BIG DATA ANALYTICS

Creating Winning Customer Experiences with Big Data Analytics

CHAD RICHESON

Big Data Analytics

Creating Winning Customer Experiences with Big Data Analytics

Cover Design by Design Commission

About the Author

Chad Richeson is CEO of Society Consulting, one of the country's largest analytics consultancies. Society provides Analytics, Big Data, and Customer Experience services to a variety of blue chip companies. Chad has built, managed, and consulted on dozens of Big Data projects for companies across North America. Prior to co-founding Society in 2011, Chad spent nearly 12 years at Microsoft building out the company's Big Data platform, and co-leading the integration of Bing and Yahoo Search data systems. Chad holds an MBA from Tulane University.

Chad can be reached at chad@societyconsulting.com, or followed on Twitter at @chadricheson.

Introduction

Big Data is driving a global revolution. Across Public and Private Sectors, and industries as diverse as Retail, Healthcare, Media, and Transportation, Big Data is impacting the lives of billions of people.

The majority of Big Data's impacts are easy to overlook. Online services such as Google, Bing, and Facebook are Big Data engines used by nearly a billion people every day. The Agriculture industry leverages Big Data to drive higher crop yields. And Government agencies employ Big Data to prevent and solve crime and cyber terrorism in countries around the world. The impacts of Big Data are substantial, yet are so unobtrusive as to go unnoticed by most people. To paraphrase Gil Scott-Heron, the Big Data Revolution will not be televised.

Big Data is also dramatically changing the relationship between brands and their customers. Across a variety of industries, brands are digitizing their Sales, Service, and Support channels to deliver more value to customers. Big Data allows these brands to understand customers as individuals, predict what they will need next, and take quick action on those needs. Large companies such as Google, Apple, Microsoft, Facebook, and Amazon already use Big Data to generate competitive advantage and higher profits; their methods are now being adopted by others.

What is fueling the Big Data revolution? The pervasiveness of smartphones, the declining cost of computing, and advances in communications all play a big role. But as well is the increased demand by customers that brands deliver ever more personalized value. These changes are happening so quickly that the only thing holding Big Data back is the knowhow required to make it all work.

That's where this book comes in. *Big Data Analytics* gives business leaders the tools to design, build, and operate a Big Data Analytics program at any company regardless of size or industry. As a companion to *Big Data Platforms* (also from Society Consulting), *Big Data Analytics* helps leaders understand how to leverage Big Data to create winning customer experiences.

Contents:

Chapter 1: The New Rules of Competition

Chapter 2: The Components of a Big Data Analytics Program

Chapter 3: Building a Big Data Analytics Team

Chapter 4: Running Big Data Analytics Processes

Chapter 5: Setting Big Data Priorities

Chapter 6: Customer Analytics 101

Chapter 7: Data Science

Chapter 8: Data Governance

Conclusion

Chapter 1: The New Rules of Competition

Companies compete for the hearts, minds, and wallets of customers. Most companies play by an older set of rules, gathering market research, setting up multi-year strategic plans, and executing those plans with little deviation. However a new generation of companies has emerged and changed those rules. Using short learning cycles and rapid execution, these New Competitors play by New Rules to deliver better customer experiences.

The New Competitors

Some of the best examples of New Competitors are Amazon, Uber, and AirBnb. These companies have risen to prominence in a short amount of time and made a significant impact on their respective industries. In less than 20 years Amazon disrupted the book industry (especially Borders and Barnes & Noble), then major retailers (Target, Wal-Mart, and others) and is now the global leader in cloud computing. Uber became a global transportation force in less than five years, putting local taxi and limo companies on defense and spawning well-funded copycats. AirBnb is revolutionizing the short-term lodging market, capturing share from hotels and changing the way people think about travel.

These companies are successful not because they have brilliant multi-year strategic plans; they are successful because they obey three New Rules: 1) they relentlessly focus on the customer experience, 2) they treat learning as a strategy, and 3) they execute on those learnings in an extremely agile fashion. It could be said that in essence *the long-term plans of the New Competitors are written by their customers every day*.

The New Rules

The New Competitors dominate because of their adherence to the New Rules. They use the New Rules to deliver more value to their customers, increasing customer loyalty and share of wallet. While the New Rules are not a big secret, their competitors have a hard time following.

Rule 1: Customer Experience is the New Brand

The New Competitors know that customers want their brands to deliver exceptional experiences. They put substantial effort into delivering a consistent experience across every touchpoint such as web, mobile, social, email, customer support, and kiosks. They focus relentlessly on how their experience shows up for the customer in a variety of situations. They know that if they offer a consistently great experience in every channel, the customer will have a high opinion of their brand. Factors such as price still matter, but other factors such as convenience, ease of use, personalization, and access to competent support now matter just as much.

The New Competitors also understand that a bad customer experience can spread like wildfire on social media. They know that their brand is only as good as their weakest channel, and that one bad interaction can color the customer’s perception of their brand.

They know they must not only deliver a consistent, personalized experience across every channel, but react competently when something goes wrong.

Rule 2: Learning is the New Strategy

The New Competitors view accumulated learning as a strategic asset. They devote considerable time and effort to building this asset. As this asset grows it becomes more difficult for their competitors to copy, which creates a sustainable competitive advantage. The New Competitors excel not only at accumulating learning, but at digesting fresh learnings in order to spot shifts in customer preferences before their competitors do.

The fuel for learning is data – the more the better – which has led to the rise of Big Data Analytics. Big Data is realtime, comprehensive feedback from the brand's customer base. Brands that collect large amounts of data, learn quickly from that data, and deploy those learnings rapidly into their customer experience are more competitive. The learnings the New Competitors generate using Big Data Analytics allow them to constantly modify and improve their experiences over the course of many short-cycle changes. The mantra of the New Competitors is to *learn fast* from their customers, and *act fast* on what they learn. Which leads to the third rule.

Rule 3: Rapid Execution is the New Differentiator

The New Competitors understand that learning is only valuable if it is applied, and they are great at applying learnings rapidly. It's not uncommon for the New Competitors to act on learnings within minutes, hours, or days, resulting in an immediate change to the customer experience.

Execution at the New Competitors is tightly coupled with customer feedback. They employ techniques such as Experimentation (a.k.a. A/B Testing) and advanced segmentation to help them learn what different types of customers want, down to the individual customer level. They do this at a speed that helps them learn faster than their competitors.

Executing these advanced techniques is difficult at high scale, so the New Competitors have built their companies specifically for rapid execution. They leverage tools such as agile development, DevOps, cloud computing, and software-as-a-service (SaaS) to enable them to deliver experiences that can evolve quickly.

Adopting the New Rules

Big Data Analytics is at the core of all three New Rules. It's impossible for the New Competitors to play by the New Rules without Big Data. Any company that wants to play by the New Rules will have to bet big on Big Data.

Betting big on Big Data begins in the C-suite. The CEO, CIO, CMO, and CFO must be aligned on the need for Big Data Analytics, and how this capability will be used to improve customer experience and business performance. The CMO and CIO in particular must have a shared vision of what the customer experience needs to become, and what their respective organizations will do to enable this experience. They should regularly review progress and reinforce key priorities with team members. The CIO should ensure their organization can implement rapid changes to the experience by adopting agile methodologies and technologies.

At a more tactical level, data processes must be re-designed with the customer at the center. Big Data should be collected at the customer level, organized at the customer level, analyzed at the customer level, and applied at the customer level. Related business processes should adapt to the customer learning process, not vice versa. For example, if the Finance budgeting process is out of sync with the customer learning process, the Finance process needs to change.

For many companies changes to culture and personnel may be needed. Data-driven cultures are not easy to create, and require analytical thinkers across every discipline. Functional leaders should hold their teams and each other accountable to understanding the customer experience, measuring the experience correctly, and constantly using data to make the experience better. Some organizations may need to import data-driven DNA from outside the company, and use it to cross-pollinate and sometimes replace existing personnel.

For most companies the changes described above are difficult and will take time to complete. But these changes are absolutely necessary to play by the New Rules.

Summary

New Competitors have emerged in several industries and are putting incumbent players on defense. These competitors are winning by playing by New Rules – focusing on every aspect of the customer experience, fueled by data-driven learning programs, applied rapidly via agile execution. Big Data Analytics plays a crucial role in helping these New

Competitors play by New Rules and win in their respective markets. Any company can adopt the New Rules, but they must make a significant commitment to Big Data, and change their processes, culture, and even personnel as a result.

Chapter 2: The Components of a Big Data Analytics Program

An effective Big Data Analytics program enables a company to play by the New Rules and create great customer experiences. But implementing such a program is not trivial; leaders must focus on the key components – people, processes, and priorities – to get the best result. When set up correctly, a Big Data Analytics program is the foundation that enables a company to play by the New Rules.

The Key Components

A Big Data Analytics program contains three key components: 1) the **people** on the team, 2) the **processes** by which the team operates, and 3) the **priorities** on which the team focuses. These components are known as the ***Three Ps of Big Data Analytics***:

1. **People.** Selecting the right people for the Big Data Analytics team is the most critical component of the Three Ps. The right people will make or break a Big Data Analytics program much more than the right technologies or even the right data. For this reason, it's crucial to place extra focus on the people component to ensure it is set up correctly. Specific roles to fill, and skills to look for when assembling the team are discussed in Chapter 3.

2. **Processes.** Implementing key learning and decision processes correctly within the team is another crucial aspect of a Big Data Analytics program. The correct processes ensure that each person on the team drives or participates in the right activities, and that each decision follows an approach that will yield the best result. Steps for implementing Big Data Analytics processes are discussed in Chapter 4.

3. **Priorities**. Establishing the correct priorities ensures the team has a clear direction and that stakeholder teams understand what the Big Data team will produce. This component integrates with key business processes to provide maximum alignment and impact across the organization. Setting and maintaining the priorities for a Big Data Analytics team is discussed in Chapter 5.

While each component is unique, all three must work together to drive a successful program. Getting the three Ps right in a Big Data Analytics program gives the program the highest chance of success.

The Other Components

Other components play an important role in a Big Data Analytics program, but their roles are subsidiary to the Three Ps:

- **Technologies.** Technologies are a major investment for a Big Data Analytics program, but they should be determined by the People, Processes, and Priorities put into place. The right technologies depend on a variety of factors, including data sizes, analytical speed required, expertise on staff, features, affordability, and so on.
- **Tools.** Analytical tools are also important to a Big Data Analytics program, but they are secondary to People and Processes. Analytical tools are one of the most fungible aspects of the entire program and can be added, removed, and modified very quickly. There is no one “right” analytics tool for every organization – in fact, the most successful Big Data Analytics teams use multiple analytical tools, each suited for a particular purpose.
- **Data.** Data itself is a function of the people that set up the instrumentation, collection, and data management processes. Data must be gathered, transformed into learning, and applied to the customer experience on a timely basis. Third party data may be incorporated, but most data from a Big Data Analytics program comes from direct measurement of the customer experience.

Basically, once the three Ps are set, Technologies, Tools, and Data follow. Placing any of these three components before People, Processes, or Priorities may cause decisions to be re-opened, core technologies to be replaced, etc.

Adapting to the Three Ps

Most companies already have analysts in Finance, Sales, Marketing, Operations, HR, and Product performing activities such as reporting, valuation, forecasting, and build vs. buy analysis. These activities are valuable and should continue.

Big Data Analytics is different, making it difficult for people and processes from existing functions to move into the Big Data Analytics world. Many companies succumb to common mistakes such as:

- Moving an analyst from Finance or Marketing into the Big Data Analytics team without understanding the analyst's technical aptitude. Many analysts are smart with data, but often lack the technical skill to work with Big Data.
- Moving a good software developer from the IT team into the Big Data Analytics team without understanding their business aptitude. Developers have strong technology acumen, but economics and customer behavior concepts may be foreign to them.
- Trying to adapt Big Data Analytics processes to existing business processes (such as a budget process.) Since most business processes were not designed under the New Rules, they will naturally drag the Big Data Analytics program into their slower-moving sphere. The better approach is to adapt the old processes to the New Rules with the Big Data program at the center.
- Trying to focus Big Data Analytics priorities on one group over another. Some stakeholder groups become favored because of more engaged leadership, more funding, etc. The impact of the Big Data Analytics program should be felt by all groups equally, with all groups preferably engaging at high levels with the Big Data program. For example, if the Marketing team is highly engaged with the Big Data Analytics program but the IT team is not, the program will be less successful.

For Big Data Analytics to truly become a strategic asset for a company, every relevant function must adapt to it. It makes no sense to build a strong Big Data Analytics program if other functions in the company don't understand or utilize its outputs, or can't do so in a timely fashion.

Summary

A Big Data Analytics program is the foundation of a great customer experience. The three components of a successful Big Data Analytics program include the people on the team, the processes by which the team operates, and the priorities on which the team focuses. These Three Ps must work in concert to drive a successful Big Data program. The Big Data Analytics program must also be the center of gravity around which related people and processes operate.

Chapter 3: Building a Big Data Analytics Team

The first step in creating a Big Data Analytics Program is to assemble the right team. While the skills required for the team are vast, and the people with these skills difficult to find, this step will make or break the entire program. A Big Data Analytics team is made up of people with diverse skill sets who work in close collaboration to generate business impact.

The Key Skills

The best Big Data Analytics team members come from a variety of backgrounds and possess a wide range of skills. There are four hard skills and four soft skills that most Big Data Analytics professionals have in common. These “critical eight” skills are often hard to find in a single person; the good news is that some of these skills can be developed.

Hard Skills

The following hard skills are critical to the effectiveness of a Big Data Analytics team:

1. **Analysis skills.** Analysis skills are the core of the Big Data Analytics program. Analysis is the art and science of setting up an analytical question correctly, bringing the right data to bear against the question, synthesizing actionable insights from the data, and presenting these insights to decision makers. The ability to understand and use the right tools & techniques for the job is critically important. *Difficulty to develop: medium.*

2. **Data management skills.** Effective data management skills help make the Big Data Analytics team more efficient, and their work product more accurate. Data management skills include organizing, annotating, normalizing, and manipulating data. Every person on the team needs some level of data management skill, and certain team members will need a lot. *Difficulty to develop: easy.*

3. **Statistical analysis and modeling skills.** Statistical analysis and modeling skills help the Big Data Analytics team better understand customer behavior and predict what customers will do or need next. These skills involve both choosing the correct technique for each analysis, and knowing how to properly execute the chosen technique. These skills can be difficult to find; the good news is that only a portion of the team needs these skills. *Difficulty to develop: hard.*

4. **Coding (programming) skills.** From data acquisition to manipulation and insight generation, code-writing skills such as Java, Python, SQL, and C# help the Big Data Analytics team in multiple ways. Big Data is unwieldy, noisy, and messy. It can often only be accessed, manipulated, or understood at a granular level. Common analytical tools such as Excel or Tableau are not built for this activity. Until better tools arrive to solve this problem, some members of the Big Data Analytics team have to roll up their sleeves and write code. *Difficulty to develop: medium.*

Soft Skills

Soft skills are also important for Big Data Analytics professionals. There are four skills or traits to look for in prospective team members:

5. **Intellectual Curiosity.** The best analysts have insatiable intellectual curiosity, both for the projects they are assigned and for related questions and problems. The very best analysts can take a subject matter area and quickly master it, generating additional questions and greatly increasing the organization's understanding of the problem space in a short amount of time. *Difficulty to develop: nearly impossible.*

6. **Business Acumen.** The best analysts understand how businesses make money. They can break the economics of a business into its "P times Q" (Price times Quantity) components on the fly. They understand the business in terms of leading indicators, lagging indicators, drivers, and KPIs. They can lead whiteboard discussions on how to set up and approach an analytical question. They understand how their work fits into the bigger picture, and how to deliver analysis in ways that can be absorbed by the business. *Difficulty to develop: hard.*

7. **Communication Skills.** The best analysts can communicate with business leaders in the leader's language. A great analyst can translate both the question and answer into terms that a business leader can understand, and upon which they can take action. The best meetings are those in which the business leader is clear on what actions need to be taken and by when. These analysts can put themselves in the leader's shoes, and see the business from the leader's point of view. *Difficulty to develop: hard.*

8. **Customer Focus.** The best analysts care as much about the end customer as anyone else. They don't see themselves as "just the analyst" on the team, they see themselves as part of a value chain that impacts the customer experience in meaningful ways. They love seeing the results of their work show up in the customer experience, and care deeply about improving that experience. They are able to put themselves in the customer's shoes, and see the brand from the customer's point of view. *Difficulty to develop: medium.*

As noted, some of these skills are hard to develop. Communication skills for instance are hard to develop or coach into people – some people just don't have the knack for storytelling or concise communication. Intellectual curiosity is downright impossible to develop – it's either there or it's not. Thus it's often easier to find people with the harder-to-develop skills and train them on the easier-to-develop skills. Of course the easiest method is to find strong people who already have all of these characteristics, but such people are in short supply.

The Key Roles

In addition to recruiting people with the right skills, it's also important to create the right roles and put the right people in those roles. There are several necessary roles in a Big Data Analytics team:

- The **Analytics Leader** is an executive who has extensive experience with analytics, data management, and business management. The analytics leader may come from either an engineering (product engineering or IT) or a business background (usually Marketing or Finance.) Their role is to recruit, develop, and retain top talent for the team, manage the team's priorities, allocate resources, and use the team's outputs to drive and influence business performance. That last aspect is important, as the total impact of the team will depend entirely on the degree to which it can influence business decisions. Thus the leader's ability to communicate or directly drive business outcomes with analytical insights is crucial. The leader's title does not matter a great deal. At smaller companies the analytics leader role may be played part-time by a CMO, CFO, or CIO; at larger companies this role may be a dedicated role at the C-level or VP level. In addition to managing analytics resources, the leader may also manage the IT resources that build and manage the Big Data Analytics infrastructure. This role is usually one of the first to designate or hire so this person can hire the rest of the team.

- The **Analytics Manager** serves as the day-to-day interface between mid-level business leaders and the analytics team. This role manages the analytics team, sets the analytical agenda in collaboration with business leaders, frames analytical questions within the appropriate business context, approves analysis for release, and communicates & evangelizes findings to decision makers. The analytics manager is usually a highly experienced analyst who also has good people management skills, stakeholder management skills, and communication skills.

- The **Analytics Product Manager** is responsible for the platforms, technologies, and tools used by the Big Data Analytics team. This role collects and synthesizes requirements, develops implementation roadmaps, and ensures correct usage of released capabilities. The people in this role are often former analysts who understand the needs of the analytics team, but also understand how to apply technology to analytical needs. The best analytics product managers understand how to balance business requirements and technical constraints to create powerful yet highly usable analytics capabilities.

- The **Analyst** understands the company's business model, is great at synthesizing linear and non-linear insights from the data, and excels at data visualization and storytelling to business audiences. The analyst should have insatiable intellectual curiosity, and always be looking for insights that will improve the customer experience and give their company a leg up on the competition. These analysts should care as much about the end customer experience as anyone else at the company.

- The **Data Quality Analyst** runs quality assurance tests on the data, prepares & organizes the data for analysis, and keeps track of previous analysis and results. The DQ analyst works closely with the Data Extraction Specialist (described below) and related IT teams to ensure that the quality of the underlying data is trustworthy. This role is primarily an analyst who understands the business and has solid data acumen, and medium to high technical acumen. This role is explained more fully in Chapter 8.

- The **Statistician** consults on each relevant analytical project and ensures the business question is being set up and analyzed correctly from a statistical perspective. Not all analytical projects require a statistical consultation, but for those that do, this role is critical. For small analytics teams, this role can be shared with other functions or played by a part-time employee or contractor. In-depth knowledge of statistical software tools such as SAS, SPSS, or R are valuable for this role.

- The **Data Extraction Specialist** knows where to find the data in the Big Data infrastructure, knows how to extract Big Data accurately and quickly, and knows how to work around data issues. The specialist is an expert at extracting, joining, cleaning, and filtering large amounts of heterogeneous data. This role typically requires strong technical acumen and the ability to write scripts and code that perform specific tasks.

- The **Computer Programmer** codes models, algorithms, and automated data pipelines. The programmer should have a high degree of technical skill. For instance, for programmers working with Hadoop, an in-depth knowledge of Java and Python can be very valuable. The programmer also works closely with product engineering teams to implement analytical results into customer-facing systems.

- The **Data Scientist** focuses on performing customer-level data analytics and devising models and algorithms that can be implemented in customer-facing systems. The Data Scientist should possess most or all of the Critical Eight attributes described above, and should be heavily involved in an ongoing process of Experimentation. The role of the Data Scientist is explained more fully in Chapter 7.

The Analytics Leader must decide how to size the team, and how many of each role they need. While these answers are unique to each company, in general analysts and extraction specialists will comprise roughly half of the organization, with the remaining slots going to the other roles described above. Many companies also recruit multi-tool players, for instance analysts who are also good at extraction, data quality analysts who are also good at business analysis, etc. These types of players give the company a lot of flexibility as their needs evolve.

The question may arise, why not just populate the whole team with data scientists? This approach is less than ideal for two reasons. First, data science skills are not needed for work such as data preparation, data management, and data extraction. Data scientists are overpaid if they are doing this work, and most will not be happy doing it. Second, finding qualified data scientists can be difficult. The demand for data scientists in the market is greater than the supply, leading to higher prices. The chances a company can recruit a whole team of data scientists is very small, and if they could it would be very expensive.

Most companies are better off creating a data science *capability* leveraging a diverse team – call it a team approach to data science. Instead of finding all of the “critical eight” skills in one person, an alternate approach is to build a team that collectively possesses all of these skills. While being more feasible, this approach has the added benefit of being less expensive overall. It alleviates some of the pressure to find true data scientists, plus it keeps each person focused on what they do best. In addition, this approach can accommodate the addition of data scientists to the team as the opportunity arises. This team approach is currently in place at several Fortune 500 companies and has been proven to work well for Big Data Analytics teams.

Summary

The most crucial step in creating a Big Analytics Program is to hire the right people and put them in the right roles. There are many skills needed in a Big Data Analytics team. Some of these skills are hard to find, and in some cases hard to develop. The right mix of roles is unique to each company and should be determined by the analytics leader. Since most companies cannot create a team of data scientists, they may be better off creating a data science *capability* composed of diverse people that collectively possess the entire data science skill set.

Chapter 4: Running Big Data Analytics Processes

Once the Big Data Analytics team is assembled, it's time to create the analytics processes that will govern how the program and team will operate. It's important to ground these processes in the New Rules – they must be customer-centric, designed to generate accumulated learning, and geared towards agile execution. An effective Big Data Analytics process follows the same general flow each time, but cycle times may vary widely. In addition, one of the most important aspects of an effective process is to ensure the right questions are being asked throughout the process.

The Analytics Process

The analytics process governs the approach the team takes to every analytical question. Employing a consistent approach ensures that the team does not overlook key steps, and enables them to correctly set expectations with business stakeholders. This results in more accurate analysis, and more action being taken by the business. An effective analytics process follows a general three-step progression:

What Is Happening?

Also known as the Observe step, this step involves collecting timely and accurate data on what is happening at the customer level. Prior to Big Data, most understanding of the customer came from samples (focus groups and surveys) combined with individual transactions, most of which were not connected back to individual customers. With the advent of Big Data, companies can track a wide range of behaviors for every customer. The goal of the Observe step is to confirm the trustworthiness and completeness of the data, analyze the data for emerging trends and patterns, and annotate external events and other isolated phenomena.

Why Is It Happening?

Once an analyst knows what is happening, they turn their attention to finding the causes. Also known as the Understand step, this step is part art and part science. To understand why something is happening, the full context of the situation must be understood and drivers (causes) exposed and tested. For example, an increase in sales could be caused by changes in pricing, new promotions being launched, customer preferences changing, competitor activity changing, and seasonality – among other factors. It's impossible to know everything that customers are thinking, everything competitors are doing, and in some cases everything the brand is doing. Since the analyst cannot account for every variable in their analysis, they must fill in the gaps with their understanding of the business. The more accumulated knowledge the organization has built up, the less intuition is required in this step. Experimentation can be a helpful tool at this stage to isolate variables and understand their effect on specific outcomes. All of these analyses require appropriate framing and scoping of the question, and some may require deep data science skills.

What Should Be Done About It?

Applying the learnings gained from analysis is the most valuable of the three steps. Known as the Act step, this step requires taking action on the understanding derived from the previous step. This step requires skills most businesses are still developing, including data strategy, algorithm & pipeline development, data governance, customer experience design, and customer operations. Some actions are simple, such as when a competitor launches a new promotion and the company must respond with a promotion of their own. But some are very complicated, such as when customer preferences are shifting and the company must anticipate where those preferences are shifting to, and what should be done about it. In the latter case the best approach is to develop a range of options with expected impacts, and then use tools such as Experimentation to test the effectiveness and value of each option.

The **Observe->Understand->Act (OUA)** process described above is fairly universal and can be applied to nearly every analytical situation. When combined with the Scientific Method (Experimentation), the OUA process becomes very powerful (more on this concept in Chapter 7.) The OUA process should be ingrained in an organization such that the three steps can be performed in quick succession. The faster the process, the faster the learning, and the more effective the Big Data Analytics program.

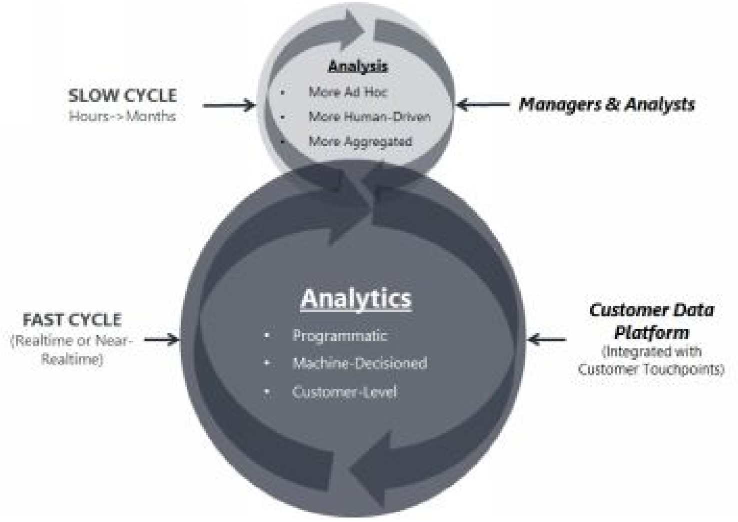
Analytics Cycle Time

One area that organizations struggle with is analytics cycle time. A mature Big Data Analytics program has two analytics cycles which intersect, yet run at dramatically different speeds.

- **Slow Cycle** “analysis” is driven by humans (managers & analysts.) Decision time can range from hours to months. The decisions to launch a new promotion, build a new datacenter, and enter a new market are examples of slow cycle time decisions.

- **Fast Cycle** “analytics” is driven mostly by machines, governed by rules or algorithms developed by humans. The decision time for fast cycle analytics is usually sub-second, much faster than a human can react. Decisions related to targeting and personalization at an individual customer level are examples of fast cycle analytics.

FIGURE 1: ANALYTICS CYCLE TIMES



The two cycles run somewhat independently, but they do intersect. For example, humans should regularly analyze and modify the rules or algorithms by which the fast cycle systems operate. Conversely, fast cycle learning such as experimentation can be fed into the analysis of new promotions, products, and markets. When operated correctly, the two cycles reinforce each other to drive higher business impact. Analytics leaders should work to understand the two cycle times in depth, and ensure the intersections are operating properly.

Key Questions to Ask

Great analytics leaders ask the right questions. They know what to analyze, and what not to analyze. Their experience guides how they frame and scope each question, such that not only can the question be analyzed, but the results can be acted upon. This is where the art and science of analysis meet.

At the analyst level, the same type of thinking also applies. The analyst should ask a standard set of questions before they begin the analysis, and after they complete the analysis.

Questions to Ask Prior to Analysis:

1. Do I understand the business question behind my analysis? What are the range of business actions that could be taken as a result of my analysis?

2. Is the analysis scoped correctly? Is it broad enough to capture all relevant questions, yet narrow enough to be completed in a reasonable timeframe?

3. Are there alternatives to doing the analysis? Has this same question been analyzed previously?

4. Do I have the data required to complete the analysis? If not, how will I acquire it?

5. Will experimentation be required to generate or validate key inputs? If so, can I get my experiments prioritized?

6. Are my data sets of high enough quality to produce confident analysis?

7. Do I have the tools, skills, and domain expertise required for this analysis?

8. Who will offer a second set of eyes on my analysis, and constructively challenge my conclusions?

9. Is this the most valuable next thing I should be doing to help the business move forward?

Key Questions to Ask After Analysis:

1. What tradeoffs had to be made during the analysis and how do I correctly expose these caveats to decision makers?

2. What alternative interpretations of my analysis exist, and which of those interpretations should be shared with decision makers?

3. What findings from past analyses should serve as context for my analysis, and

how do I best present my findings within that context?

4. How will decision makers take action on the results of my analysis? When will I need to follow up with them?

5. Will the recommendations I propose have potential adverse effects on other metrics, and if so how closely should those effects be monitored?

6. What changes need to be made to fast cycle systems, and what impact will those changes have on the customer experience?

7. How will I know if the actions taken were successful? What are the success metrics and what goals should be set?

8. How does my analysis add to the accumulated learning of the organization, and have I appropriately documented and shared this learning with affected stakeholders?

9. What additional analysis is needed to further the organization's understanding on the topic? What priority should that follow-on analysis take?

Analytics leaders should ensure these questions are being asked and answered by their analytics team. Doing so promotes consistency and completeness of analysis, which allows the program to grow stronger with each cycle.

Summary

A great analytics process is another crucial component of a Big Data Analytics program. An effective process follows the same general flow each time, though cycle times may vary depending on the type of analysis or analytics being performed. In addition, analysts should ask themselves standard questions both before and after the analysis to promote consistency and completeness.

Chapter 5: Setting Big Data Priorities

One of the greatest challenges for a Big Data Analytics team is the sheer volume and difficulty of requests that stakeholders can generate. While all of these requests may be important, they are not *equally* important, so the analytics team must develop a framework to prioritize each request and give it the appropriate amount of attention. A strong analytics prioritization framework can mean the difference between a burnt-out analytics team having too little impact, and an energized team having an outsized impact.

Creating an Analytics Prioritization Framework

Business priorities change constantly, and as a result analytics priorities will too. In fact, an effective Big Data Analytics program will drive changes to business priorities, which will in turn drive changes in priorities back to the analytics team. Thus the Big Data Analytics team should have a flexible prioritization process that keeps its analytics priorities in constant lockstep with business priorities.

Every analytics priority needs a clearly defined business question, an owner, a set of expected outcomes, assigned resources, and expected time to completion. Every priority should state who is going to use the results of the analysis and how they will use it. The analytics prioritization framework should contain clear rules on what can change and who is authorized to make changes. All priorities should be reviewed with stakeholders at least monthly or whenever significant changes occur.

There are 3 main steps to operating a prioritization framework for a Big Data Analytics program. Step 1 is a configuration step and performed less often, whereas steps 2 and 3 are operational steps that should be performed regularly in a tight iterative loop.

Step 1: Connect company goals to analytics goals. The business goals set at the company level should be shared with and understood by the Big Data Analytics team. The analytics team should break the business goals down into their component metrics, and determine which metrics are leading indicators, lagging indicators, drivers, and KPIs. From there the team can determine which metrics will receive the most analytical attention. Since the universe of analytical projects is endless, it's best to focus on fewer drivers that can be moved via analytics, and understand how each piece of analysis will contribute to moving that driver towards a numerical goal.

Step 2: Create an analytics agenda. The analytics agenda sets the priorities for the Big Data Analytics team and should be reviewed by executive stakeholders at least monthly. The analytics agenda is the customer learning plan, and should also be the vehicle that captures and shares accumulated learning across the organization. The agenda should be built around the big questions that need to be answered, over what timeframe, and the smaller questions that ladder to the bigger questions. The analytics agenda is co-owned by business leaders and analytics leaders, but the Big Data Analytics team should also incorporate its own thinking into its priorities given its unique perspective on the business.

Step 3: Follow through on actions. Taking action is part of the prioritization process because Big Data Analytics is a constant feedback loop. Many analyses require implementation to drive the learning that will lead to more understanding and new questions to answer. To turn data into action, there must be a mechanism that translates insights into changes in the customer experience (touchpoints such as web, mobile, kiosks, etc.) Each touchpoint presents an opportunity to learn more about the customer, experiment with new features and designs, and feed those learnings back into the prioritization process. For example, if five different actions were taken and the lowest-ranked action is outperforming the higher-ranked actions in a touchpoint, this information should be fed back into the prioritization process. The lowest-ranked action may need to move up the list, and may also generate new questions to add to the learning agenda.

A mature, well-run analytics prioritization process is foundational to running a business by the New Rules. The approach described above establishes Big Data Analytics as a strategic asset at the center of every customer interaction, which creates more value for the customer and the business.

The RBI Model for Big Data

One helpful way to manage the myriad of potential Big Data Analytics projects is to borrow a concept from baseball. In baseball, the goal is to score runs, whether from stringing together singles, doubles, and triples, or by hitting home runs. Every successful at-bat is considered a hit (absent a walk, error, or hit by pitch.) While it's tempting to swing for the fences on every at-bat, most teams employ a strategy of getting runners on base and driving those runs in by hitting RBIs (runs batted in.)

A similar approach can be applied to a Big Data Analytics program, where every project is treated as an at-bat or Attempt. The goal of each project is to get a Hit (single, double, triple, home run) and where possible string projects together for even greater impact. Just like in baseball, the final score doesn't count only home runs, it counts all runners (projects) that crossed the plate (drove cumulative impact.)

Just like baseball, the bigger the Hit the bigger the risk taken. Home run hitters typically strike out more often than hitters that hit for average. Just like a baseball team, a Big Data Analytics team is comprised of all types of players who may need coaching on which types of risks to take and when. The team should allocate their time to different types of Attempts in order to optimize their chances of driving large cumulative impacts. They should also avoid leaving projects (runners) stranded, meaning delivering promising early analysis on a topic but not following through with the remaining analysis required to drive the most cumulative impact.

Table 1: Big Data Projects in the RBI Model

Project Type	% of Attempts	Typical ROI*	Example Projects
Singles	50%	1-3x	• Consolidating data sources • Discovering new trends and seasonalities
Doubles	25%	3-5x	• Launching a new feature before a competitor • Performing cross-experience analytics • Deploying new

Triples	15%	5-10x	customer segments across touchpoints • Implementing a cross-channel optimization program
Home Runs	10%	>10x	• Implementing advanced targeting and personalization algorithms • Launching a new product before a competitor

** ROI is measured as the cost of the Big Data team performing the analysis, the cost of implementing the findings in terms of engineering and technology costs, and maintenance costs.*

As table 1 shows, a Big Data Analytics team should understand the risks of the projects they will undertake, and how to string projects together to generate the most impact. They should review and re-allocate risk regularly to ensure they are always making the best use of their Attempts. This approach resembles portfolio management, but differs in that it puts more focus on the cumulative impact that can be realized from projects delivered in succession. Cumulative learning is the very nature of Big Data Analytics.

Summary

A well-managed analytics prioritization framework can make a big difference in the impact the analytics team has on the business. To operate an analytics prioritization framework, first connect business goals to analytical goals, then create an analytics agenda with regular stakeholder reviews, then follow through to action and incorporate the resulting learnings into the future agenda. The team should also balance the risk of the projects they tackle to ensure they are spending time on the activities that generate the most cumulative impact on the business.

Chapter 6: Customer Analytics 101

Introduction

Customer analytics is one of the key activities performed by the Big Data Analytics team. The insights derived from effective customer analytics ultimately enable a brand to delight their customers and deliver more profits for the business. Even though customer analytics is different than traditional business analysis, most organizations will find they can make substantial progress on this activity by following a handful of tried and true techniques.

Cracking the Customer Analytics Code

Customer analytics is not easy, but parts of it *are* easier than most leaders think. When leaders think of customer analytics, they tend to think of data science, which is only partially true. Data scientists can create compelling insights from customer-level data, but there are also a variety of insights that can be produced by analysts who are not data scientists.

Here are 5 tips that can help non-data science analysts unleash the power of customer analytics.

Tip #1: De-Average Customer Data

Most traditional BI metrics are averages such as “average revenue per customer” and “average cost per transaction.” Such metrics are useful for dashboards and executive reporting, but are wholly insufficient for customer analytics. When data is averaged, information is lost that may be useful to understanding customer behavior.

Enter distribution analysis, which from a BI standpoint is the act of de-averaging customer data. Distribution analysis is one of the most powerful yet often overlooked techniques in analytics. De-averaging customer data exposes phenomena which are lurking below the surface that can drive additional insights into customer behavior.

Distribution analysis involves breaking a metric into frequencies along the customer axis. An example is the analysis of number of transactions per customer during the past year. While the average might be 10 transactions per customer, the range might be from 1 to 1000, and thus the distribution of this data is much more interesting than the straight average. 80% of customers may perform 1 transaction per year, while the other 20% perform an average of 50. Visualizing these differences in behaviors opens up whole new avenues of inquiry and understanding.

A common form of distribution analysis is to break the data into ten equally distributed parts called deciles and then average customer behavior in each decile. Deciles do a good job of exposing interesting customer phenomena, while grouping customers together to make the analysis more manageable and readable. If the analyst decides that looking at 10 deciles is too many, it can easily be collapsed into 5 groups (quintiles) or 4 groups (quartiles.) A decile analysis tends to highlight significant behavioral differences across the customer base, leading to shorthand concepts such as the 80/20 rule (80% of transactions are generated by 20% of customers.)

Decile analysis is also wonderfully easy to explain to non-analyst stakeholders.

Tip #2: Create Threshold Metrics

Distribution analysis provides a great foundation for creating threshold metrics. A threshold metric measures a percentage of customers that meet or exceed a particular pre-set threshold for a given metric. Examples include: what % of customers transacted 5 or more times last month? What % of customers bought more than $100 worth of goods? And so on. Threshold metrics can be applied in a variety of places throughout the customer experience. The goal of threshold metrics is to get decision makers and execution teams oriented on driving metrics above certain thresholds, with the purpose of bringing up the overall averages of the metric itself or related metrics (e.g. revenue.) Threshold metrics are different than customer segments, which pre-score customers at the beginning of the period and then keep that scoring throughout the reporting period. Threshold metrics are dynamic in that any customer can meet the threshold in a given time period, thereby giving business leaders an incentive to drive better performance throughout the period.

Tip #3: Segment Everything

Segmentation is a close cousin to distribution analysis and threshold metrics. Segmentation is the act of grouping customers together based on common behaviors and/or attitudes. Segmentation does not have to be overly sophisticated; any logical grouping of customers is better than none at all. One of the most common segmentation approaches is to group customers by low/medium/high engagement (or number of transactions per time period.) More on this topic in Chapter 7.

One of the powerful aspects of segmentation is that it can be written into the analytics infrastructure, which opens the door to even more powerful analysis. Once customers are assigned a segment in the analytical infrastructure (i.e. scored), these segments can be used as dimensions and cross-referenced with other key metrics. For example, a segment of “Young Moms” can be compared to a segment of “Rich Retirees” across every metric available in the analytics infrastructure. This type of analysis will show where similarities and differences occur, and will likely spark ideas on how to improve the experience for both segments. Marketing funnels, product & site usage, pathing, engagement, loyalty – everything can be analyzed via distribution analysis and multiple segmentation schemes simultaneously. In an ideal state, the same segmentation used by the Big Data Analytics team is also utilized by customer-facing touchpoints, allowing for analytical insights to drive consistent direct changes to the customer experience.

Tip #4: Think 3 Pivots Deep

After a few years on the job most analysts start to dream in dimensions. They begin to look at data in terms of what relates to what, and how they can use those relationships to uncover fascinating new insights.

Business leaders, however, rarely think in terms of dimensions (or "pivots" as they're known to Excel geeks.) They ask a business question but don't automatically think about the related questions that will pop up, or what related data will be needed to fully answer the question. Thus when an analyst is assembling an answer to a business question, they should immediately be thinking 3 pivots deep. It's almost always within 3 pivots deep that the full explanation to a business question can be found, and failing to go that deep may leave valuable insights uncovered.

Thinking 3 pivots deep means the analyst is looking 3 steps upstream at the drivers of the business question being asked. For instance, the business leader may ask, "why did we sell fewer widgets this month than last month?" The answer can usually be found within 3 pivots. First pivot widget sales by geography. Then within each geography, pivot by salesperson. Then within each salesperson, pivot by customer. In 90%+ of the cases, the answer will be found 3 pivots deep.

Applying this same approach to customer analytics is a little more challenging, but the principles still hold true. Big Data starts to come into play when the question becomes, "why did the buying behavior of Customer Segment X change?" Such a question involves going 3 pivots deep into the data and employing one or more of the techniques outlined in Chapter 7.

Tip #5: Use Sampling

One of the great things about Big Data is that it's big! Meaning that even small samples of it can be statistically valid. Sampling can be a powerful tool in an analyst's arsenal. There are certain analyses that lend themselves well to sampling, and effective sampling can greatly speed up the analytical process. Sophisticated Big Data Analytics teams perform the following steps to use sampling in their analytical process.

- They build & maintain robust sampled data sets that can be used across multiple projects.

- They balance and re-calibrate their samples regularly against the total population.

- They validate their sample-based analysis against the total population for major decisions.

Maintaining a balanced sampled data set allows analysts to quickly answer a variety of questions without tying up excess system resources each time. Sometimes all that's needed for an analysis is a quick, directional answer. Sampled data sets can be very valuable in this scenario. Other times, an analyst will start with a broad set of sampled data to help them narrow in on the data that will be most useful to analyze. At that point, they can decide whether to continue analyzing sampled data, or switch to the full population.

When setting up sampled data sets, it's important to enlist the help of a qualified statistician to make sure the data is sampled correctly and stays balanced over time. If the sample gets unbalanced, it will yield incorrect results (for example, it's common for heavy users to end up over-represented in the sample.) To understand the behavior of the population, or when comparing segments of users, the sample must represent all users in correct proportion. It's also important for analysts to explain their methodology to decision makers when they are presenting insights based on sampled data.

The Role of Storytelling

Many analysts think they have to become good storytellers to communicate their insights and recommendations effectively. Storytelling can certainly be helpful when communicating insights, but the degree to which this skill matters depends on company culture. The goal of the Big Data Analytics team is to generate insights that drive actions. The method of driving action is different within each company. Some companies have a high level of data savvy and don't want to hear a narrative – they want to study the data and come to the same conclusion as the analyst. Amazon and Facebook are such cultures. Other companies have the opposite cultures – where the analyst's ability to paint a picture with data helps reinforce their insights and generate action. Cultures can also vary between functions, for instance presenting to a CMO versus a CFO may require very different approaches. Both the storytelling and non-storytelling approaches have their pros and cons, but at the end of the day the analyst should tailor their approach to whatever generates the most action.

Summary

Customer analytics may sound daunting, but several components of this key activity are actually easier than most leaders think. While certain types of analysis require data scientists, other types such as distribution analysis, threshold metrics, segment analysis, deep dive analysis, and sample-based analysis can be performed by non-data science analysts. Storytelling can be a valuable tool for analysts, and should be used in a way that drives action within the company culture.

Chapter 7: Data Science

Data science is one of the hottest topics in Big Data Analytics yet probably one of the least understood. Data Science is the act of performing advanced analytics on deep (often customer-level) data. When performed correctly, data science activities can drive tremendous value for a Big Data Analytics team.

The Role of the Data Scientist

A data scientist is typically one of the most skilled analysts on the team, often possessing advanced degrees and significant domain expertise. The data scientist is also one of the highest paid analysts on the team, making their time very valuable. Thus the data scientist should focus on the types of analytical activities that only they can do; other activities (especially data preparation) should be performed by lower-paid analysts to the extent possible. This frees the data scientist to focus on the highest value add work.

Experimentation should be a top priority for data scientists. The scientific method involves creating a hypothesis and then testing it in the real world. Data scientists are uniquely qualified to perform this role (or why call them scientists?) Data scientists play a key role in driving a rapid learning agenda for their organization, using the scientific method.

Apart from Experimentation, Data Scientists should spend their time on Data Preparation (as little as possible), Segmentation, Longitudinal & Cohort Analysis, Optimization, Attribution Modeling, Path Analysis, Targeting & Personalization, and Predictive Analytics. While it's not possible to cover each activity in depth, the following sections describe each of these activities at a high level.

Data Preparation

Data preparation is a necessary activity for the Big Data Analytics team, but too often much of this load falls on data scientists (the New York Times estimates that 80-90% of a data scientist's time is spent on data preparation.) Analytics leaders should structure data science roles to focus on the highest value add activities, while minimizing data preparation time. Reducing the amount of time data scientists spend on data preparation can be accomplished by implementing better extraction tools, ensuring data quality earlier in the process, and staffing data extraction resources to assist data scientists with data preparation.

Segmentation

Segmentation involves assigning a customer to a group based on similar attitudes and behaviors. Customers in each segment can then be treated differently throughout the customer experience. Most segmentation schemes are applied to a company's entire user base and persist for a year or more, though exceptions exist. Data scientists are typically best at deciding which attributes to use for segmentation, and how to cluster or group customers into the right segments. And remember, the best customer segmentation is the one that gets used!

Some companies deploy multiple segmentation schemes against their customer base. This can be a risky proposition as it may result in the same customer being treated differently at different touchpoints. However, for companies that have strong internal coordination processes, multiple segmentation schemes do work (examples include different schemes for different geographies and for different customer bases.)

A strong segmentation scheme is especially powerful when incorporated into Big Data Analytics infrastructure. For segmentation schemes based primarily on behavioral data, customer-level data in the Big Data infrastructure can be scored with each customer being assigned a segment. The analytics team can then look at differences in all kinds of behaviors by segment, whether or not those behaviors were used as inputs for the segmentation. The analytics team can also study how customers move between segments over time, and analyze what factors influence those moves.

Combining Segmentation with Experimentation is even more powerful, as different types of experiments can be created for different segments, and treatment and control groups can be set up within the same segment to understand what certain segments of customers like and don't like. This approach helps an organization learn even faster, and helps them put an emphasis on the customers that bring them the most value. Experimentation is explained later in this chapter.

Longitudinal and Cohort Analysis

Longitudinal analysis and cohort analysis are two related analytical techniques that allow a Big Data Analytics team to better understand what is happening at the customer level. While not the same as a Segmentation (which is applied to an entire population), longitudinal and cohort analysis involve analyzing subsets of individual customers over time to understand what factors influence or contribute to certain behaviors.

Longitudinal analysis involves studying the same individuals over a given period of time and trying to capture as much information as possible to explain changes in their behavior. Longitudinal analysis can be used to uncover changes in behavior from customers that have been affected by a change in features, advertising, experiment treatments, etc.

Cohort analysis groups similar customers together typically around a single event that impacted each customer, to understand the impact that event had on behavior or for other analytical purposes. For example, an analyst could take a group of customers who signed up for a new service and track their behavior 3 months, 6 months, and 12 months after signing up. It doesn’t matter in which month each customer signed up, for every customer the month of signup is set to month 0 for the analysis. In the wireless carrier industry, churn by tenure is a common application of cohort analysis.

Cohorts are typically not written into the analytics infrastructure in the same way segmentation schemes are, but exceptions may exist for commonly-repeated analyses.

Experimentation

Experimentation (a.k.a. A/B Testing, Split Testing, Multivariate Testing, or Flight Testing) is a technique that allows an organization to make the experience more compelling for customers by varying small aspects of the experience, promoting what works and demoting what doesn't. Experimentation is an important part of the continual learning process, allowing an organization to learn fast and keep up with ever-changing customer demands.

Most experiments are performed on web sites, though an increasing number are deployed into mobile and kiosk environments. Classic examples of experiments include changing the size and colors of buttons, swapping different stock photos, and removing potential distractions from the page. An experiment involves identifying two groups of similar users, called a Treatment group and a Control group. The Treatment group is given the new experience (for example, a button is changed from orange to green); the Control group receives no change. After an acceptable period of time, the behaviors of the two groups are analyzed (for example, did the green button drive a higher clickthrough rate than the orange button?) If the Treatment group outperformed the Control group, the experience given to the Treatment group is promoted to become the primary experience. If the size of the customer base is substantial and analysis resources abundant, many experiments can be run at the same time.

One persistent challenge with Experimentation is minimizing confounding variables. Even with a well-controlled environment, it's often difficult to eliminate all other factors that may influence customer behavior and attribute precise behavioral changes to a particular treatment; however, directional and approximate effects can usually be observed and understood. The key to good experiments is to maintain good control groups, and to keep a sharp eye on external variables that could impact behavior in both the treatment and control groups.

Experimentation should be part of any DevOps or Agile development process. These development processes rely on constant customer feedback, and experimentation is a method of proactively generating quantitative, immediately actionable customer feedback.

Optimization

Optimization is an activity designed to balance all variables in a given portion of the journey to ensure each variable is functioning at its optimum level to drive specific outcomes (revenue, conversions, etc.) A good example is the optimization of a marketing conversion funnel. A marketing conversion funnel has several moving parts, all of which must work together to produce conversions at the end of the funnel. When the number of conversions are below goal, it's tempting to just feed the top of the funnel with more traffic. This may or may not be the right solution, as a larger problem could be traffic falling out of the funnel prior to conversion. An optimization approach looks at the entire funnel and through a process of analysis and experimentation determines which metrics need to change at every step of the funnel and by how much. Once those optimal values are attained for each metric the funnel is considered optimized, though the target values tend to change over time as customer demands change, as competition responds, etc. – making Optimization an ongoing activity.

One of the newer tools in the optimization arsenal is automation. Computers are very good at solving equations, and optimization is just an equation always looking to be solved as inputs change. Many companies optimize the SEO/SEM components of their marketing funnels using automated or semi-automated means; this approach is also expanding to cover more of the funnel. Big Data infrastructure is well suited for such activities as it allows a higher number of inputs to be analyzed and optimized.

Attribution Modeling

Attribution Modeling (a.k.a. Marketing Attribution Modeling) helps companies understand which of their marketing channels and creative assets are contributing to the customer's buying decision, and by how much. This information helps a company spend its marketing dollars more effectively. As with most online phenomena, customer demands, market conditions, and competition change quickly, and thus a marketing attribution model must be constantly updated with new data and learnings. Attribution Modeling can be combined with Experimentation to test new channels, messages, creative, and buying paths.

Path Analysis

Path analysis helps an organization understand how customers are traversing a journey and how to modify navigation to make such movement more efficient. Pathing is particularly useful in marketing funnel analysis, site analysis, and product analysis. However, customer paths can be challenging to analyze due to the sheer number of unique paths that can be taken. While basic pathing is available in standard web analytics tools such as Adobe Analytics and Google Analytics, complex experiences require analysis by a data scientist. Analyzing paths by customer segment is a very revealing activity.

Targeting and Personalization

Targeting and personalization are two of the hottest topics in marketing right now. Neither topic is new; both have been discussed regularly since the early days of the Internet. What is new is that technologies and techniques have gotten to the point where targeting and personalization at scale have become economically viable for most organizations.

Targeting has been occurring in the advertising space for over 10 years, first with content targeting, then with re-targeting, and then with customer-level behavioral targeting. Companies such as Google and BlueKai drove significant innovations in this space, and allowed marketers to reach customers in a context or mode where they are more willing to buy.

Personalization is very similar to customer-level targeting in that it tries to deliver the right experience to each customer, but personalization tends to be more of a two-way street where the customer agrees to be tracked (or even provides inputs) in exchange for a more personalized experience. Personalization also customizes the “non-commercial” portions of the experience (e.g. a list of menu shortcuts the customer prefers), resulting in an overall more pleasing experience. When performed correctly, personalization provides more positive benefits for the customer and makes them more loyal and likely to buy more.

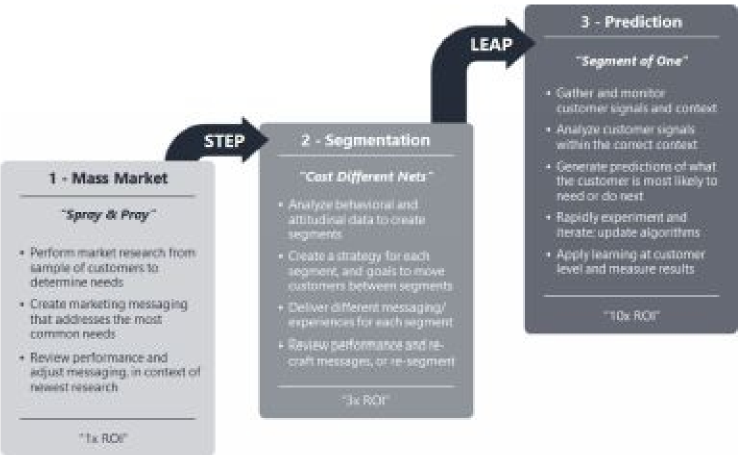
Data science teams play an active role in creating targeting and personalization algorithms, combining customer-level data from Big Data Analytics infrastructure and many of the techniques explained earlier in this chapter.

Predictive Analytics

The holy grail of Big Data Analytics is Predictive Analytics. Predictive Analytics is the ability to understand what the customer will need next, when they will need it, and what they will be willing to pay for it. Knowing these things allows an organization to programmatically deliver the right thing at the right place at the right time for the right price to the right customer. This results in happy customers and even happier shareholders. But being a holy grail, it's of course very challenging to attain.

Predictive Analytics is best approached as a process that starts with segmentation, and through analysis and experimentation evolves into finer and finer grain segmentation, until eventually each segment becomes a "Segment of One." The key is to discover which variables have the most predictive power, and how to harness these variables via various models. Tools such as machine learning become important at this stage, but all computer-driven processes require some level of human thought and intervention (e.g. machine learning is only as effective as its scope and the inputs provided; it's up to humans to figure out the right scope, which inputs to prioritize, and whether the inputs are even accurate in the first place.) The companies who will win are those who are able to feed their algorithms lots of good, clean data, and use a human-machine partnership at scale to develop the right algorithms for every customer.

Figure 2: The Path to Predictive Analytics



While Predictive Analytics algorithms are being fed by fresh, clean customer data, cycle time should continually be reduced to keep up with the speed of customer decisions. While moving from mass market to segmentation could be considered a step forward, the movement from segmentation to prediction is a giant leap forward, in terms of capability, difficulty, and payoff. Thus it should be treated as a long term quest and built atop a solid Big Data Analytics foundation.

Many other applications for Predictive Analytics exist which are beyond the scope of this book. The important thing to know is that Predictive Analytics is not possible without good, clean, comprehensive data available in a highly functional and highly governed Big Data Analytics infrastructure.

Summary

Data Science is the act of performing advanced analytics on deep (often customer-level) data and is a very valuable activity in a Big Data Analytics environment. Data scientists should focus on the types of analytical activities that only they can do, including Experimentation, Segmentation, Optimization, Attribution Modeling, Pathing, Targeting & Personalization, and Predictive Analytics – while minimizing time spent on Data Preparation.

Chapter 8: Data Governance

The final important aspect of an effective Big Data Analytics program is Data Governance. A Data Governance program ensures critical data sources and systems are operating with high degrees of accuracy, integrity, coordination, and compliance. A Data Governance program spans four distinct areas: 1) Data Quality, 2) Master Data Management, 3) Data Security, and 4) Data Privacy. A sound Data Governance program can make the difference between a mediocre Big Data Analytics program and a great one.

Data Quality

Inaccurate or incomplete data is the arch-enemy of a Big Data Analytics program. Bad data is at best valueless, and at worst harmful to the decision process as it can mislead decision makers. Since Big Data is inherently noisy and hard to manage, data quality can become a big issue in a Big Data Analytics environment if not controlled. Given the importance of maintaining accuracy in a Big Data Analytics environment, a dedicated Data Quality team is usually warranted.

A Data Quality (DQ) team is a dedicated analytics team whose charter is to proactively and reactively scrutinize granular data at every stage of the process on a regular basis. The DQ team keeps granular data clean and organized, and serves as first responders to data issues. Their mindset is that all data is guilty until proven *accurate*.

The DQ team is typically embedded in an analytical business function so they can maintain close ties to the business. For instance, when a new marketing campaign launches, it's expected that certain metrics will change. The Data Quality team should be monitoring these metrics closely as the campaign launches to ensure all measurements are working as expected, and that changes to data values have a correlation to the expected business impact of the campaign. Having a business foundation underlying DQ activities gives decision makers more confidence that the data correctly reflects actual business performance.

Justifying the cost of a Data Quality program is fairly easy – simply place a value on two things: 1) the time spent by downstream analysts cleaning dirty data, and 2) the time spent by decision-makers debating which data source is correct. Some organizations also factor in the impact of incorrect decisions and so-called decision paralysis. Assembling these costs typically makes for an easy business case for one or more dedicated DQ resources, and in some cases an entire team.

Data Quality Team Roles

A high-functioning Data Quality team is composed mainly of analysts who have strong technical skills, as well as dedicated specialists. The team breaks down into four main roles:

- **Data Quality Manager** – a seasoned analyst with good people management skills, strong technical skills, good business acumen, & good communication skills.

- **Data Quality Analyst** – a business analyst with medium to high technical skill (typically database skills) and medium to high business acumen.
- **Data Quality Automation Specialist** – a technical specialist experienced in building monitoring systems, alerting systems, and issue tracking systems.
- **Data Quality Process Specialist** – a business process specialist versed in methodologies that reduce errors in complex systems. Six sigma training is a plus here.

A typical team structure consists of 1 DQ Manager, 2 DQ Automation Specialists, 1 DQ Process Specialists, and a handful of DQ Analysts. DQ Analysts often specialize in business verticals and thus the number of analysts may be higher or lower depending on the size and complexity of the businesses they support. All other DQ roles should be designed to span multiple businesses or the entire organization.

Data Quality personnel should be hired for their understanding of measurement, their analytical prowess, and their technical skill. But they should also be hired for their mindset. They should be intellectually curious about how the business works, how business decision-makers interpret the data, and how the business makes decisions. DQ personnel don't just study the data systems, they also study how the data is interpreted and used by downstream users (including how it's consumed by customer touchpoints.) DQ personnel obsess about downstream users being efficient and not having to worry about data quality. They are the first to know when a quality issue emerges, and won't rest until the issue has been resolved.

Data Quality Team Responsibilities

The DQ team is charged with the regular scrutiny of granular data, and owns the diagnosis and resolution of data quality issues. The team's duties specifically include:

- Mapping out the end-to-end system of data flows to understand all sources and uses of data in the organization.
- Developing standards for data accuracy, and service level agreements (SLAs) with downstream users for key data sources and metrics. Users can be both human decision-makers, and customer touchpoints that require data from the Big Data Analytics environment.
- Developing automated monitoring processes that produce alerts when certain thresholds are exceeded, when data is missing, when data appears to be corrupt, etc.

- Regularly spot checking data looking for inconsistencies or values that vary from expectations.
- Diagnosing and fixing data quality issues in concert with IT teams; triaging, categorizing, and tracking data quality issues.
- Communicating data issues and estimated time to resolve to all affected stakeholders.
- Recommending long term data quality fixes to product owners and ensuring those fixes get prioritized appropriately.

Data Quality Process Mapping

Process mapping is a necessary part of the data quality remit, and helps determine where to perform regular analysis and where to implement automated monitoring. It can also help identify weaknesses in the system. The usefulness of process mapping also extends to Data Privacy & Data Security (described below.)

Data process mapping begins by identifying the sources & uses of data in the system, and producing solid documentation of each data source and what it is being used for. This exercise can be very illuminating and lead to productive discussions between system owners and stakeholders. From there, a more detailed version of the mapping is produced that shows the technical details of which data fields are passed, at what granularity, transformations that occur, who has access, etc. Organization-wide data mapping can be very valuable.

Automated Monitoring

Automation is one of the most effective tools in the DQ toolkit and absolutely necessary to monitoring a Big Data Analytics environment. The DQ team should set up a monitoring system with automated triggers and thresholds that can be fine-tuned over time. Various methodologies such as Six Sigma can be employed to help drive continual improvement and provide quantitative measurement of Big Data quality.

Data Quality Scorecard

A Data Quality Scorecard should be developed to track accuracy, completeness, latency and other key metrics at various points in the data value chain. These metrics are monitored hourly by the DQ team, and reported monthly to the executive stakeholders of the Big Data Analytics program. This scorecard helps instill confidence in the entire program and decision making organization wide. Over time, poll-based metrics such as data trustworthiness can be incorporated to illustrate downstream users' trust of the data.

Master Data Management

Master data management is the management of the attributes associated with the data itself, often referred to as “the data about the data”. This concept carries forward directly from the BI paradigm, and includes two main components: Metadata Management, and Data Mastering.

Metadata Management

Metadata management primarily covers how the data is defined, organized, and archived. Metadata management was already important in small data environments, but it becomes even more important in Big Data environments. At a minimum, an organization-wide Data Dictionary should be developed; as well, hierarchies and taxonomies should be coordinated across all relevant stakeholder teams.

Metadata management efforts should be spearheaded by the Big Data Analytics team but should have participants from across all functional data stakeholders (IT, Product, Sales, Marketing, Finance, HR, etc.) This permanent virtual team controls the Data Dictionary, hierarchies, and taxonomies, and administers a disciplined change control process to keep all functions aligned.

Data Mastering

A close cousin to Metadata Management is Data Mastering. Data Mastering is the act of assigning management attributes to data when it is created; a good example is assigning a numerical ID to a customer when they sign up for a service. If Data Mastering is not coordinated across product teams for instance, the same customer will end up with multiple IDs, which will complicate every cross-product analysis performed by the Big Data Analytics team and thwart attempts to provide a consistent customer experience.

IDs are also typically assigned to transactions, sessions, journeys, etc. All of these IDs must relate to one another, so that a single view of the customer and all of their activity can be reassembled in the Big Data Analytics environment. If an organization has multiple views of the same customer, they will not be able to play by the New Rules.

Data Privacy

Data Privacy is an important topic for every company that collects and utilizes customer-level data. Privacy initiatives are typically spearheaded by the CMO to drive policy, with the CIO driving technical implementation.

There are two types of private information that most Big Data Analytics teams deal with:

- **Personally Identifiable Information (PII)** is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context. A customer's name, social security number, IP address, birthdate, and telephone number are examples of PII.
- **Personal Information.** Personal information is different from PII, in that it is connected to a person, but not their identity. A good example is an anonymous web cookie, which many Internet services use to tell users apart on their service. An anonymous web cookie tells the service it's the same person that visited before, and may record certain preferences and personal settings, but the cookie is not tied to the person's identity (name, IP address, etc.)

Certain industries have more stringent privacy requirements, such as Healthcare, Financial Services, and companies that deal with children's data (under the age of 13.) Companies in these industries must take special care to ensure compliance with the requirements for their industry at a legal, business policy, and technical level.

Customer-level data must be used in a way that respects customer privacy. Some brands take a conservative approach to privacy and only take action on customer data when given explicit permission to do so; others take a more liberal approach and use data from all customers who haven't explicitly opted out. Brands should develop policies and practices consistent with the laws of their jurisdiction, their industry, and their brand promise.

Before taking action on individual customer data, it's always a good idea to benchmark against competition and engage qualified legal counsel. Since brands are required to publish their privacy policies on their websites, it's typically easy to survey the market to see how brands in certain categories treat customer privacy. In general, the better job a brand does using personalization and targeting to deliver experiences that are in the best interest of the customer, the more willing customers will be to allow their data to be used. Brands that abuse this privilege are typically penalized by customers, in the form of customers being less willing to share their data.

The IT team plays a critical role in implementing Privacy policies correctly. If certain data is not supposed to be available for certain purposes, the IT team should create technical barriers to prevent such access or usage from occurring. This is an ongoing effort since systems, data flows, and authorized users change over time. Relevant infrastructure needs to be monitored and tested regularly to ensure it stays in compliance. Sophisticated organizations run a barrage of test cases against their Big Data infrastructure regularly to see if any Privacy vulnerabilities have emerged.

Data Security

Big Data infrastructure should contain multiple defenses to ensure the data it contains remains safe and secure. Just like a company would protect any other asset, Big Data requires protection from both external threats and internal threats. Breaches can lead to identity theft or other pain for customers, and significant negative PR for the company.

The four lines of defense seen most often in Big Data infrastructure include:

- **External Firewalls.** The first line of defense is to prevent unauthorized external access to data stores via firewalls. Firewalls exist at virtually every company and make it difficult for external hackers to gain access to company infrastructure. All major cloud providers also offer the latest in intrusion prevention technologies.
- **Security Model.** The second line of defense is to establish a Data Security Model. A Data Security Model is implemented within the Big Data Analytics infrastructure and sets forth the permissions required for individual users and groups of users to access specific files, fields, and dimensions. The security model should be constantly updated as users come and go from the organization, and as files, fields, and dimensions change.
- **Encryption.** The third line of defense is to employ encryption of sensitive fields (e.g. PII.) Encryption prevents unauthorized users from seeing the contents of a data field, even if they gain access to it, unless they also obtain the correct decryption keys. Decryption keys should be kept in a separate location under separate security, and access to keys should be highly regulated.
- **Internal User Training & Monitoring.** The last line of defense is to train internal data users and monitor their activity as it pertains to access and usage of sensitive data. In rare cases an internal user will purposely steal or manipulate data, but most unauthorized access or use is inadvertent. Either case should be prevented, with processes designed to ensure that only authorized personnel are able to access data and use it in an approved fashion. The most common problem seen with data access is that users receive access to sensitive data for one project, and after the project is done the access is not removed. At a large company this bad practice can quickly add up to dozens or hundreds of users having unauthorized access to sensitive data, which presents a significant security risk.

As with Privacy, all of these Security defenses should be tested and reviewed regularly, to ensure that new vulnerabilities have not emerged, and that all users with access are still authorized. While maintaining robust security for a Big Data Analytics environment can be expensive, the potential costs of a breach are much higher.

Summary

A strong Data Governance program is an important part of a strong Big Data Analytics program. A Data Governance program spans Data Quality, Master Data Management, Data Security, and Data Privacy. All of these areas play an important role in keeping data accurate, organized, useful, and secure.

Conclusion

Big Data is here to stay. New Competitors, playing by New Rules, were the first to take advantage of Big Data's potential, but any company with the desire and appetite to harness this valuable new tool can reap its benefits. Companies who wish to play by the New Rules should focus on creating consistent, data-driven customer experiences, establishing an ongoing learning strategy, and executing in an agile fashion. The impact of these changes will delight customers and drive more long-term revenue and profit for the company.

Creating an effective Big Data Analytics program is not easy, but the payoffs can be profound, thus giving nearly every company in every industry the incentive to adopt a Big Data Analytics program.

For assistance in creating and operating a world-class Big Data Analytics program, please contact Society Consulting at info@societyconsulting.com or visit www.societyconsulting.com

The author can be reached at chad@societyconsulting.com