharing of ideas. Explore options outside of the client's office, such as a hotel conference room or a partner's conference room. We once conducted a brainstorming session on a wind turbine farm, just to get the participants out of their comfort zone. So be creative.
- Minimize clutter by getting rid of tables and setting up chairs in a horseshoe style (avoid lecture hall or classroom set ups).
- Tape multiple flip charts on the walls around the room to capture ideas.
- Place a "parking lot" flip chart on the wall that can be used to capture discussions that may be interesting but threaten to derail the brainstorming process. It's a polite way of saying that you need to move on.
- Randomly place the Post-it notes on the multiple flip charts. Don't worry about grouping the Post-it notes as you place them on the flip charts. You'll use a grouping process in the next step in the Ideation Workshop process.
- Ensure that everyone works individually. When participants work in groups, it's not unusual that one person dominates the conversation and many good ideas from other group members never get recognized or recorded.
- Capture one idea per Post-it note. If you get a Post-it note with multiple questions, divide it into multiple Post-it notes.
- Read out loud what others have written as you place the Post-it notes on the flip charts. Reading the notes as you post them helps to fuel the creative thinking.
- Run the brainstorming session as long as anyone is still generating ideas. In fact, let silence work to your advantage by continuing to encourage folks to think of new questions, statements, and ideas.
- Give participants a heads-up that you'll stop capturing Post-it notes at 5-, 3-, and 1-minute intervals. Don't feel obligated to stick to those particular timeframes. Again, let the process run as long as it's productive.

Aggregation or Grouping

The goal of the aggregation or grouping step is to group the questions, statements, and ideas captured on the Post-it notes into common themes. Have the participants

huddle around the flip charts and look for common themes amongst the Post-it notes. Move the business questions and statements into common "themes" (use cases), for example, revenue analysis, customer up-sell, customer churn, and branch performance analysis. It is not unusual to have multiple Post-it notes that are very similar because many of the business stakeholders are asking the same questions, although they may use different metrics or dimensions. For example, the Sales department might want to see sales performance by sales reps and sales territories, while the Marketing department might want to see sales performance by campaign and promotion, and the Product Development department might want to see sales performance by products and product lines. Every group is interested in sales performance, just by different dimensions of the business.

Once you have established a "theme" and have grouped the common Post-it notes together around that theme, use a marker to draw a large circle around that group of Post-its and give it a label, such as customer acquisition, customer churn, or up-sell. Keep the title or description short (three- to four-word descriptions). Later in the documentation phase, you'll flush out the themes with a more descriptive title and more details gathered from the Post-it notes associated with that theme.

Typically, the targeted business initiative will break down into multiple (6 to 12) use cases. For example, "leverage customer behavioral insights to optimize the customer lifecycle engagement processes" might break down into the following "themes" or use cases:

- Reduce churn
- Most important customer segments
- Competitive churn benchmarks
- Product usage characteristics
- Network performance trends
- Customer acquisition
- Customer profiling and segmentation
- Package audience segments
- Location-based services

The end result of the brainstorming process will be several flip charts covered with Post-it notes with the common themes or use cases grouped together (see Figure 9-6).

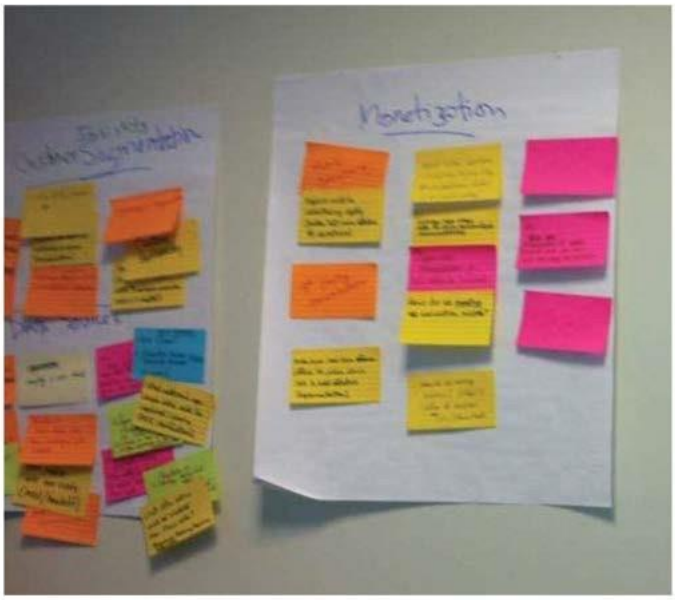


Figure 9-6: Using Post-it Notes for the brainstorming process

Finally, create a separate Post-it note for each identified theme or use case. These Post-it notes will be used in the prioritization exercise.

Step 4: Ideation Workshop: Prioritize Big Data Use Cases

Finally, you will guide the workshop participants through a prioritization process where each use case is judged based on its relative business value vis-à-vis its implementation feasibility. During this process, you will capture details regarding the business value drivers (for instance, why one business opportunity was valued more highly than another) and the reasons behind the feasibility determination (such as why one business opportunity is more difficult to implement than another). The end result of the prioritization process is a matrix like that shown in Figure 9-7.

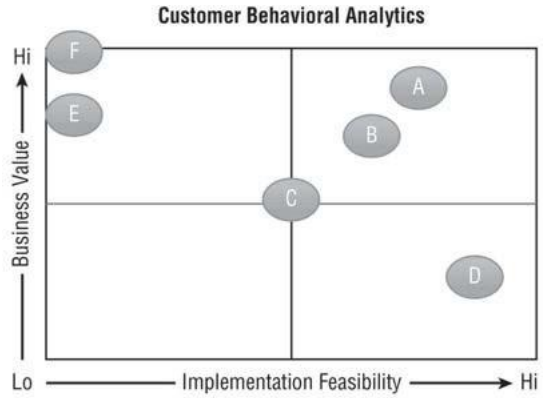


Figure 9-7: Sample prioritization results

Use Cases:

A. **Churn**: Couple smartphone app usage data with customer financial and demographic data to improve Churn Predictive Model Effectiveness
B. **Product Performance**: Drive changes to network bandwidth based upon customer's usage and customer profitability
C. **Network Optimization**: Optimize Network investments to reduce congestion based upon customers' app usage patterns
D. **Standardization**: Standardize tools, processes, analytic models, and hiring profiles across analytics teams
E. **Recommendations**: Create customer-specific product and service recommendations based upon their smartphone app usage patterns
F. **Monetization**: Leverage smartphone app usage data to drive new location-based services business opportunities

I will cover how to facilitate the prioritization process later in this chapter, as it is the key capstone activity of the ideation workshop, and turns all the prior research and brainstorming into an executable action plan.

Step 5: Document Next Steps

As the last step, you will summarize the identified and prioritized business opportunities, and recommend steps for deploying advanced analytics in support of the targeted business initiatives. You will document the results of the envisioning process which include:

- Key interview findings as related to the targeted business initiative including key business questions, business decisions, and required data sources

- Analytic use cases that came out of the brainstorming step
- The Prioritization Matrix results including details on the placement of each use case, business value drivers, and implementation risk items
- Recommended next steps

The final stage of the vision process workshop is a presentation of the findings and recommendations, as well, as the detailed insights from the envisioning exercise, to executive management. The findings and recommendations will confirm the relevance of big data to help drive the targeted business initiative and determine next steps for implementation.

The Prioritization Process

One key challenge to a successful big data journey is gaining consensus and alignment between the business and IT stakeholders in identifying the initial big data business use cases that deliver sufficient value to the business, while possessing a high probability of success. One can find multiple business use cases where big data and advanced analytics can deliver compelling business value. However, many of these use cases have a low probability of execution success due to:

- Unavailability of timely, accurate data
- Lack of experience with new data sources like social media, mobile, logs, and telemetry data
- Limited data science or advanced analytics resources or skills
- Lack of experience with new technologies like Hadoop, MapReduce, and text mining
- Architectural and technology limitations with managing and analyzing unstructured data, and ingesting and analyzing real-time data feeds
- Weak working relationship between the business and IT teams
- Lack of management fortitude and support

I have found one tool for driving business and IT collaboration and agreement around identifying the right initial use cases for your big data journey—those with sufficient business value and a high probability of success. This tool is the *Prioritization Matrix*. Let me share how the Prioritization Matrix works to not only prioritize

the initial big data use cases, but how to use it to foster an atmosphere of collaboration between the business and IT stakeholders.

The prioritization process is the single most important step in the envisioning process. While I expect that most readers would think the brainstorming process is the most important, the truth is that many use cases are probably already known ahead of the brainstorming session. The brainstorming session is useful in validating and expanding on those known use cases and helping to fuel the identification of additional use cases.

But if you cannot gain group consensus on the right use cases on which to start your big data initiative, then the big data initiative has a greatly diminished chance of success. To be successful, the big data initiative needs the initial support *and on-going leadership* of both business and IT stakeholders in order to drive the potential business transformation. Let's start the prioritization process lesson by first understanding the mechanics of the Prioritization Matrix.

The Prioritization Matrix is a 2×2 grid that facilitates the interactive process and debate between the business and IT stakeholders to determine where on the matrix to place each use case in relation to the other use cases. The use cases are placed on the matrix based on:

- **Business value**: the vertical axis of the matrix. The business stakeholders are typically responsible for the relative positioning of each business use case on the Business Value axis. The Business Value axis reads from low business value at the bottom to high business value at the top as shown in Figure 9-8.
- **Implementation feasibility**: the probability of a successful implementation considering availability, granularity and timeliness of data, skills, tools, organizational readiness, and needed experience. Implementation feasibility is the horizontal axis of the matrix. The IT stakeholders are typically responsible for the relative positioning of each business use case on the Implementation Feasibility axis. The Implementation Feasibility axis reads from low implementation feasibility on the left (higher probability of failure) to high implementation feasibility on the right (higher probability of success).

As a reminder, you are not looking for the exact valuation of each use case from a Business Value perspective. Instead, you want to know the relative business value of each use case and some level of justification from the business stakeholders as to the reasoning behind the placement of the use case.

2

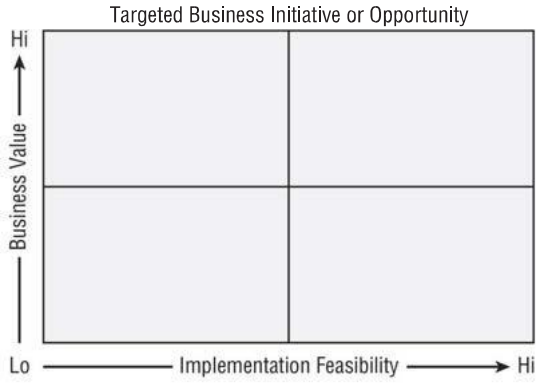


Figure 9-8: The Prioritization Matrix

The Prioritization Matrix Process

Focusing the Prioritization Matrix process on a key business initiative—such as reducing churn, increasing same store sales, minimizing financial risk, optimizing market spend, or reducing hospital readmissions—is critical as it provides the foundation upon which the business value and implementation feasibility discussion can occur.

The Prioritization Matrix process starts by placing each use case identified in the brainstorming and aggregation stages on a Post-it note (one use case per Post-it). The group, which must include both business and IT stakeholders, decides the placement of each use case on the Prioritization Matrix by weighing business value and implementation feasibility, vis-à-vis the relative placement of the other use cases on the matrix.

The business stakeholders are responsible for the relative positioning of each business case on the Business Value axis, while the IT stakeholders are primarily responsible for the relative positioning of each business case on the Implementation Feasibility axis (considering data, technology, skills, and organizational readiness).

The heart of the prioritization process is the discussion that ensues about the relative placement of each of the use cases (see Figure 9-10), such as:

- Why is use case [B] more or less valuable than use case [A]? What are the specific business drivers or variables that make use case [B] more or less valuable than use case [A]? (See Figure 9-9.)
- Why is use case [B] less or more feasible from an implementation perspective than use case [A]? What are the specific implementation risks that make use case [B] less or more feasible than use case [A]?

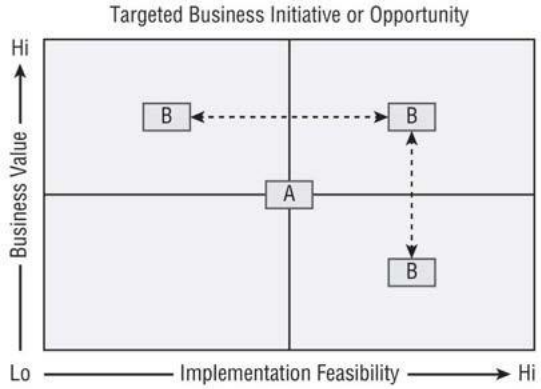


Figure 9-9: Prioritization process drives group alignment

It is critical to the prioritization process to capture the reasons for the relative positioning of each use case, in order to identify the critical business value drivers and potential implementation risks.

Prioritization Matrix Traps

One of the keys to affectively using the Prioritization Matrix is to understand the potential discussion traps and to guide the workshop participants around those traps. In particular, you want to avoid use cases that fall into the following matrix zones (see Figure 9-10):

- "Zone of Mismanaged Expectations" are those use cases with huge business value but little chance of successful execution (for example, solve world

hunger). It is not uncommon for a senior executive to have a pet project that is grand in vision and scale. The Prioritization Matrix will highlight the specific reasons why that might be a poor use case against which to start your big data journey. The Prioritization Matrix process will also highlight what steps need to be taken to move the use case into a more highly feasible situation.

- "Zone of User Disillusionment" are those use cases which are easy to execute but provide little business value. These types of use cases tend to be technological science experiments, where the IT group has developed some skills in a new technology or has gained access to some new data sources and are desperately trying to find a use case against which to apply their new capabilities. Don't go there. While there is always room within IT for experiments in order to develop more knowledge and experience, don't make your business stakeholders guinea pigs in those experiments.
- "Zone of Career-limiting Moves" are those use cases that have little business value and have a low probability of success. These sorts of use cases should be self-evident and no one on either the business or IT sides of the room should want to target one of these use cases.

Figure 9-10: Prioritization Matrix traps

Use cases that fall into one of these zones should be avoided because they either don't provide enough business value to be meaningful to the business stakeholders, or are too risky to IT from an implementation perspective.

It is important to note that understanding where each use case falls, and the open discussion between the business and IT stakeholders about why each use case is positioned where it is, is key to understanding the implementation and business risks and avoiding surprises once the project is implemented—Eyes wide open!

Finally, the end result of the Prioritization Matrix process will look something like that shown in Figure 9-11. All the use cases have been placed on the Prioritization Matrix and justification for both the business value and implementation feasibility discussed and agreed upon. The use cases in the upper-right quadrant of Figure 9-11 end up being the "low-hanging fruit" for your initial big data engagement.

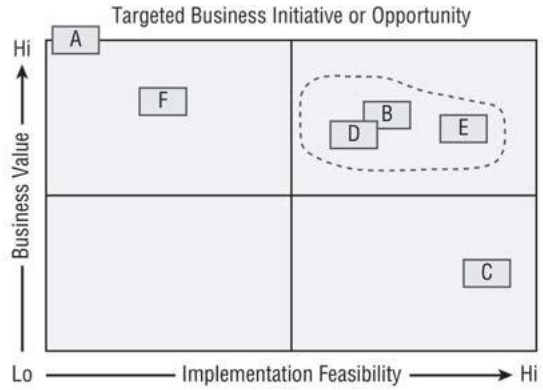


Figure 9-11: Prioritization Matrix end result

The prioritization matrix is a marvelous tool for facilitating a conversation between the business and IT stakeholders about where and how to start the big data journey. It provides a framework for identifying the relative business value of each business use case (with respect to the targeted business initiative) and for identifying and understanding the implementation risks. Out of this prioritization process, both the business and IT stakeholders should know what use cases they are targeting and the potential business value of each use case. Participants also have their eyes wide open to the implementation risks that the project needs to avoid or manage.

Using User Experience Mockups to Fuel the Envisioning Process

Developing simple user experience mockups is a powerful way to help the business users "envision the realm of what's possible." Organizations can combine big data concepts with user experience mockups to help break out of their current mental boxes—to think differently—and identify new ways that big data can power the organization's value creation processes. The new customer, product, and operational

insights gathered from these mockups can also help identify new revenue or monetization opportunities. Let's review a few examples of how a simple mockup can help drive the envisioning process.

The following example takes an organization's website or mobile apps and poses some challenging questions about how the organization could improve the website or app to drive a more engaging customer experience. The mockup shows a credit union that released a smartphone app to support their new "MyBranch" customer engagement initiative (see Figure 9-12). (Note: All of the information used to create this mockup was retrieved from the credit union's public-facing website.) The new smartphone app supports the following customer transactions:

- View current and available balances across all the customers' accounts
- Transfer funds between accounts or make loan payments
- View transaction history and access details on specific transactions
- Electronically pay bills anywhere and anytime
- Get directions to the nearest branch or ATM
- Set alerts on account balances, debit card transactions, and withdrawals

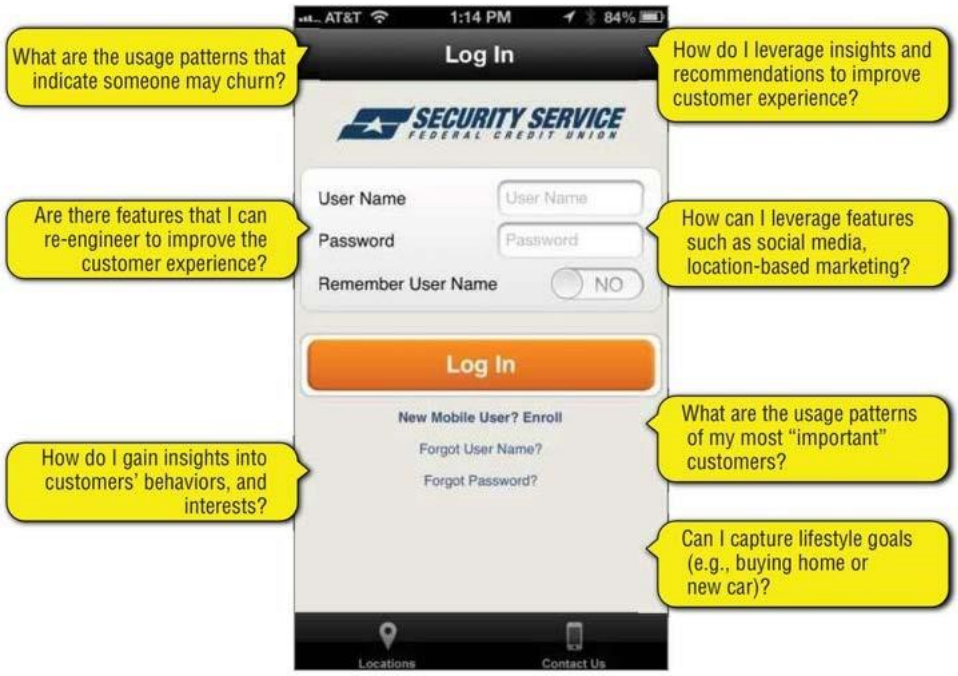


Figure 9-12: Mobile app functionality mockup

These customer transactions are a ripe source of customer insights and product preferences that can be mined to provide a more compelling and relevant user experience. That same user experience can also yield new customer and product insights that can be converted into new monetization opportunities such as new services and products. With this mockup in hand, the business users can now be taken through a series of envisioning exercises to explore and brainstorm the following types of questions (see Figure 9-13):

- What are the usage patterns of my most valuable customers?
- What are the usage patterns that indicate someone may be churning?
- How do we leverage personalized insights and previous activities to improve the customer experience?
- How can we provide additional features, such as social media, to capture more information about our customers' interests, passions, associations, and affiliations?
- How can we leverage these insights, coupled with the GPS features of our smartphone apps, to offer location-based customer services?
- How can we leverage recommendations to enhance the customer experience?
- How can we capture lifestyle goals, such as saving to buy a home or new car?
- Are there instrumentation opportunities we can use to gain insights into our customers' behaviors, preferences, and interests?
- Are their combinations of features that we can re-engineer to improve the customer experience?

I hope you can realize how powerful even simple mockups can be in helping the vision workshop business and IT stakeholders identify how big data can power an improved customer experience and uncover new monetization opportunities. A simple customer experience mockup can bring to life the potential of big data to:

- Identify additional opportunities to capture customer usage and product preference data through additional instrumentation of the website and smartphone apps
- Leverage advanced analytics to uncover customer-specific insights, recommendations, and benchmarks to power a more relevant and compelling user experience

- Leverage experimentation techniques to tease out more customer and product insights by presenting different recommendations to see which audiences respond to which offers and recommendations

The mockup in Figure 9-13 is a little more advanced and explores how a cellular provider could leverage their subscribers' app usage data to improve the subscriber's user experience—make the experience more relevant and actionable—in order to improve customer engagement processes and uncover new monetization opportunities.

This example evaluates how a cellular phone company could leverage a subscriber's app usage data, and the app usage behaviors of similar customers, to develop personalized e-mail recommendations that might be beneficial to that subscriber. In the process, the cellular provider will learn more about their subscribers' preferences—what they like and what they don't like—that can yield even more subscriber and product insights. This is a counter-example to the "unintelligent" user experience presented in Chapter 8.

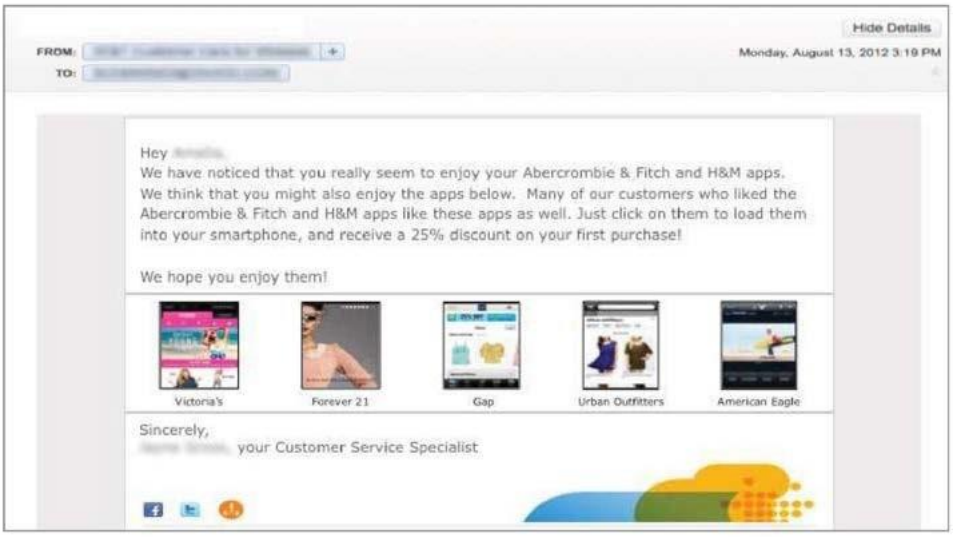


Figure 9-13: Leverage apps usage data to improve a subscriber's user experience

There are numerous "what if" questions that could fuel the brainstorming process for this mockup, such as:

- What if we could leverage our subscribers' app usage patterns to recommend apps that move the user into a more profitable, high-retention usage category

(for example, from "Moderate Female Teenage Browser" to "Female Teenage Shopaholic")?

- What if we could score the customers' app usage patterns to identify and act on potential churn situations more quickly?
- What if we could integrate app performance data across our subscriber base to recommend apps that provide a superior customer experience and help create more loyalty to the cellular provider?
- What if we could aggregate subscriber app usage insights across the entire network to create new monetization opportunities, such as app developer referral fees and co-marketing fees?
- What if we could integrate customers' app usage insights with real-time GPS location information to offer personalized location-based services?

Creating mockups is an effective technique for fueling the creative thinking process during the ideation workshop. Don't be concerned about the professional level of mockup (my mockups look like I drew them with a crayon). It's more important that the mockups challenge the current conventional thinking of the business stakeholders. The mockups can push the business stakeholders out of their current thinking ruts to contemplate the realm of what might be possible by leveraging all their customer and product insights to optimize existing customer engagement processes and uncover new monetization opportunities.

Summary

This chapter reviewed in detail the vision workshop or envisioning process. I described each of the five steps in the vision workshop methodology and provided details on each step using real-world examples.

You spent quite a bit of time on the data preparation and analysis work required to transform business initiative-specific data into an envisioning exercise that can be used as part of the ideation workshop. This is an important part of the vision workshop methodology because it helps the envisioning process *come to life* for the workshop participants. I provided several examples of creating customer-specific envisioning exercises.

You learned about the brainstorming and aggregation process of the ideation workshop. You also reviewed how to use the Michael Porter value creation processes—Value Chain and Five Forces Analysis—as well as the business initiative-specific envisioning exercise to tease out new business opportunities as part of the envisioning process.

Unit 6

10 Solution Engineering

You are now ready to tie all the previous exercises an
a solution. But what is meant by a "solution," and wh
cesses are necessary to architect a solution?

The trouble with big data is that there is no one shiny
can't just install Hadoop, predictive analytics, or a data a
will provide a big data solution. The data industry has stru
before, as data warehousing and business intelligence techn
within organizations over the past 10 to 15 years. To be suc
advanced analytics—like its brethren data warehousing a
before it—requires a new engineering skill, something cal

There's engineering for many disciplines—system
engineering, mechanical engineering—so why not solutio
engineering would be defined as:

> *A process for identifying and breaking down an organiz
> initiatives into its business enabling capabilities and su
> ogy components in order to support an organization's d
> data monetization efforts.*

Let's take a look at the steps of the solution engineering

NOTE Some key materials and graphics have been d
chapters in order to make this a stand-alone chapter.

The Solution Engineering Proces

Surprisingly, the solution engineering process is similar
bricks. The most successful LEGO projects are those that
thoroughly defined, well-scoped solution. Do I want to b
castle, or the space station? With LEGO bricks, I can bui

many more. However, each solution requires a different set of bricks in different configurations and a different set of instructions. Much like LEGO bricks, it is critical to ing your big data business success to identify up-front what solution your organization is trying to build, and then to assemble and integrate the right data and technology capabilities with the right instructions or roadmap to deliver a successful solution.

In this section I outline a 6-step solution engineering process for identifying, architecting, and developing a business solution (see Figure 10-1). This six-step solution engineering process encompasses the following:

1. Understand how the organization makes money
2. Identify your organization's key business initiatives
3. Brainstorm big data business impact
4. Break down the business initiative into use cases
5. Prove out the use case
6. Design and implement the big data solution

This process requires an up-front investment of time and creative thinking to grasp how your organization makes money. This means that you need to invest the time to identify your organization's *strategic nouns*; that is, those strategic business entities—like customers, stores, employees, and products—around which your organization builds differentiated business processes (e.g., acquisition, retention, optimization, management). You need to understand the role that these strategic nouns play in your organization's value creation processes. You need to identify the organization's key business initiatives and understand the desired impact of these initiatives on the organization's strategic nouns. This knowledge and information will guide and focus your solution engineering efforts. These are all activities for which approaches and methodologies were provided in Chapter 7.

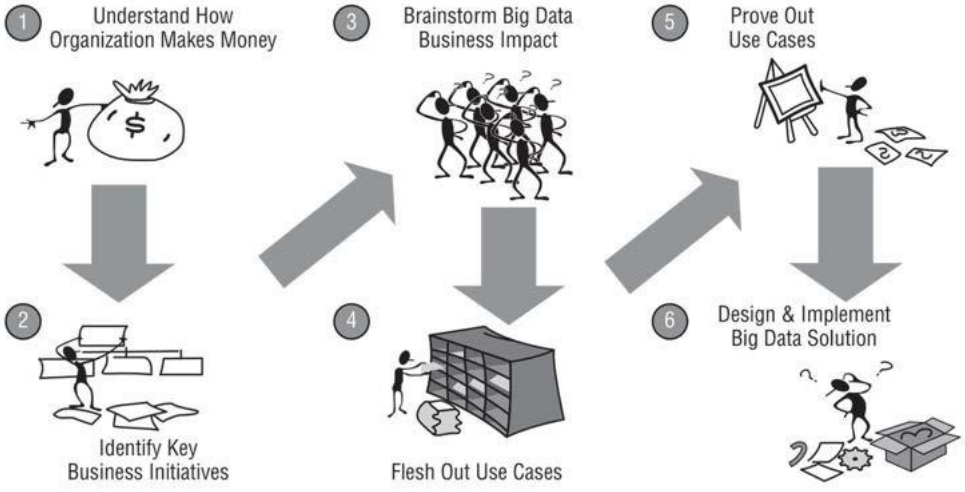


Figure 10-1 Solution engineering process

Step 1: Understand How the Organization Makes Money

Pretend that you are the general manager of the organization. Take the time to contemplate how the organization can make more money. For example, what can the organization do to increase revenues, decrease costs, reduce risks, or increase compliance?

There are many levers an organization can pull in order to make more money. Increasing revenue, for example, can include initiatives such as increasing the number of premium or gold card customers, increasing store or site traffic, reducing customer churn, increasing revenue per shopping occurrence, increasing private label sales as a percentage of market basket, increasing cross-sell/upsell effectiveness, and optimizing promotional effectiveness (see Figure 10-2). Reducing costs can include reducing inventory and supply chains costs, reducing fraud and shrinkage, improving marketing spend effectiveness, consolidating suppliers, improving on-time pickups and deliveries, optimizing merchandise markdowns, and improving asset utilization and turns.

Sales & Marketing

- Acquire more customers
- Retain existing customers
- Cross-sell/up-sell
- Increase market basket
- Increase store traffic
- Optimize pricing and yield
- Increase conversion rate
- Improve ad effectiveness

Operations

- Optimize network performance
- Predict maintenance problems
- Eliminate shrinkage
- Predict utilization/capacity
- Increase fill-rates
- Reduce out-of-stocks
- Consolidate suppliers

Finance

- Rationalize products
- Close unprofitable channels
- Increase inventory turns
- Increase asset utilization
- Reduce DSO
- Reduce SG&A
- Reduce T&E
- Reduce fraud and waste

Figure 10-2: High potential big data business opportunity

Next, spend the time to identify and understand your organization's strategic nouns, and ascertain how those nouns drive the moneymaking capabilities of the organization. For example, if you're in the airline industry, hubs are a very important noun used in your business, and any way that you can increase the number of flights per hub (such as decreasing airplane turnaround times or improving terminal and ramp efficiencies) means more flights per day, which equals more money. If you're

in the movie theater business, then concessions is a very important noun and any way you can increase the concession market baskets of theater guests (for example, buying a soda with the popcorn or buying the large water bottle versus the small water bottle) equals more money.

Finally, invest time actually using your organization's products or services. Experience first-hand how your organization's product or products work. Become a customer, become familiar with the user experience and understand the product's value propositions to its customers and partners. This will help you identify and understand the organization's key moneymaking and value creation activities that could be impacted by big data.

With these observations in hand, you are now prepared to put pen to paper and start to envision ideas on where new sources of customer, product, and operations insights could power your organization's ability to make more money. For example, if your organization is in the Business-to-Consumer (B2C) market, you can easily imagine how the organization could leverage both internal customer engagement data (such as e-mail, consumer comments, service logs, or physician notes) as well as external customer engagement data (such as social media postings, service ratings like Yelp, or blogs) to uncover insights that can help to optimize the customer engagement processes (for instance, profiling, segmentation, targeting, acquisition, activation, maturation, retention, and advocacy) in order to create more "profitable" customers (see Figure 10-3).

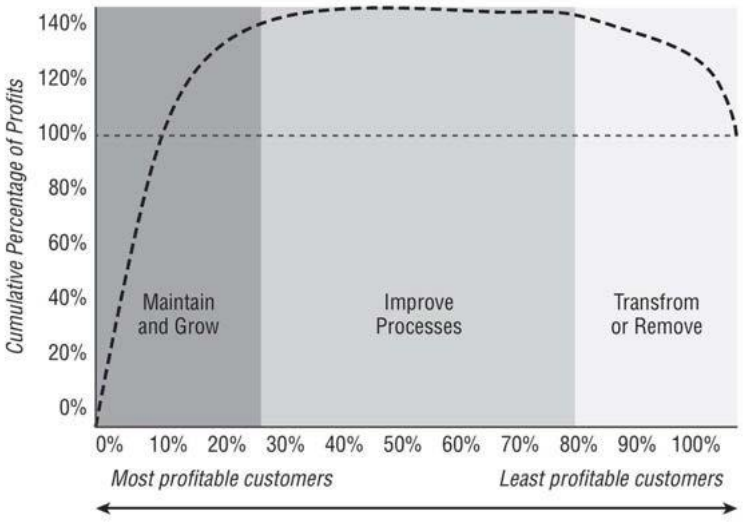


Figure 10-3: Customer profitability distribution

Source: R. S. Kaplan and S. Anderson, "Time-driven Activity-based Costing," *Harvard Business Review* (November 2004)

As we discussed in Chapter 8, across most industries:

- 0 to 25% of customers drive greater than 100% of profits
- 50 to 60% deliver no profits
- 10 to 25% deliver negative profits

Consequently, the solution engineering challenge in B2C industries is to determine how to leverage big data analytics to:

1. Move customers "up" the profitability curve (e.g., cross-selling them additional products, up-selling them more profitable products and services, replacing branded products with private label products to increase market basket profitability)
2. Service unprofitable customers in a more cost-effective manner (e.g., via the web, using self-service, through partners)

Step 2: Identify Your Organization's Key Business Initiatives

The next step is to do some primary research to understand your organization's key business initiatives. This includes reading the annual report, listening to analyst calls, and searching for recent executive management speeches and presentations. If possible, interview senior business management to understand their top business initiatives and opportunities, as well as their perceptions of the key challenges that might prevent the organization from successfully executing against their top business opportunities. (I will provide suggestions and examples on how to read an organization's annual report or public statements to uncover big data business opportunities later in this chapter.)

For each identified business initiative or opportunity, capture key information such as:

- Business stakeholders and their roles, responsibilities, and expectations
- Key performance indicators and the metrics against which success of the business initiative will be measured
- Timeframe for delivery
- Critical success factors
- Desired outcomes
- Key tasks

NOTE Review Chapter 6 on using the big data strategy document to break down your organization's business strategy into its key business initiatives, desired outcomes, and critical success factors.

156

Step 3: Brainstorm Big Data Business Impact

The next step in the solution engineering process is to brainstorm how big data and advanced analytics can impact the targeted business initiative. As discussed in Chapter 7, there are four ways (the four "Big Data Business Drivers"; see Table 10-1) that big data and advanced analytics can power an organization's key business initiatives:

1. Mine the more detailed transactional (dark) data at the lowest level of transaction granularity, which enables more granular and detailed decisions. For example, analyze the detailed transactional data, such as customer loyalty transactions, to enable more granular decision making and uncover new data monetization opportunities at the individual customer, seasonal/holiday, and location/geography levels.
2. Integrate new unstructured data sources to enable more robust and complete decisions. This includes internal unstructured data sources such as consumer comments, call center notes, e-mail, physician notes, and service bay logs, as well as external unstructured data sources such as social media posts, blogs, mobile/smartphone apps, and third-party or public data sources. I would also include sensor-generated data, such as smart grids, connected cars, and smart appliances, in this category. These new diverse data sources provide new variables, metrics, and dimensions that can be integrated into analytic models to yield actionable and material business insights and recommendations.
3. Provide real-time/low-latency data access where you reduce the time delay between when the data event occurs and the analysis of that data event, which enables more frequent and timely decisions and data monetization. This could include the creation of on-demand customer segments (based on the results of some major event like the Super Bowl) as well as real-time location-based insights from smartphones and mobile apps.
4. Integrate predictive analytics into your key business processes to provide new opportunities to uncover causality (cause and effect) buried in the data. Predictive analytics enable a different mindset with your business stakeholders, encouraging them to use new verbs—like optimize, predict, recommend, score, and forecast—as they explore new data monetization opportunities.

.

Table 10-1: Big Data Business Value Drivers

Big Data Business Drivers	Data Monetization Impacts
Transactional (dark) Data: More Detailed Transactional Data (e.g., POS, CDR, RFID, Credit Card)	Enable more granular, more detailed decisions (localization, seasonality, multi-dimensionality)
Unstructured Data: Diverse Internal (e-mail, consumer comments) and External (social media, mobile) Data	Enable more complete and more accurate decisions (new metrics, dimensions, and dimensional attributes)
Data Velocity: Low-latency ("Real-time") Data Access	Enable more frequent, more timely decisions (hourly versus weekly; on-demand analytic model updates)
Predictive Analytics: Causality, Predictors, Instrumentation, Experimentation	More actionable, predictive decisions (Optimize, Recommend, Predict, Score, Forecast)

Later in this chapter I will provide some examples across different industries of how to leverage these four big data business drivers to create business solutions.

Step 4: Break Down the Business Initiative Into Use Cases

The next step is to conduct a series of interviews and ideation/envisioning workshops to brainstorm, identify, define, aggregate, and prioritize the use cases necessary to support the targeted business initiative. As discussed in the Vision Workshop section of Chapter 9, you want to capture the following information for each identified use case:

- Targeted personas and stakeholders, including their roles, responsibilities, and expectations
- Business questions the business stakeholders are trying to answer, or could be trying to answer if they had access to more detailed and diverse data sources

e

- Business decisions the business stakeholders are trying to make, and the supporting decision processes including timing, decision flow/process, and downstream stakeholders
- Key performance indicators and key metrics against which business success will be measured
- Data requirements including sources, availability, access methods, update frequency, granularity, dimensionality, and hierarchies
- Identify analytic algorithms and modeling requirements such as prediction, forecasting, optimization, and recommendations
- Capture user experience requirements, which should couple closely with the user's decision-making process

This is great time to deploy the prioritization matrix methodology to use group dynamics to prioritize the different business use cases, and build IT and business consensus and support to move forward on the top use cases.

NOTE Step 3 and step 4 are great opportunities to use the previously discussed ideation workshop process described in Chapter 9 to brainstorm new ideas, aggregate the ideas into business relevant use cases, and prioritize the use cases based upon a weighting of business value and implementation feasibility.

Step 5: Prove Out the Use Case

Now is the time to deploy data and technology to validate the analytic feasibility of the solution. This is a good time to introduce a Proof of Value analytic lab to prove out the business case (financial model, ROI and analytic lift) using the full depth of available data and full breadth of available technology capabilities. You have a detailed definition of the desired solution including key decisions, business questions, key performance indicators, and all the other solution details captured in step 4. At this point, you should also have a solid understanding of the required data (such as data sources, key metrics, levels of granularity, frequency of access, dimensionality, and others) and the necessary technology and analytic capabilities to build out the Proof of Value. This Proof of Value analytic lab process should include:

- Gathering required data from both internal and external data sources, and integrate the data into a single data platform. You want the detailed data, not the aggregated data, because you're going to want to mine the detailed data

n i n n

to uncover the material, significant and actionable nuances buried in the data. You should also explore the use of third-party data, some of it publicly available from sources like `www.data.gov`, to help broaden the quality of the analytics. This is also a great time to bring in social media data, especially if you are dealing with a customer-centric use case.

- Defining and executing data transformation processes necessary to cleanse, align, and prepare the data for analysis. This most likely will include several data enrichment processes in order to create new composite metrics—such as frequency (how often an event occurs), recency (how recently the event occurred), and sequencing (in what sequence did the events occur)—that may be better predictors of business performance.
- Defining the analytic test plan including the test hypotheses, test cases, and measurement criteria.
- Developing and fine-tuning analytic models against defined key performance indicators and critical success factors. The data scientists involved in this step will likely continue to explore new data sources and new data transformation techniques that may help improve the reliability and predictability of the analytic models.
- Defining user experience requirements—in particular, understanding the downstream constituents of the analytic results and how the analytic results will need to be consumed by those constituents.
- Developing mockups and/or wireframes that help the business stakeholders understand how the resulting analytic results and models will be integrated into their daily business processes.

The goal of the Proof of Value analytic lab is to prove out the business case (including financial return or ROI, business user requirements, and critical success criteria), as well as to create and validate the underlying data models and analytic models that will provide the analytic lift. You want to validate that the integration of massive amounts of detailed structured and unstructured data coupled with advanced analytics can result in a more predictive, real-time analytic model that can deliver material meaningful and actionable insights and recommendations for the targeted business solution.

Step 6: Design and Implement the Big Data Solution

Based on the success of the Proof of Value analytic lab process, it's now time to start defining and building the detailed data models, analytic models, technology architectures, and production roadmap for integrating the analytic models and insights

n i n n

into the key operational and management systems. The implementation plan and roadmap will need to address the following:

- **Data sources and data access requirements**: This should include a detailed plan and roadmap for prioritizing what data to capture and where to store that data (both from a data access, as well as an analysis perspective). This plan will need to address both structured and unstructured data. It also needs to address external data sources, which means that the data plan will need to be updated every 4 to 6 months to accommodate the many new data sources that are becoming available.
- **Instrumentation strategy**: It is likely that additional data about your customers, products, and operations will need to be captured, mainly out of the existing business processes. The instrumentation strategy will need to cover how additional tags, cookies, and other instrumentation techniques can be used to capture additional transactional data.
- **Real-time data access and analysis requirements**: Certain use cases are going to require real-time (or low-latency) data access, analysis, and decision making as data is flowing through the business. These real-time requirements must be addressed across your entire technology and architectural stack including your Extract, Transform, and Load (ETL) and Extract, Load, and Transform (ELT) algorithms, data transformation and enrichment processes, in-memory computing, complex event processing, data platform, analytic models, and user experience.
- **Data management capabilities:** The big data industry has gained lots of experience and has developed many excellent tools and methodologies for helping organizations in the data management space (such as, master data management, data quality, and data governance). However, organizations also need to address when the data quality is good enough given the types of decisions and business processes that are being supported. Organizations need to carefully think through this question so that time is not wasted trying to make imperfect data perfect, especially when the decisions and the business processes that the data will support do not need perfection (for example, ad serving, fraud detection, location-based marketing, and markdown management). This part of the solution requires understanding and answering the "When is 90-percent accurate data good enough?" question.
- **Data modeling capabilities:** Data modeling requirements need to encompass all the traditional data warehousing architectural approaches—operational data store, data staging area, data marts, enterprise data warehouse—plus many of the new data platform and data federation tools and techniques

n i n n

that are available. The data modeling plan will need to consider data schema design and the role of NoSQL databases (where NoSQL stands for "Not Only SQL"), Hadoop and the Hadoop Distributed File System (HDFS).

- **Business intelligence**: Most organizations have an existing business intelligence or Business Performance Monitoring (BPM) environment in place that addresses key performance indicators, reporting, alerts, and dashboard requirements. This is the time to determine how to enhance that investment with new big data capabilities such as unstructured data, real-time data feeds, and predictive analytics. As discussed in the Business Model Maturity index section of Chapter 1, organizations have already invested a considerable amount of time, money, and people resources to build a BI environment around many of their critical internal business processes. Now is the time to develop a plan for how best to leverage and expand on those BI investments.
- **Advanced analytic capabilities (statistical analysis, predictive modeling, and data mining)**: This is the realm of the data science organization and much has already been discussed about the importance of creating an environment where the data science team is free to do their jobs. (I'll also discuss some high-level architectural components of an analytic sandbox in the next chapter.) Organizations should also start to develop an experimentation strategy that calls out the areas of the business where experimentation is going to be used to gain additional insights about customers, products, and operations.
- **User experience requirements**: The user experience plan needs to include the wireframe and mockup processes to ensure an understanding of how the analytic results and models will manifest themselves into the business users' daily operations and the management reports and dashboards. Use this opportunity to understand the user experience requirements of your internal users, external customers, and business partners, and to capture how analytic insights will be integrated into those user environments.

Solution Engineering Tomorrow's Business Solutions

While solution engineering might not be tomorrow's sexy job, it will become more and more important as the amount and variety of data continue to evolve, technology capabilities continue to expand (fueled by both venture capitalists and the explosive growth of the open source movement), and as mobile devices and smaller

form-factor mobile apps redefine the user experience. As the data and technology sands are shifting under your feet, it will become even more important that you are focused on delivering business solutions that have a high return on investment and a short payback period.

So how do you leverage these big data business drivers to explore or envision how big data and advanced analytics can help you define and deliver solutions that drive your key business initiatives? Let's walk through some examples.

Customer Behavioral Analytics Example

The big data opportunity in customer behavioral analytics is to combine your detailed customer transactions (such as sales, returns, consumer comments, and web clicks) with new social media and mobile data. The goal of combining these is to uncover new customer insights that can optimize your customer engagement lifecycle processes, such as profiling, targeting, segmenting, acquisition, activation, cross-sell/up-sell, retention, and advocacy. These same customer and product insights can ultimately lead to personalized marketing, especially when coupled with the real-time customer activity and location data that can be obtained via mobile apps. To gain new insights about your customers' behaviors, your organization could implement the following solutions:

- Integrate all of your detailed customer engagement transactions, such as sales history, returns, payment history, customer loyalty, call center notes, consumer comments, e-mail conversations, and web clicks into a single or virtual data repository.
- Use advanced analytics to analyze the detailed customer engagement transactions to model and score your most valuable customers and customer segments, create more granular behavioral categories, and use these behavioral categories and customer scores to refine your target customer profiles and customer segmentation strategies.
- Integrate and cleanse all of your prospect data—such as name, company, and contact information—gathered via lead generation events and third-party market sources.
- Augment the customer and prospect data with third-party data, from vendors such as Acxiom, Experian, BlueKai, and nPario, with customer demographic information, such as age, sex, education level, income level, and household information.
- Capture and aggregate relevant social media data about your products, services, and company from sites such as Facebook, Twitter, LinkedIn, Yelp, Pinterest, and others.

- Search, monitor, and capture relevant product and company comments from product and company advocates and dissenters located on blog sites such as WordPress, Blogger, and Tumblr.
- Use text analytics and/or Hadoop/MapReduce to mine the social media and blog data to uncover new insights about your customers' interests, passions, affiliations, and associations that can be used to refine your target customer profiles and customer segmentation models.
- Leverage mobile app capabilities to uncover real-time insights about your customers' locations, purchase behaviors, and propensities in order to drive real-time location-based promotions, offers, and communications.

Predictive Maintenance Example

Maybe the most significant opportunity for business-to-business (B2B) companies in the big data space is the opportunity to provide predictive maintenance services to their business (and possibly consumer) markets. Big data analytics can leverage sensor-generated data from appliances, equipment, implements, and machinery to analyze, score, and predict the maintenance requirements in real-time. Any industry that operates machinery—such as automobiles, airplanes, trains, farming machinery, construction equipment, appliances, energy equipment, turbines, servers, business equipment—can benefit from predictive maintenance that is enabled by the combination of sensor-generated data coupled with real-time analytics. To gain new predictive maintenance insights, organizations could architect the following solution:

- Capture raw unstructured appliance, equipment, implement, and machinery sensor-generated logs and error codes in real-time *as-is* (with no data preprocessing required, and no predefined data schemas) into Hadoop and HDFS.
- Use advanced analytics against your historical performance data to build predictive models of what constitutes "normal" appliance, equipment, and machinery performance at the individual unit and component levels. Six Sigma techniques, such as control charts, can be very useful to identify unusual product performance. Plus, Six Sigma is a methodology that is typically well understood within manufacturing industries.
- Leverage advanced data enrichment techniques, such as frequency, recency, and sequencing metrics to identify combinations of events or event thresholds that may be indicative of maintenance needs. Think about creating a "basket"

4 u t a

of activities that can be mined using market basket or association analytic modeling algorithms.

- Integrate external dynamic data sources such as weather (temperature, rainfall, snow, ice, humidity, and wind), traffic, and economic data sources, to identify new variables that can enhance the predictive models. For example, ascertain what impact humidity might have on the performance of your wind turbines or the impact rain and snow have on the on-time performance of your trains.
- Leverage a real-time analytics environment to monitor real-time streaming sensor data, to compare the feeds in real-time to your performance models and control charts, and then flag, score, and rank any potential performance problems.
- Send out automated alerts to concerned parties (such as technicians or consumers) including recommended maintenance information (like location, projected replacement parts, projected maintenance crew skills, and maintenance best practices documentation), and create optimized service schedules, calendars, and crew scheduling.
- Capture product or component wear data from replaced parts at the time of maintenance to continuously refine predictive maintenance models at the individual appliance/machinery and component levels.
- Aggregate and analyze wear data in order to create, package, and sell performance insights back to appliance, machinery, product, and component manufacturers.

Marketing Effectiveness Example

Every company spends money on marketing and increasingly portions of that spend are being spent in highly measurable digital media channels. Quantifying the effectiveness of marketing spend across the online—as well as offline channels such as TV, print, and radio—is a difficult challenge. Organizations that can more accurately quantify and attribute credit to the marketing channels and marketing treatments that are driving business and sales performance are better positioned to optimize marketing spend. To better measure marketing effectiveness, organizations could craft the following solution:

- Aggregate all marketing spend, at the lowest level of detail, across all digital channels (impressions, display, search, social, and mobile) as well as offline channels (TV, print, and radio).
- Integrate all sales activities and transactions (calls, bids, proposals, and sales losses and wins) with online conversion events, and associate these activities back to the different marketing activities and spend.
- For digital data, capture and aggregate into a market basket the impressions, displays, key word searches, social media post, web clicks, mouse overs, and associated conversion events at the individual user level (cookie-level detail).
- Calculate advanced composite metrics associated with marketing treatment frequency, recency, and sequencing in order to quantify the effectiveness of the different marketing treatments (attribution analysis).
- Augment campaign data with external data such as weather, seasonality, local economic, local events, and other similar data to improve campaign modeling and predictive effectiveness.
- Benchmark current campaign performance against previous and similar ("like") campaigns to identify and quantify previous campaign performance drivers.
- Leverage prospect data captured via third-party direct marketing campaigns to build out your prospective database against which you will run your direct marketing campaigns.
- Acquire new sources of customer digital insights from DMP vendors like BlueKai and nPario.
- Develop an experimentation strategy to test the effectiveness of different marketing treatments, messaging, and channels.
- Analyze social media data to capture consumer interests, passions, affiliations, and associations that can improve profiling, segmentation, and targeting effectiveness.
- Capture real-time social media feeds to analyze, monitor, and act on campaign and product sentiment trends in real-time.
- Use the insights from the marketing performance analytics and insights to drive both pre-campaign media mix allocation recommendations (such as, how to allocate marketing spend between different marketing channels like TV, print, online, display, keywords, and others) and in-flight campaign performance recommendations (such as, how to reallocate digital media spend between ad networks, keywords, target audiences, and similar items).

Fraud Reduction Example

Big data provides new and innovative technologies to identify potentially fraudulent activities in real-time. New data sources (such as social media and detailed web and mobile activities) and new big data innovations (like real-time analytics) are enabling organizations to move beyond the traditional static fraud models to create dynamic, self-learning fraud models that flag behavioral and transactional activities as they are occurring, as well as combinations of activities, that are potentially fraudulent. Here is an example of a big data fraud detection solution:

- Deploy a real-time data platform that can capture and manage a high volume of real-time data feeds (such as purchases, authorizations, and returns) from multiple internal and external data sources.
- Use in-database analytics to accelerate the development and refinement of fraud prediction models based on historical transactions.
- Use predictive analytics to analyze real-time transactions to score unusual transactions, behaviors, and tendencies across thousands of dimensions and dimensional combinations, and compare those scores to historical norms to flag potential fraud situations.
- Employ advanced data enrichment techniques—such as frequency, recency, and sequencing of activities and transactions—to create more advanced profiles of potentially fraudulent activities, behaviors, and propensities.
- Integrate mobile data with location-based analytics to dynamically identify and monitor locations, businesses, and other places that have a higher than normal propensity for potentially fraudulent behaviors (for example, gas stations, discount retailers, and convenience stores).
- Integrate real-time fraud detection models into operational systems (such as point-of-sale systems, call centers, and consumer messaging systems) to enable real-time challenging of specific transactions and groupings of transactions, with the goal of challenging those transactions while they are still in process.
- Leverage social media data to identify networks or associations of potential fraud cohorts.

Network Optimization Example

Whether you operate a network of devices (servers, ATM, switching stations, or wind turbines) or outlets (stores, sites, or branches), there are invaluable sources of new customer and product data that can be leveraged to ensure you have the "right nodes in the right locations at the right time" to provide an "exhilarating" customer experience. Over and undercapacity are always key challenges to networks, and

those capacity requirements and needs can change rapidly based on customer and product behaviors and tendencies. To optimize your network operations, here is what a network optimization solution might include:

- Aggregate your network "node" data at the lowest level of detail (such as log files) across all of your different network components and elements. Keep lots of history at the detailed transaction level.
- Integrate social and mobile consumer data to identify and quantify changes in customer, network, and market preferences and behavioral tendencies.
- Augment data assets with external data sources such as weather, local events, holidays, and local economic data to provide new predictive metrics that can improve the predictive capabilities of capacity planning and resource scheduling models.
- Use advanced analytics to project network capacity requirements (at the node, time of day, and day of week levels, and so on) and calculate key network support variables such as personnel, inventory, replacement parts, and maintenance scheduling.
- Leverage real-time analytics to reallocate network capacity (resource scheduling, ramping up or ramping down cloud resources, and others) to support daily, hourly, and location usage pattern changes.

Being a solutions engineer requires not only a strong understanding of the business problems that your organization is trying to address, but also a strong understanding of the capabilities of new big data and advanced analytics innovations. Applying the six-step solution engineering process helps to ensure that you "deploy the right technology capabilities at the right time to solve the right business problem." Without a detailed understanding of the problem you are trying to solve and a solid foundation in the big data analytics capabilities, you will quickly fall back into the old methods of leading with technology "in search of a business problem."

Reading an Annual Report

I'm a big advocate of leveraging an organization's public documents (such as annual reports and quarterly 10-Q financial filings) and announcements (such as press releases and executive presentations) to uncover big data business opportunities. This section of the book will provide some real-world examples of where to look in an annual report to identify potential big data opportunities, and how to do a quick assessment of how big data might be used to power those opportunities.

I'm always surprised by how few people take the time to read their company's annual report, or search out public statements and presentations being made by senior members of the organization's leadership team. In particular, the "President's Letter to the Shareholders" (or the CEO's letter in some cases) is a gold mine. It is within this section of the annual report that the president talks about all the great things they did for the company over the past year. That usually takes up about 75 percent of the letter and, in my humble opinion, can largely be ignored. It's the last 25 percent of the letter that is most informative because it is here that the president talks about the key business initiatives for the next year. Let's review a few annual reports to see what I mean.

Financial Services Firm Example

Following is an excerpt from a letter to the shareholders of a financial services firm:

> **This year we have crossed a major cross-sell threshold. Over banking households in the western U.S. now have an average of 6.14 products with us. For our retail households in the east, it's 5.11 products and growing.** *Across all 39 of our Community Banking states and the District of Columbia, we now average 5.70 products per banking household (5.47 a year ago). One of every four of our banking households already has eight or more products with us. Four of every ten have six or more. Even when we get to eight, we're only halfway home. The average banking household has about 16.* **I'm often asked why we set a cross-sell goal of eight. The answer is, it rhymed with "great." Perhaps our new cheer should be: "Let's go again, for ten!"**

This section of the letter highlights a business initiative to improve customer cross-sell effectiveness in order to reach a goal of 10.0 products per banking household, an increase from the current 5.7 products per banking household. While 10.0 may be a BHAG (Big, Hairy, Audacious Goal!), it is clear that some executive in the organization (likely in Marketing) has been chartered with increasing cross-sell effectiveness.

Here are some examples of how big data could help their cross-sell effectiveness business initiative:

- Use detailed customer financial data on the number and types of accounts held by household, combined with key account information (such as length of account ownership, account balance, and account balance trends) and household demographics data, to create more granular household segments.

- Run analytic models to score these new household segments by their likelihood to buy a specific additional financial product. For example, households who hold these products and are in this demographic group have a certain percentage-likelihood of buying this additional product.
- Develop different models for different combinations of products and household demographics.
- Use social media data from sites such as Facebook, Pinterest, Yelp, and Twitter to identify trends in financial products that might be candidates for cross-product promotions. For example, mortgage refinancing is hot, so look to bundle mortgage refinancing with a home equity line of credit. Run these trends against your household/product cross-sell models to identify direct marketing targets.
- Instrument or tag all of the direct marketing and digital marketing campaigns to see what messaging and offers work best for which audience segments.
- Develop an experimentation strategy for identifying what offers to test with what audience segments. Capture the results in real-time and make in-flight campaign adjustments.

Retail Example

This example uses information gained from a retailer's 2011 annual report. There are at least two sections of the annual report where the company could integrate structured data (such as point-of-sale, inventory, returns, and orders transactions) with unstructured data (such as social media, web log, and consumer comments) to drive their key business initiatives. The following highlights the first section of the letter to shareholders:

> *Our strategy is to provide our members with a broad range of high quality merchandise at prices consistently lower than they can obtain elsewhere.* **We seek to limit specific items in each product line to fast-selling models, sizes, and colors.** *Therefore, we carry an average of approximately 3,600 active stock keeping units (SKUs) per warehouse in our core warehouse business, as opposed to 45,000 to 140,000 SKUs or more at discount retailers, supermarkets, and supercenters. Many consumable products are offered for sale in case, carton or multiple-pack quantities only.*

This section highlights an opportunity where big data could help drive store assortment optimization. In particular, big data could help in the following ways:

- Integrate demographics data and product sales data to forecast optimal store assortment (at an individual store level), and update optimal store assortments

e

more frequently (perhaps weekly) based on local events such as Cinco de Mayo or San Francisco Giants home games.
- Integrate social media insights with consumer comments (such as those gained from call centers, e-mail, and websites) with store and product sales data to calculate and track the net promoter scores and consumer sentiment for particular stores (by product category or season). Then use this information to identify and act on underperforming stores, products, and product categories.
- Leverage social media data to identify product and market trends (by store and product category) that can impact pricing, in-store merchandising, and store assortment planning.
- Test different store assortment options in different stores, capture the results, and make recommendations to optimize store assortment at the individual store and department levels.

The second example from this retail company that follows, highlights the business value of increasing private label sales effectiveness (e.g., increase private label sales from 15 percent of products sold to 30 percent over the next several years).

> *We remain focused on selling national brand merchandise while developing our Private Label brand to enhance member loyalty. After 19 years, our Private Label products now represent 15% of the items we carry, but 20% of our sales dollars.* **We believe that we have the capability of building our sales penetration of our Private Label products to 30% over the next several years,** *while continuing to provide our members with quality brand name products that will always be a part of our product selection.*

Here are some examples of how big data could be used to increase private label sales effectiveness:

- Integrate sales and inventory data with social media data to score product categories that are the most likely opportunities (highest probabilities of success) to introduce private label products. Scores are created by store, geography, and product category.
- Mine social media data to identify consumers' areas of interest that can be used for direct marketing and in-store promotions and merchandising around promoting private label products.
- Integrate historic private label sales data and correlate them with local geographic variables including economic conditions, unemployment rate, changes in home sales and home values, traffic conditions, and similar items.

- Test different private label strategies at different geographic and store levels to determine if certain geographies are more receptive to certain private label product categories.

Brokerage Firm Example

This third example comes from a brokerage firm's 2010 annual report. Following is an excerpt from the letter to shareholders.

> *Client feedback is essential if we hope to see the world through [the] client's eyes. Last year, we continued our Client Promoter Score (CPS) program in which we survey clients and ask them to rate us, from 0 to 10, on their willingness to recommend us. The CPS calculates the number of "promoters" minus the number of "detractors" to arrive at a net indicator of client loyalty.* **Our CPS for individual investors reached a record 37%, with significant gains for our value proposition, investment help and guidance, and customer service.** *CPS scores also remained strong for independent investment advisors, who praised our responsive service, and for retirement plan sponsors.*

The annual report highlights the company's plan to focus on driving their Client Promoter Score (CPS) program. In 2010, the firm was able to achieve their highest score to-date, a record 37 percent. (It would be nice to know if they had a goal and a timeline for their CPS, but maybe that's something that can be determined via interviews.) Here are some examples of how big data can be used to drive the CPS program.

- Leverage social media sites and blogs to create a more current and comprehensive CPS that is a better predictor of clients' feelings and perspectives (for example, their likeliness to recommend the firm to others).
- Integrate unstructured customer conversations from call centers, consumer comments, and e-mail received.
- Build analytic models that incorporate social data and different financial transactions to break out and track CPS by most "valuable" customer segments. Match customer financial transaction patterns with sentiment analysis to flag potential CPS score drops by customer segment.
- Create analytic models that analyze CPS variables such as broker, brokerage firm, broker location, financial topic, day of week, time of day, and so on. Triage analytic results to uncover any correlations between CPS and broker engagement variables.

- Use the CPS to segment key customers. Leverage Twitter and Facebook data to monitor sentiment trends across their most valuable customer segments in order to more quickly identify and quantify (score) potential customer attrition and corresponding performance drivers.
- Capture key broker demographic (background, education, certifications, years of experience) and performance (client performance, satisfaction scores) data to model the correlation between broker demographics and customer CPS.
- Create real-time tracking control charts that are constantly monitoring key broker engagement variables for potentially troubling situations. Create control charts at the different broker engagement levels, such as broker, broker location, brokerage firm, and financial topic.

You've considered the business transformational power of big data—the power to tease out new customer, product, and market insights that can be used to drive higher-fidelity, more frequent business decisions. But in order to drive business transformation, you need to "begin with an end in mind" (to steal from Stephen Covey). You need to invest the time to understand your organization's key business initiatives, and contemplate the "realm of the possible" with respect to the big data business drivers (for example, more detailed structured data, new unstructured data sources, real-time/low-latency data access, and predictive analytics). There is no better place to start your big data journey than by targeting the key business initiatives that can be found in your company's annual report.

Summary

This chapter introduced you to the concept of solution engineering and provided a six-step process for going from opportunity identification to solution implementation. I provided several examples across different industries, highlighting how a business solution could leverage new sources of data and new big data technology innovations.

You then learned how to read an annual report (and other publicly available data sources) to identify an organization's business initiatives where big data can provide material financial impact. You then reviewed several examples of reviewing annual reports across different industries to identify how big data could impact those organizations' key business initiatives.

e

Unit 7

Big data Architecture

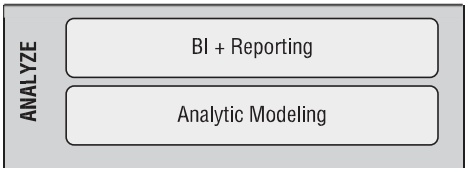
This is the part of the book that most everyone has been waiting for—the technology discussion. As you can probably guess, there is a reason the technology discussion is near the end of the book: a technology discussion can only be productive if it has the benefit of a prior understanding of the business drivers and the targeted business solution around which to scope the technology discussion. Too many times in this industry the first discussion is about the technology features. The reason why organizations want to talk about the technology is because it's easy to talk about its general features and capabilities (the "feeds and speeds") as compared to taking the time to understand what business challenges or opportunities an organization is trying to address with the technology. As is typical in this industry, we're always in search of the "silver bullet."

So this chapter briefly introduces some of the new big data technologies, and provides links and suggested readings in case you want to dive deep into the technologies—there are plenty of outstanding and freely available resources that talk about the new big data technologies. The remainder of the chapter then focuses on exploring the architectural ramifications of big data, especially for organizations that have already made significant investments in their data warehousing and business intelligence (BI) capabilities. As discussed in Stage 1 of the Big Data Business Maturity Index, those business processes around which organizations have already built their data warehouses and BI capabilities are good starting points for the big data journey.

Big Data: Time for a New Data Architecture

For the past 15 to 20 years, organizations have been operating with the data architecture that was built on Online Transaction Process (OLTP)-centric relational database technologies. This architecture worked just fine when dealing with gigabytes and low

terabytes of structured data in batch mode, and business users became accustomed to data request turnaround times measured in weeks or months. But there wasn't much "intelligence" in the BI tools with, and very little predictive analytics and data mining capabilities. Reports and dashboards that monitored performance with a rearview mirror view of the business were state of the art (see Figure 11-1).



t n

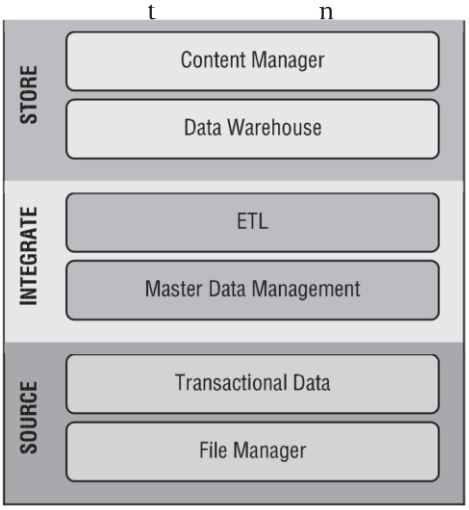


Figure 11-1 Traditional business intelligence/data warehouse reference architecture

- Batch oriented, with high latency
- Monolithic, with inflexible layers
- Brittle and labor intensive (metadata jailhouses)
- Focused on structured data
- Performance and scalability challenged
- Information integration requires significant hand-coding
- Data stored in aggregated tables for specific reports

However, we all know the story by now—Internet companies like Google, Yahoo!, and Facebook couldn't make this architecture work. They explored traditional data management and analysis tools from the traditional data and BI vendors, but even

t n

t n

Introducing Big Data Technologies

These big data technologies have the potential to dramatically re-invigorate your existing data warehouse and BI investments with new capabilities and new architectural approaches. Organizations have an opportunity to extend their

11 Big Data Architectural Ramifications

t n

This is the part of the book that most everyone has been waiting for—the technology discussion. As you can probably guess, there is a reason the technology discussion is near the end of the book: a technology discussion can only be productive if it has the benefit of a prior understanding of the business drivers and the targeted

t n

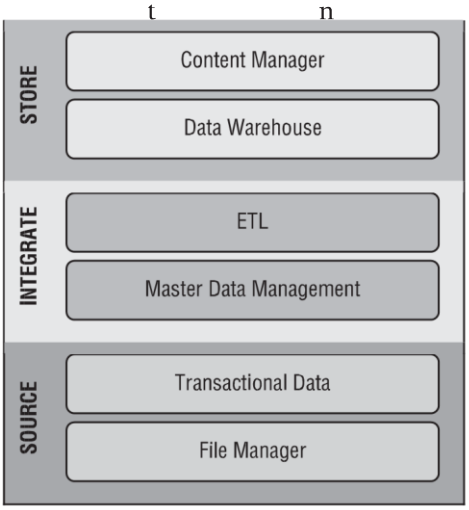


Figure 11-1 Traditional business intelligence/data warehouse reference architecture

- Batch oriented, with high latency
- Monolithic, with inflexible layers
- Brittle and labor intensive (metadata jailhouses)
- Focused on structured data
- Performance and scalability challenged
- Information integration requires significant hand-coding
- Data stored in aggregated tables for specific reports

However, we all know the story by now—Internet companies like Google, Yahoo!, and Facebook couldn't make this architecture work. They explored traditional data management and analysis tools from the traditional data and BI vendors, but even

t n

doing things like trying to dynamically tinker with their software kernels to accommodate real-time analysis across hundreds of terabytes and petabytes of data didn't

Introducing Big Data Technologies

These big data technologies have the potential to dramatically re-invigorate your existing data warehouse and BI investments with new capabilities and new architectural approaches. Organizations have an opportunity to extend their

t n

existing data warehouse and BI environments by leveraging the following big data capabilities:

t n

by the framework. It enables applications to work with thousands of computation-independent computers and petabytes of data. The entire Apache Hadoop "platform" is now commonly considered to consist of the Hadoop kernel, MapReduce, HDFS, and a number of related projects including Apache Hive, Apache HBase, and others.

NOTE The Apache Software Foundation is a community of developers and users organized for the purpose of coordinating a portfolio of open source projects, and promoting the development and use of open source products, many of which will be described in this chapter. You can learn more about the Apache Software Foundation at `http://www.apache.org/`.

Hadoop MapReduce

MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. A MapReduce program comprises a `Map()` procedure that performs filtering and sorting (such as, sorting students by first name into queues, one queue for each name) and a `Reduce()` procedure that performs a summary operation (such as, counting the number of students in each queue, yielding name frequencies). The *MapReduce System* (also called *infrastructure* or *framework*) orchestrates the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, providing for redundancies and failures, and managing the overall process. Figure 11-3 shows how the MapReduce function works.

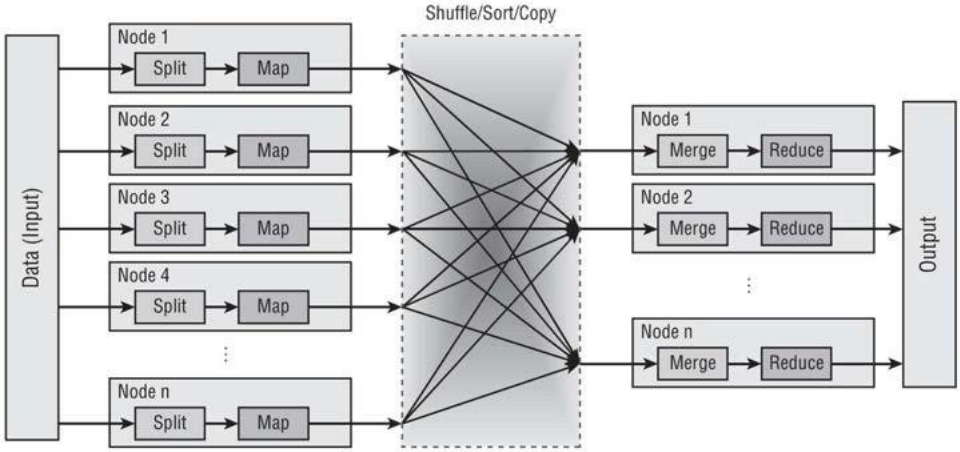


Figure 11-3 MapReduce process flow

Apache Hive

Apache Hive is a data warehouse infrastructure built on top of Hadoop that provides data summarization, query, and analysis. While initially developed by Facebook, Apache Hive is now used and is being enhanced by other companies, such as Netflix. As of the writing of this book, Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic MapReduce on Amazon Web Services. Apache Hive supports analysis of large data sets stored in Hadoop-compatible file systems. It provides an SQL-like language called HiveQL while maintaining full support for MapReduce. To accelerate queries, Hive provides indexes, including bitmap indexes.

Apache HBase

HBase is an open source, non-relational, distributed database model written in Java. It was developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS. HBase provides a fault-tolerant way of storing large quantities of sparse data. HBase features compression, in-memory operation, and Bloom filters on a per-column basis. Tables in HBase can serve as the input and output for MapReduce jobs run in Hadoop, and may be accessed through the Java API.

Pig

Pig is a high-level, natively parallel data-flow language and execution framework for creating MapReduce programs. Pig abstracts the MapReduce programming language into a higher-level construct, similar to how SQL is a higher-level construct for relational database management systems. Pig can be extended using user-defined functions, which the developer can write in Java, Python, JavaScript, or Ruby and then call directly from the language.

Figure 11-4 shows a typical Hadoop architecture or ecosystem configuration including many of the components discussed above.

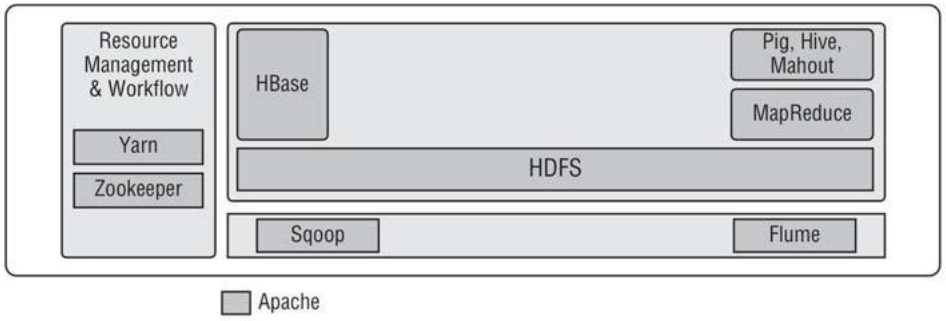


Figure 11-4 Standard Hadoop architecture

NOTE There is a bevy of content on these technologies available on the Apache Hadoop website (`http://hadoop.apache.org/`). I also recommend reading the book *Hadoop: The Definitive Guide* by Tom White (O'Reilly and Yahoo! Press, 2009). It is the definitive book for those seeking to learn more about Hadoop and the Hadoop ecosystem.

It is worth noting that many vendors are investing heavily to extend the Hadoop functionality to make it easier to leverage Hadoop within an organization's existing data warehouse and BI environments. As of the writing of this book, some vendors such as Pivotal, are adding the capability to access data stored on HDFS directly with industry standard SQL query tools and SQL-trained personnel (see Figure 11-5). I expect this trend to continue.

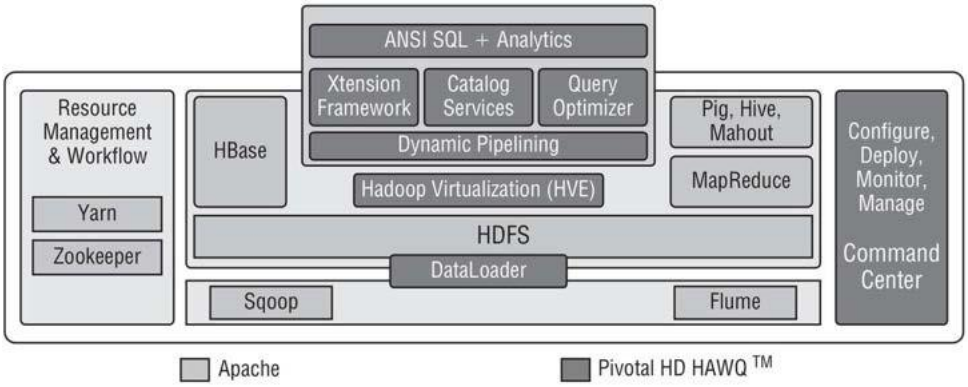


Figure 11-5 Extended Hadoop architecture

New Analytic Tools

There has also been a rash of new development in the area of analytics and data visualization tools fueled by big data. Some of the more interesting tools include:

- R, an open source programming and statistical language that is rapidly gaining popularity among universities and startup companies alike. R is a GNU project, so the software is freely redistributable. Plus it has the advantage of literally thousands of developers using and extending the R capabilities. Go to `http://cran.r-project.org/` to learn more about R. And I personally love the RStudio integrated development environment, which opens R to a larger constituency of users. Go to `www.rstudio.com/ide/download/` to download and learn more about RStudio.

- Apache Mahout, another Apache Software Foundation project, provides scalable machine learning algorithms on top of the Hadoop platform. Mahout provides algorithms for clustering, classification, and collaborative filtering implemented on top of Apache Hadoop using MapReduce. Go to http://mahout.apache.org/ to learn more about Apache Mahout and see the wide range of analytic algorithms supported by Mahout.
- MADlib, an open-source library that supports in-database analytics. It provides data-parallel implementations of mathematical, statistical, and machine-learning methods that support structured, semistructured, and unstructured data. Go to `http://madlib.net/` to learn more about MADlib.

New Analytic Algorithms

Finally, I don't want to leave out the many innovations that are taking place in the development of new, advanced analytics capabilities. The discussion about the capabilities of these new algorithms is beyond the scope of this book. However, I have provided a partial list below of some of my data scientist friends' favorite new algorithms, along with a link where you can learn more:

- Support Vector Machines are based on the concept of decision planes that define decision boundaries and a decision plane that separates sets of objects having different class-membership (`www.statsoft.com/textbook/support-vector-machines/`).
- Random Forest consists of a collection of simple tree predictors, each capable of producing a response when presented with a set of predictor values (`www.statsoft.com/textbook/random-forest/`).
- Ensemble Methods is a model testing and verification technique that tests multiple models to obtain better predictive performance than could be obtained from any one analytic model (`http://en.wikipedia.org/wiki/Ensemble_learning`).
- Champion/Challenger is another model testing and verification technique where you classify your current analytic model as the "Champion," then challenge the champion with different analytic models where each "Challenger" differs from the Champion in some measurable and defined way (`www.edmblog.com/weblog/2007/04/adaptive_contro_1.html`).
- Confusion Matrix is a specific table layout that allows visualization of the performance of an algorithm (`http://en.wikipedia.org/wiki/Confusion_matrix`).

Sample Questions on Introduction to Big Data

1.Why is big data analytics important?

2. Differentiate between adata analysis and data analytics

3.What are the five V's of Big Data?

4.List some tools Used for Big Data?

5.Which are the best tools that can be used by Data-Analyst?

6.What is Data Cleansing?

7.What are the sources of Unstructured data in Big Data?

8.Define term Outlier in Big Data analytics?

9.What is the K-mean Algorithm?

10.What is Clustering?

11.Mention some statistical methods needed by a data analyst?

12.What is the difference between Data Mining and Data Analysis?

13.What is the Function of a collaborative filtering algorithm?

14.What is the difference between linear regression and logistic regression?

15.What are the most common analytical technique categories?

16.What does P-value signify about the statistical data?

17.What is Machine Learning?

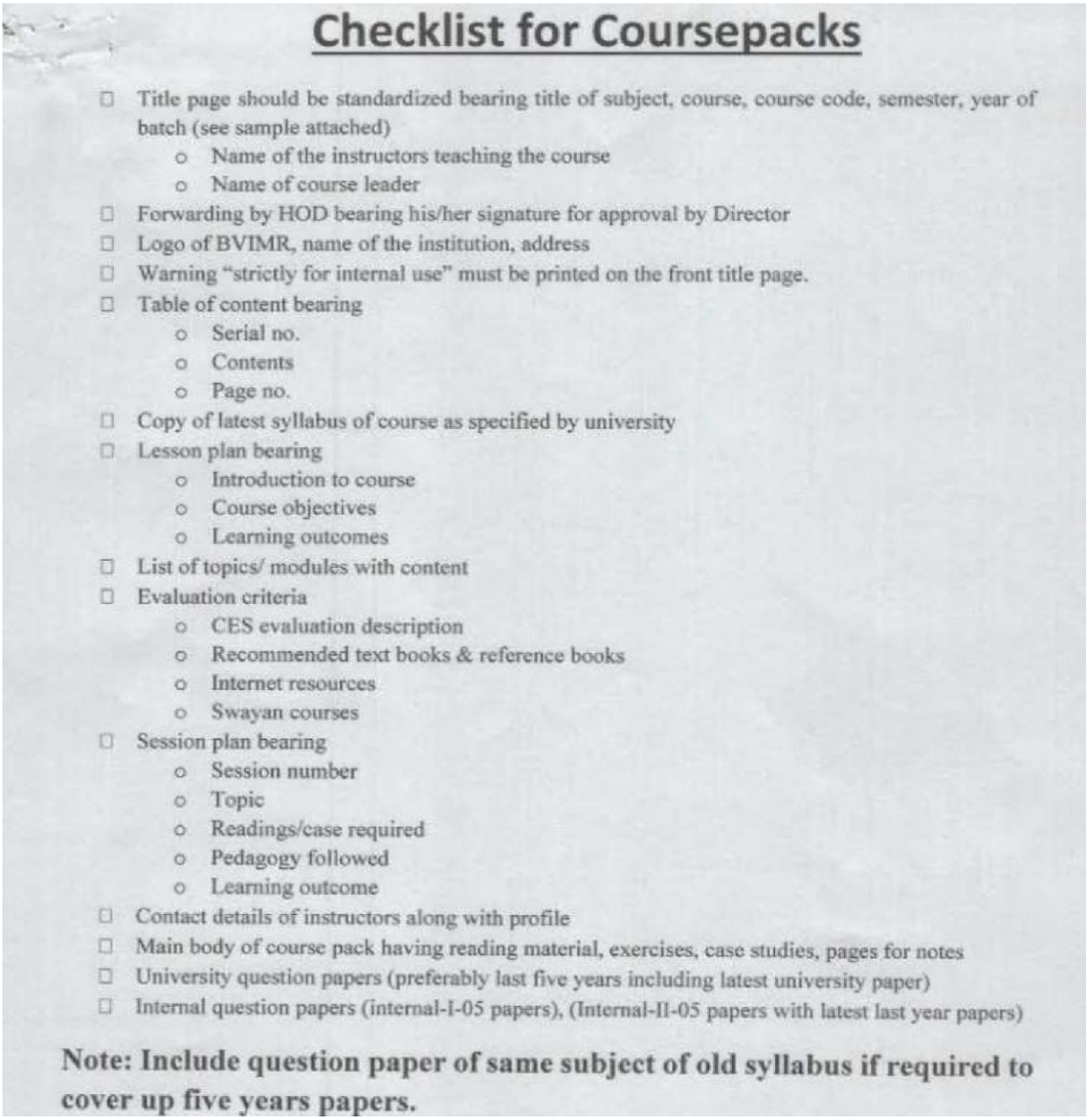
18. What is the difference between data mining and data profiling?

19.What are an Eigenvalue and Eigenvector?

20 What are the data validation methods used in data analytics?

21 Explain some programming languages used in Big Data Analytics?

22.What are the primary responsibilities of a data analyst?



Checklist for Coursepacks

- ☐ Title page should be standardized bearing title of subject, course, course code, semester, year of batch (see sample attached)
  - Name of the instructors teaching the course
  - Name of course leader
- ☐ Forwarding by HOD bearing his/her signature for approval by Director
- ☐ Logo of BVIMR, name of the institution, address
- ☐ Warning "strictly for internal use" must be printed on the front title page.
- ☐ Table of content bearing
  - Serial no.
  - Contents
  - Page no.
- ☐ Copy of latest syllabus of course as specified by university
- ☐ Lesson plan bearing
  - Introduction to course
  - Course objectives
  - Learning outcomes
- ☐ List of topics/ modules with content
- ☐ Evaluation criteria
  - CES evaluation description
  - Recommended text books & reference books
  - Internet resources
  - Swayan courses
- ☐ Session plan bearing
  - Session number
  - Topic
  - Readings/case required
  - Pedagogy followed
  - Learning outcome
- ☐ Contact details of instructors along with profile
- ☐ Main body of course pack having reading material, exercises, case studies, pages for notes
- ☐ University question papers (preferably last five years including latest university paper)
- ☐ Internal question papers (internal-I-05 papers), (Internal-II-05 papers with latest last year papers)

Note: Include question paper of same subject of old syllabus if required to cover up five years papers.

Declaration by Faculty

I Dr Jitendra Singh Designation Visiting Faculty Teaching Introduction to Big Data subject in BCA Gen, course vth sem have incorporated all the necessary pages section/quotations papers mentioned in this check list above.

(Santanoo Pattnaik)