O'REILLY®

Compliments of Qubole®

Creating a Data-Driven Enterprise in Media

DataOps Insights from Comcast, Sling TV, and Turner Broadcasting

Ashish Thusoo &
Joydeep Sen Sarma

Creating a Data-Driven Enterprise in Media

DataOps Insights from Comcast, Sling TV, and Turner Broadcasting

Ashish Thusoo and Joydeep Sen Sarma

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Creating a Data-Driven Enterprise in Media
by Ashish Thusoo and Joydeep Sen Sarma

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Nicole Tache
Production Editor: Nicholas Adams
Copyeditor: Octal Publishing, Inc.
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

March 2018: First Edition

Revision History for the First Edition
2018-02-23: First Release

This work is part of a collaboration between O'Reilly and Qubole. See our *statement of editorial independence*.

978-1-491-99797-0

[LSI]

Table of Contents

CHAPTER 1

Data-Driven Disruption in the Media and Entertainment Industry: Trends, Challenges, and Opportunities

Until fairly recently, the media and entertainment industry's struggle to reach target audiences could still be characterized by the proverbial John Wanamaker quote. "Half the money I spend on advertising is wasted," he said more than a century ago. "The trouble is, I don't know which half."

It had almost become the industry's tagline.

But that is shifting—rapidly—because of big data and analytics. Media and entertainment companies have begun their data-driven journeys. For the first time, data is being used on a large scale to deliver the right content to the right people on the right platform at the right time.

A huge factor in this transformation is that media companies are focusing intently on consumers. Data is being used to personalize customers' consumption experiences by getting the precisely right content to them when and where they want it, on whatever device they happen to be using at the time. Data is also being used to keep the network performing as required by customers—even the so-called "last mile," which is the part of the network that actually delivers the content into consumers' homes, and which can be beyond

some media companies' control. And, most important, data is key to transforming the way media companies measure the success of their efforts.

The latter is a truly revolutionary change. Media firms—which include traditional broadcast and cable companies, digital outlets, and social media—are transforming the way they sell ads as well as create and program content. Rather than depending on outdated proxy metrics like gross rating points (GRPs), click-throughs, or impressions, they use big data and advanced analytics to sell business results. Instead of going for the highest number of eyeballs, they're going for increases in actual revenue.

Now *that's* revolutionary.

In this report, you'll learn about the trends, challenges, and opportunities facing players in the media and entertainment industry. You'll see how big data, advanced analytics, and a move toward *DataOps* (a concept we define in the next chapter) are influencing how three major media and technology companies—Sling TV, Turner Broadcasting, and Comcast—are proceeding on their data-driven journeys. And, you'll take away important best practices and lessons learned.

A Fragmented—but Growing—Industry

The global entertainment and media (E&M) industry reaped $1.9 trillion in revenues in 2016, and will increase revenues at an approximate 4.4 percent compound annual growth rate (CAGR) through 2020, to reach just under $2 trillion this year, according to PwC's Global Entertainment and Media Outlook for 2016-2020. This growth will be driven by E&M companies diversifying their offerings and channels as well as consumers' increasing strident demand for new content to consume, says PwC.

According to Deloitte, the way in which people consume media has changed dramatically over the past decade, creating both challenges and opportunities for traditional broadcasters and publishers and emerging digital players alike. Millennials today spend more time streaming content over the internet than watching it on television, and more than 20 percent of them habitually view videos on their mobile devices. Streaming services like Hulu and Netflix continue to flourish, with approximately 60 percent of consumers subscribing to

them. By 2021, 209 million people will be using video-on-demand services, up from the 181 million viewers in 2015. But it's a complicated scenario as well, which is keeping media companies on their toes. The latest Deloitte research shows that consumers will spend half a trillion dollars in 2018 alone streaming content live—with content being delivered on demand leveling off.

Other hot spots for media growth include ebooks, especially in education; digital music; broadcast and satellite television; and video games—including PC- and app-based as well as those written for online consoles.

But with consumers in the proverbial driver's seat, traditional business models are running out of gas. And a surprising number of people in the marketing community still don't necessarily see that anything's broken. They're about to get a wake-up call.

How Data Is Changing the Media Game

Broadcast television and traditional print media used to be easy ways for hundreds of billions of dollars to change hands. For a long time, those delivery channels worked. They created jet-turbine streams of demand for brands, enabling them to reliably reach virtually all targeted eyeballs.

Then, of course, customers ruined that. They fragmented their consumption habits. First through cable, and then streaming, and then spending more and more time using various digital devices to consume both video and textual content. Suddenly the reliable revenue machines of broadcasting and publishing began sputtering.

For these reasons and more, media companies are now under extraordinary pressure to turn to data-driven strategies. Then there are the following three issues that have made changing the existing business operating models an imperative:

Media companies increasingly lack control over last-mile delivery mechanisms and platforms

Unlike traditional media and entertainment scenarios, today's media companies often have little to no control over how their content reaches consumers. People could be using any combination of device and transport mechanisms to read or view content. Because of this, it is essential that media companies collect, analyze, and deploy operational data to flag potential problems

with a partner—whether a carrier, a device manufacturer, or an over-the-top service provider—that could affect the consumer. Putting data-driven self-healing systems in place using machine learning technologies is an increasingly common proactive stance media companies must take today to ensure that users can consume content when and how they want to without hiccups. (Note that among the companies profiled in this report, Comcast can be seen as a bit of an outlier. As a leading provider of entertainment as well as information and communications services, Comcast technically *does* own the last mile. Although Comcast owns NBCUniversal, this report discusses Comcast's broader data-driven initiatives as a media and technology company.)

Advertising budgets require hard ROI
: The latest CMO Survey found that 61 percent of CMOs are under pressure from their CEOs to prove that marketing adds value to the business. Media companies, in a chain reaction, are under the gun to provide hard evidence that placing advertising with them represents good business investments. In Jack Marshall's *Wall Street Journal* blog post, Facebook's vice president of measurements and insights, Brad Smallwood, is quoted as saying, "We're pushing the industry to actually think about business outcomes, and the causation marketing is driving as a success metric, as opposed to proxy metrics that aren't even particularly good to look at."

Data and analytics technologies are rapidly evolving
: From cloud infrastructure management solutions capable of helping media companies scale capacity, to advanced analytics that allow them to anticipate demand for advertising inventory, to AI-based corrections that make it possible for servers or network devices to meet performance service-level agreements (SLAs), technologies are emerging every month to help media companies accelerate their data-driven journeys. And new innovations are right around the corner. In fact, one of media companies' challenges will be tracking such innovations closely to see which ones might benefit them, and how.

But old ways die hard. Marketers are still following their budgets across stages of the customer journey from awareness and branding and acquisition, to retention and loyalty and the like. They're still treating each of those as separate and distinct stages as opposed to

part of a smooth continuum. And they're still treating their channels independently across display and video and mobile and social and native—and all of digital—relative to traditional. Each channel is tracked using separate key performance indicators (KPIs) that are really about inputs, not about results. With target rating points (TRPs) over here and click-through rates (CTRs) over there, media businesses aren't able to immediately grasp what the effects of content are on business results—and have begun to realize that all of the glowing prophecies of the promise of the digital age haven't caught up with reality.

As a result of wanting clearer, results-oriented metrics, most media companies are beginning to organize themselves around the customer—and to become omni-channel by design. They are beginning to understand that behind all those screens is just one person, and that they need to change their KPIs to reflect that. And they are finally at the point where they can think about attribution as a product. There's real appetite for this kind of sophistication—to point all available machinery at metrics that matter.

Three Areas of Opportunity for Media Companies

This new data-driven era offers opportunities to media companies in three technology areas in particular: cloud infrastructure, artificial intelligence, and analytics.

New (Cloud) Infrastructure Required

Although startups have the option of beginning with a clean infrastructure slate and can go directly to the cloud without stopping at "Go," updating legacy IT infrastructure is a challenge for older and larger media organizations. Why? In a word: scalability. The sheer size of the data, and the massive compute required to perform advanced analytics on this data, makes the cloud inevitable. Recent statistics from Ovum reflect this showing a rapid acceleration of cloud changes (see Figure 1-1).

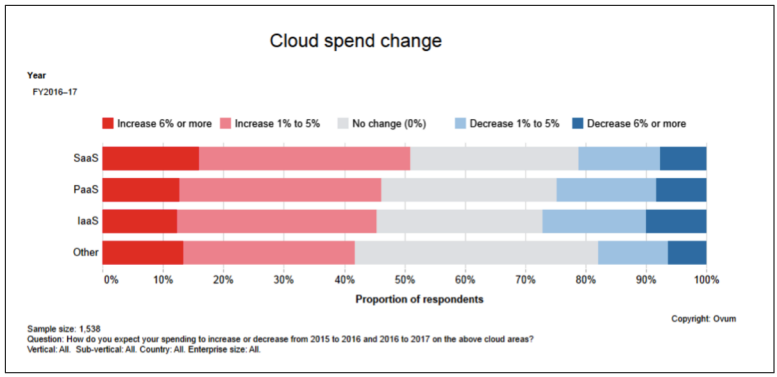

Figure 1-1. Cloud spend is anticipated to grow across industries

Of course, you still need to support your legacy environment during the transition to cloud and open source and the new way of thinking about data. Yet, don't be too slow about doing this. Failing to adapt quickly enough to infrastructure requirements of the new data-driven world will cause media companies that today are profitable to flounder.

Because, let's face it: the infrastructure on which the traditional industry model was built wasn't intended to handle today's data and analytics load. It's creaky. You have layers and layers of new flooring on an old 1940s house, and somebody has to get in there and rip it out and rebuild it. With its headers and pixels, and redirects, and Java scripts, it wasn't built for today's media business. That's not the way you build a trillion-dollar industry. You need to replatform your data environment in a modern, cloud-based infrastructure.

The fact that this is all still relatively new complicates matters. Innovations in big data and analytics and cloud technologies are emerging every day. Which to deploy? In many cases, your data environment is a sort of a Frankenstein's monster of pieces connected to other pieces connected to pieces that are beyond your control.

Media companies also need to carefully consider being creative about the potential of new, external sources of data. Social media generates terabytes of nontraditional, unstructured data in the form of conversations, photos, and video (Figure 1-2). Add to that the streams of data flowing in from sensors, monitored processes, and external sources ranging from local demographics to weather forecasts. One way to prompt broader thinking about potential data is to

ask, "What decisions could we make if we had all the information we need?"

Figure 1-2. Social media generates terabytes of nontraditional, unstructured data in the form of conversations, photos, and video (Source: Pixabay)

Artificial Intelligence: An Extraordinarily Promising Innovation

Artificial intelligence (AI) is obvious—and even a cliché at this point—when it comes to analyzing data and making predictions about everything from systems performance to consumer behavior. But there's an opportunity to go beyond what's been done with AI thus far and to begin using it for much more insight.

How much to pay for an impression is obviously incredibly important, but you also want to empower people with real insights around how to expand the boundaries of the product or service. Now that you know you can reach a certain persona, or audience segment, can you build something new for them? Can you speak to this audience in a different way than you had before? Now that you can break down averages and get to the individual behaviors and preferences themselves, AI can help you do your job better. For example, we'll see a programmatic-first approach to both media spend and programming content, where machine learning will drive optimization. You will be able to know and reach your customers with surgical

precision. It's about delivering a more seamless, relevant user experience that drives true outcome-based results.

Using Analytics to Drive True Personalization

The third and final opportunity is that offered by advanced analytics, which we can use to drive true personalization (Figure 1-3). Until now, the industry has done a pretty marginal job of making content compelling, personalized, and transparent. It's time to do that right.

Figure 1-3. We can use advanced analytics not only to glean valuable business insights, but to drive personalized recommendations for customers (Source: Pixabay)

Just look at the way screens have evolved. They've shrunk from auditorium-sized movie screens, to living-room television set screens, to PCs, to laptops, to tablets, and finally, to phones and watches. The phone is the ultimate personal screen. And it generates a ton of data that is just begging to be analyzed and put to use.

Think about it. A phone is meant to be used by one person. Your users watch it alone. It knows everything about them—their location, search history, even the hour-by-hour activities they've scheduled. When set up properly and within legal bounds of privacy, all of this rich information can be made sense of by using advanced ana-

lytics, which we can then take advantage of to send highly personalized content, advertising, and marketing directly to each user.

Initiating a Cultural Shift Across the Organization

In addition to the technological investments needed, it's also essential to pay attention to your organizational culture. Successful deployment of emerging data and analytics technologies is one thing; aligning them to the way people in your organization actually make decisions is another.

Make sure that you get your business users collaborating with your data scientists and analysts. Make sure your data infrastructure team works hand in hand with them, too. This is what big-data-as-a-service provider Qubole calls "DataOps," and it's an essential part of the puzzle. We discuss DataOps in more detail in Chapter 2. Yes, you will need sophisticated tools for data modeling, but you will also need intuitive reporting mechanisms for your users—and your management team—along with the right kind of training. The bottom line is that becoming data-driven needs to be carefully planned for true organizational change to occur.

Keep in mind that even with the most simple and intuitive tools, your users will probably need to enhance their analytical abilities. And management must make data a non-negotiable part of presenting in meetings as well as in explaining decisions and strategies.

Change is difficult for organizations, and becoming a data-driven company is the ultimate test of your change-management capabilities. This shift represents an upending of media marketing from an intuition-based discipline to a science-based discipline. From inputs to outcomes: from "I went into media marketing because I hate math," to, "If I don't have math skills, I can't do media marketing." The industry is changing, and that's scary for some people and exciting for others.

Yet the process of becoming data driven is a complicated one. Are your people organized the right way? Have you made it easy to move your data around? Is it easy to attach data and analytics to real use cases that show the value of what they do? Not just to infer the value—but to *prove* that the use of data and analytics actually improves

the quality of the customer experience and, ultimately, the profitability of your business as a whole?

This isn't easy. You need to attack the challenges of implementing DataOps on several levels—including training—and by using incentives to encourage the data-driven behaviors you want. And this cultural shift has to happen across the organization. Management must pay more than lip service to it, and must be prepared to act as role models to everyone else in the organization.

The best way to get started is just to get started. Baby steps.

First, you need to empower somebody to effect change. To take 1% of the budget and do things differently. And then effectively move it from 1% to 2% to 4% to 8% to 16%. If you just put one foot in front of the other, in five years you will have changed your organization.

Another, more effective, approach is to create a closed loop of a small initiative that does something different using data, proves that it works, and creates the justification for the next, bigger, step. Then, use that as evidence that data and analytics work, that change is possible, and that with more attention and more resources you can do more. Don't overreach and go from paralysis to overreaction. Be methodical in making the changes you desire.

Getting the Industry Up to Speed

All of this is happening at scale and very quickly right now. It's a boom time for big data and marketing in the media industry.

According to Joe Zawadzki, CEO of MediaMath, "When we started MediaMath in 2007, we knew the scope of the problem. But calling something 10 years too early is just as bad as not calling it at all." So, it was an issue of getting the industry from the current state to the future state, to chart the path of the company in such a fast-moving universe. "And, of course, we knew it was going to be disruptive for every department in every company in the world. It was just a question of how and when," he says.

Shortly after founding MediaMath, Zawadzki saw an interesting confluence of forces take shape, where the first of the technology-based media software models were appearing—and being acquired. Google bought DoubleClick. Yahoo bought Right Media. Microsoft bought aQuantive, Inc. and Advantium.

"We also saw data and media disaggregated, with the launch of BlueKai—which assembles media and data in the moment, as opposed to selling and buying it as a funneled solution," says Zawadzki.

And, finally, he began to see the emergence of an advertiser and agency community that was getting the sense that the current model wasn't sustainable. "The economics didn't make sense," Zawadzki said. "Because it was very hard to point to value using traditional metrics in this new world of digital and social and mobile."

Zawadzki thought, if businesses could start using data, now disaggregated from impressions, and start pointing it at deeper business goals rather than click-through rates, they could showcase a more effective model. "And once you deliver 10 times ROI relative to business as usual, you are on the threshold for positive disruption," he says.

How far has the world come in the last decade? "Maybe we're in the third inning of the ballgame; almost halfway to where we need to be," Zawadzki says. "We're at the end of the beginning, but it's going to take another decade to really embed data into everything we do."

Today, people are changing their organizations and their metrics, and they're willing to do things differently. "Call it a greed motivator, or an existential crisis, but media companies are finally realizing that unless they figure out how to change the way they use data—to create that direct connect between their products and services, and the human beings that will discover and ultimately be consuming them—they won't be around for long," says Zawadzki.

Even as this report was being written, entire marketing organizations are being rebuilt from the inside to make this a reality. And there are new configurations of partners that the world couldn't have imagined before, where media technology companies, data companies, agencies, and brands are all in the room together, sharing a common set of objectives as opposed to the fire brigade model of one person talking to one person, passing it onto another person, passing it onto another person. "Now, everyone is working off the same script," says Zawadzki. "It's not particularly comfortable. People wouldn't do it unless they were aware that the consequences of *not* doing it were extremely serious."

Get in the Game, or Get Out

Investing in modern data technologies and revamping corporate culture and business processes to reflect data-driven objectives are no longer in the category of "nice to have." Media companies are truly facing an existential moment.

If they aren't using data in their decision-making—to enable human beings to make higher-quality decisions, or to enable machines to do the same—they will struggle. Worse, they will fail as businesses.

Thus, the opportunities are huge, the technology has become available, and if you don't get in the game, you're dead. Those are all good motivators. Most media companies have realized this. They are engaging in one-to-one conversations with consumers, customers, and prospects across display, social, mobile, and video channels. And they're focused on real business outcomes rather than user clicks.

Next, let's define exactly what we mean by "data-driven organization" and how such organizations have used DataOps to get where they are today. Then, we'll learn how three real-world media companies are endeavoring to become data driven.

CHAPTER 2

A Brief Primer on Data-Driven Organizations and DataOps

If you're reading this report, you probably already agree that data is important to your business. You might already have a data-driven organization and are simply curious about how companies in the media and entertainment are coping with big data.

Or, you could be starting your data-driven journey. Either way, this report will be highly informative and useful to you.

Let's first define what a data-driven organization is:

> A data-driven organization is one that understands the importance of data. It has an organizational culture that requires all business decisions to be backed up by data.

Note the word *all*. In a data-driven organization, no one comes to a meeting armed only with intuition. The person with the superior title or largest salary doesn't win the discussion. Facts do. Numbers. Quantitative analyses. Stuff backed up by data.

Why become a data-driven company? Because it pays off. The MIT Center for Digital Business asked 330 companies about their data analytics and business decision-making processes. It found that the more companies characterized themselves as data driven, the better they performed on objective measures of financial and operational success.

But how do you become a data-driven company? This is something that we address in our book *Creating a Data-Driven Enterprise with*

DataOps. As we discuss in that book, despite the benefits of becoming a data-driven culture, actually getting there can be difficult. It requires a major shift in the thinking and business practices of all employees at an organization. Any bottlenecks between the employees who need data and the keepers of data must be completely eliminated. This is probably why only two percent of companies in the MIT report believe that attempts to transform their companies using data have had a "broad, positive impact."

The Emergence of DataOps

Once upon a time, corporate developers and IT operations professionals worked separately, in heavily armored silos. Developers wrote application code and "threw it over the wall" to the operations team, who then were responsible for making sure the applications worked when users actually had them in their hands. This was never a great way to work, for obvious reasons. But it soon became impossible. The internet had arrived. Businesses were now developing web apps. In the fast-paced digital world, they needed to roll out fresh code and updates to production rapidly. And it all had to work seamlessly.

Unfortunately, it often didn't.

So, organizations are now embracing a set of best practices known as *DevOps* that improve coordination between developers and the operations team. DevOps is the practice of combining software engineering, quality assurance (QA), and operations into a single, agile organization. The practice is changing the way applications—particularly web apps—are developed and deployed within businesses.

Now a similar model, called *DataOps*, is changing the way data is collected, stored, analyzed, and consumed.

Here's a working definition of DataOps:

> DataOps is a new way of managing data that promotes communication between, and integration of, formerly siloed data, teams, and systems. It takes advantage of process change, organizational realignment, and technology to facilitate relationships between everyone who handles data: developers, data engineers, data scientists, analysts, and business users. DataOps closely connects the people who collect and prepare the data, those who analyze the

data, and those who put the findings from those analyses to good business use.

The aspirations for a data-driven enterprise are similar to those that follow the DataOps model. At the core of the data-driven enterprise are executive support, a centralized data infrastructure, and democratized data access. All of these things are enabled by DataOps.

Two trends in particular are creating the need for DataOps:

Organizations need to possess more agility with data
: Businesses today run at a very fast pace, so if data is not moving at the same pace, it is simply eliminated from the decision-making process. That's obviously a big problem.

Data is becoming more mainstream
: This ties back to the fact that in today's world there is a proliferation of data sources because of all the advancements in collection: new apps, sensors on the Internet of Things (IoT), and social media. There's also the increasing realization that data can be a competitive advantage. As data becomes mainstream, more businesses see that they must democratize and make it accessible.

DataOps has therefore become a critical discipline for any IT organization that wants to survive and thrive in a world in which real-time business intelligence is a competitive necessity.

The Data-Driven Maturity Model

How do companies move from traditional models to becoming data-driven enterprises using DataOps? Big-data-as-a-service provider Qubole has created a five-step maturity model that outlines the phases that a company typically goes through when it first encounters big data. Figure 2-1 depicts this model, followed by a description of each step.

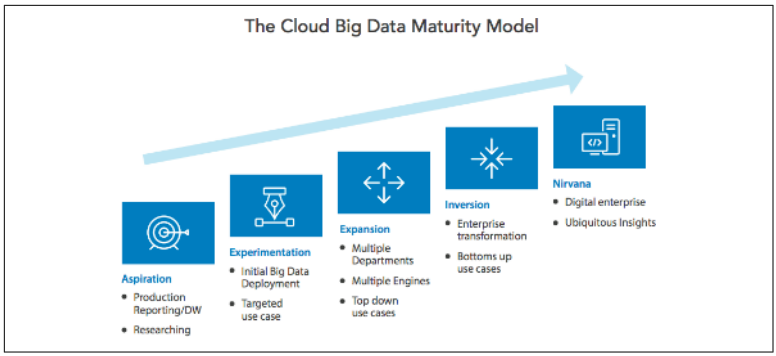

Figure 2-1. The Qubole Data-Driven Maturity Model (Source: Qubole)

Stage 1: Aspiration

At this stage, a company is typically using a traditional data warehouse with production reporting and ad hoc analyses. The classic sign of a Stage 1 company is that the data team acts as a conduit to the data, and all employees must to go through that team to access data. The key to getting from Stage 1 to Stage 2 is to not think too big. Rather than worrying about how to change to a DataOps culture, begin by focusing on one business problem you have that might be solved by a big data initiative.

Stage 2: Experiment

In this stage, you deploy your first big data initiative. This is typically small and targeted at one specific problem that you hope to solve. You know you're in Stage 2 if you have successfully identified a big data initiative. The project should have a name, a business objective, and an executive sponsor.

Stage 3: Expansion

In this stage, multiple projects are using big data, so you have the foundation for a big data infrastructure. You have created a roadmap for building out teams to support the environment. You also face a plethora of possible projects. These typically are "top-down" projects—that is, they come from high up in the organization, from executives or directors.

Stage 4: Inversion

It is at this stage that you achieve enterprise transformation and begin seeing "bottom-up" use cases—meaning employees are identifying projects for big data themselves rather than depending on executives to commission them. All of this is good. But

there is still pain. You know you are in Stage 4 if you have spent many months building a cluster and have invested a considerable amount of money, but you no longer feel in control.

Stage 5: Nirvana
: If you've reached this stage, you're on par with the Facebooks and Googles of the world. You are a truly data-driven enterprise with ubiquitous insights. Your business has been successfully transformed.

Where Are You in the Maturity Model?

After you determine where you sit on the maturity model, what do you do? For answers, we asked three leading media and entertainment companies—Sling TV (a Dish company), Turner Broadcasting, and Comcast—to tell us about their data-driven journeys. Read about them in the following chapters.

CHAPTER 3

Sling TV: Providing "Big Data on Demand" for Users and Systems

Sling TV is a leading over-the-top (OTT) live streaming content platform that delivers live TV and on-demand entertainment instantly to a variety of smart televisions, tablets, game consoles, computers, smartphones, and streaming devices. Sling TV, a subsidiary of DISH Network Corporation, is a *virtual multichannel video programming distributor* (vMVPD)—allowing today's most popular channels to be viewed through the Sling TV application.

As the cloud-native and big data evangelist at Sling TV, Brad Linder leads three teams: Big Data & Analytics, Cloud Native Engineering, and a Client Middleware Development team. In total, that accounts for about 50 employees. There are currently 12 people on the Big Data team.

"We are doing some really cool stuff with some pretty awesome technology," says Linder. "There is no better example than Sling TV to demonstrate a unique and interesting use case where cloud-native and big data come together." Sling TV has millions of devices in the field talking back to it as it delivers video over the internet. As such, it faces a lot of unknowns. "When we deliver video over the internet, constant two-way communication is occurring. Millions of devices are talking to our backend systems, necessitating our infrastructure to be highly elastic and responsive. Our system must adapt," says Linder. "It is very exciting."

Sling TV depends on data to keep its competitive edge. It was first to market, having disrupted a legacy industry, and enjoyed having the "whole blue ocean" to itself for a while, according to Linder. "Now that it's a proven success, the competitors are coming. And they are not small fish," he says. Linder continues:

> The data is very much going to drive how we react. A lot of the focus from our product management organization is therefore on building customer-centric features. It is like, "Hey, we have this idea, let's try it out and see what folks think." We are trying to get to that true customer-first approach to product development. Data and analytics are key to that. Data will also drive A/B testing of application features and the evolution of our roadmap.

Linder is responsible for—among other things—next-generation platforms to enable web scale at Sling TV. This is no easy task. Launched in February of 2015, Sling was the aggregation of a number of acquisitions over the years by DISH Network, and has a number of legacy backend systems. The organization was also constantly accumulating more data to manage. This means it ended up with a number of disparate systems that were not connected in the right way. "This legacy environment is not a competitive advantage for us as a company," says Linder.

In fact, legacy systems are a common problem in the media and entertainment industry. Sling TV has a number of these major backend legacy systems, and all of them report back into a "semi-common" big data platform that provides data to run both technical and business operations for the company. But the data, at this point, is still difficult to access. Sling TV also has data siloed off within various functions within its organization. "We are ensuring the right groups have access to the data they need to make the best decisions for the company and our customers," says Linder. "We also want to use it to power the personalization and user-experience plans on our roadmap."

Sling TV's Current Data Landscape and Plans for Next-Generation Data Pipeline

On the technical operations side of Sling TV's data environment, the team has created a series of dashboards that enable the tech ops team to gain insights into the current state of the environment. It also has a number of log tools that help it drill into the data. "We

monitor the systems constantly so we are not having to *react* to any issue, rather, we are able to catch it before it becomes a customer-impacting problem," says Linder. Sling TV is beginning to get into predictive analytics, where, for example, the systems can detect an anomaly or particular kind of issue, and signals that a certain device needs attention, "versus waiting to find out the hard way when a major incident occurs that impacts our customers," says Linder. "Eventually, this anomaly detection engine will fire off an automated response to the issue identified and hopefully fix it without involving a human."

On the business operations side, some of the same dashboards are used for connected devices that give the team an overall picture of what is happening in the business. "We are trying to understand how customers are using our service, so we can expand what we offer and give them an overall better experience," says Linder, who adds that personalized content is the overall goal: enabling customers to find what they want to watch as quickly as possible with the highest quality of service.

The ultimate objective: use common platforms so data flows to the systems and people that need it with the appropriate access controls in place. Right now, Linder's team is in the process of standing up the first version of its next-generation data pipeline. "We are implementing industry-standard best practices," he says, adding that you would find the "usual suspects" among his data platform and tool choices: "Confluent for our Kafka services and the Elastic stack make up the core of our next generation big data pipelines. From a data lake perspective, we are evaluating a number of tools. At this point, we are trying to keep it simple and build a good foundation to build on," says Linder. "The data lake will be a big part of what we deliver in 2018 if things go to plan."

"Of course, there are always issues," Linder says. "Today, we are not able to access data right away." Linder's team is in process of changing that. "By enabling Kafka as the enterprise data bus, we enable appropriate groups and systems to get in line and ask for what they need," he says. "That is where we are starting on the systems side, so we can get data to the folks that need it, which in the end will help our company and our customers."

On the personalization side, Linder's team is compiling data so that Sling TV can present options to customers of what they might want

to view in the future. "Based on customers' previous viewing habits and behaviors, we can make intelligent recommendations based on what we think might be interesting as well as what is trending currently in the world," says Linder. "And we are trying to boil all of that data into a common resource that can be presented on demand via our cloud-native, next-generation enterprise service layer."

The Cloud as an Enabler of Infrastructure Elasticity

Sling TV is in the process of structuring all of its next-generation systems using a highly elastic, cloud-native architecture and infrastructure. "If we cannot do that, we are not going to be successful in our big data endeavors," says Linder. The next-generation pipeline will enable elastic workloads that go on and off the cloud as needed.

Data is everywhere in the Sling TV environment. Making the right data available in a cost-effective and highly scalable way is the goal. "We are trying to leverage the rapidly advancing cloud-native world to drive our big data roadmap. There is no question that elastic workloads are the way to go, but finding the right way to accomplish this is the main goal here," says Linder.

Helping Users Help Themselves

Sometimes, users know what data they want. Sometimes, they have an end result in mind and ask Linder's team specifically for certain data. However, sometimes Linder's team has to help them articulate what they're trying to do and reverse engineer the request to identify specifically what data they need.

Hiring the right team makes all the difference to the success of the Sling TV big data initiatives. Linder's data team is a relatively small one and hierarchically very flat. "We do not have a lot of levels because we are trying to find people who are driven enough that they do not need to be told what to do," he says. "Finding the right people is key, as is nurturing a highly collaborative environment. The folks we have found so far are wonderful and I cannot wait to see what else they will come up with."

Linder recently hired his first data scientist into his organization to help unite teams and bring the other data scientists in the broader

organization together. "I expect to expand this role so that people can come to us with business questions, and not have to understand engineering or advanced big data and artificial intelligence concepts to get what they need," Linder says. "We want the business user to feel comfortable saying, 'I wish I could know more about X or Y,' and we can say, 'Here is what we have. Is this enough?' It is all about building that collaborative environment."

The ultimate objective, of course, is self-service with proper access control. "We are trying to enable a 'Chinese food menu' of data for our internal customers and systems to order from. At the end of the day, we want to provide the best experience possible for our internal customers and systems that need to share the data we manage," says Linder.

Eventually, says Linder, Sling TV will have a data-driven culture that permeates all layers of the organization. "Right now, we are dependent on the expert knowledge of a few and the opinions of many," he says. "We do have some data available, but we need to get to the place of letting the changes in data trigger workflow." Data is also critical to the cloud-native data environment that Sling TV is trying to build. "We will need compute capacity on demand, and that will be triggered automatically based on near real-time data when we are done."

At any company, you are going to have marketing, engineering, and other types of data residing in individual silos, Linder says. "If we can bring all that together, that will be the best of all worlds. More data equals more value in my eyes. More value means a better customer experience, and that is why we are working on any of this. Encouraging and enabling various teams to share data easily is a major goal that we have."

On Not Owning the Last Mile

The media landscape has changed radically in the last decade. Many of the traditional businesses in this industry controlled the entire medium—from originating content, to the customer consuming it. For example, if you look at traditional DISH Network satellite television, the content comes in, hits a satellite, and comes down to the customer's house. "DISH owns the whole experience, down to the last mile," says Linder. "But in the case of Sling TV, content comes in, goes out over the internet, and then we have to rely on a lot of

factors outside of our control—such as the internet working properly—to deliver a good customer experience." Most emerging media and entertainment firms are in the same position, "trying to build resilient and scalable services that react to the unknowns we have to deal with is what we are working on," says Linder. "It is a really fun and exciting place to be."

Because of this lack of control, Sling TV cannot operate without automation, and without artificial intelligence tools to tell it, for example, that customers are having problems on the West Coast, and that it is probably one ISP (internet service provider) causing the trouble. "Although you obviously do not own it, you have to consider what you can do about it," says Linder. "We are looking into data and analytics solutions to help us make decisions in those cases."

On Jumping into the Data Lake

At this point, Sling TV is just starting to get its head around what it's enterprise data lake would look like. It is considering both Hadoop and some Amazon Web Services (AWS) tools. "The goal is to find us a robust, scalable solution to get us much more longevity," says Linder, who added that Sling TV is still in the early stages on that particular journey. "But the goal is to try to bring all of the disparate silos of data together into an enterprise data lake that we can leverage as a company," he says. "The data lake is our first step to truly robust machine learning applications, which will enable the personalized experience we want for our customers."

There is no cookie-cutter plan to follow for how to approach this big data challenge with the technology moving as fast as it is currently. "If we tried to sit down and figure it all out first, we would never get started," Linder says. "Taking an iterative approach to building this environment will get us to the places we want to go quicker and with less risk," he says. "So, we are taking a small, iterative approach. We try something, see what we learn, and improve it," he says. "We are measuring KPIs [key performance indicators]. We ask questions. And see what we can figure out together."

Using Data to Drive Business Decisions

Sling TV has already had some data-driven successes. It recently launched its cloud-based DVR service, and depended on big data and analytics to make sure it got it right.

"Once we got to the point of bringing this long-running, complex project to market, we had to test it thoroughly and truly understand the customer experience," says Linder. "This was a brand-new technology, and we needed to understand how well it worked."

First, Sling TV launched the DVR in a beta test to a limited number of customers. As they used the system, the team was constantly querying the beta users and collecting data on their responses. "What are you thinking? How is it working?" On the backend, as customers used the system, the team kept its eyes on key KPIs: How did the system work with one hundred people using it? One thousand? And then more. "As it grew we were able to learn from the data on both the business and the systems perspectives, iterate on changes, and tackle a number of challenges before we officially brought it to market," says Linder.

Encouraging a Data-Driven Culture

Linder has always remembered something he was told as an undergraduate student by his statistics professor. "He told the class that, 'If you do not have real data, you just have an opinion,'" says Linder. "And I believe that to my core, because even if I am the subject-matter expert, and I know 'everything' about this part of the business or this part of the system, I could be wrong because things are changing so quickly."

One of Linder's top goals is to help Sling TV avoid dependency on a few core subject matter experts (SMEs) to troubleshoot and solve all the issues it has. "Because although there are people who certainly know our service exceptionally well, wouldn't it be even better if their knowledge could be transcribed into key KPIs and metrics that everyone could understand and easily see?" he asks. That way, everyone in the organization would be aware that if a KPI suddenly jumps two or three standard deviations, that there is a problem. "Or, better yet, maybe that triggers an automated response to self-heal the system," says Linder.

At Sling TV, most people have by now realized the value of becoming a data-driven organization, says Linder. They are past the aspirational stage, and into the experimentation stage (Figure 3-1). "That is the phase we are in, taking the strategic vision and turning it into reality," says Linder, who says two of his teams are on the verge of the expansion stage, but they are an exception for the larger organization.

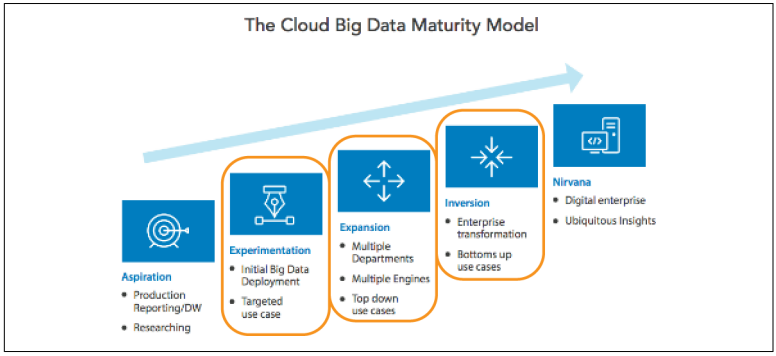

Figure 3-1. Sling TV is currently in the Experimentation stage of the Data-Driven Maturity Model

Numerous challenges accompany the changes required to move through this maturity model. "It all comes down to the fact that you are dealing with humans," says Linder. "People are people, and if you tell someone who has been doing things a certain way for 5, 10, maybe 20 years that everything they thought they knew is now different, you are going to get some resistance."

Because of this, Linder's teams are trying to take a partnership approach to getting people to accept the data-driven culture. It is important, he says, to include everyone as early on as possible so they do not feel that they are being dictated to.

Additionally, from a legacy-platform perspective, the current big data pipeline processing Sling TV data uses software installed many years ago. This software is going to be retired and replaced within three months of the new pipelines launching into production. "As I start to talk about it and evangelize it, people are getting excited," Linder says. "That is a good thing, but also leads to a problem I am okay having. People are getting excited before I can actually deliver and that is a challenge we are currently facing. However, you could have much worse problems, so I am not complaining!"

"We have to make 'data friends,' that is the best way I can look at it," he says. To make this a priority and bring the organization and its data together, a dedicated product owner has been tasked with finding data across the organization and bringing people together to hopefully get all the data into the enterprise data lake that is under design. "We have offices globally that need to be aligned and integrated. So, we are starting conversations about bringing their data together with ours. We need to make the most of these interdependencies and build those relationships."

Then There's Automation…

Linder is adamant: If you are not managing and processing your data in an automated way; if you are not taking the cloud-native approach to managing your application deployments, configurations, and runtimes; at some point, you are going to trip up. This is even more true if you are working at web scale. The question is not if, it is when.

"If you are in an environment that is highly elastic, like the Sling TV world, where you have huge expansions of traffic coming through an uncontrolled medium like the internet, you need to know that if there is a hiccup in the matrix, that you can handle it," he says.

What all this comes down to, he says, is: *you need automation*.

"We are putting together the building blocks to leverage machine learning," Linder says. "Our working strategy is to get as much data as possible into our environment. Once we have the makings of a proper enterprise data lake, we will be able to build some pretty amazing models and toolsets. I cannot wait to see what we can actually enable to provide a better customer experience."

Starting on Your Journey

How should you begin your data-driven journey? First, says Linder, you need to take an honest look in the mirror and understand what you have got. "If you cannot be honest with where you are currently, you will never get where you want to go. There is nothing wrong with where you are; everyone has to start somewhere. Be careful with any systems documentation you may have as well, they are typically outdated the minute they are completed. Also remember that these types of initiatives are not one size fits all. You need to 'get

your fingerprints on it' to make it your own and be successful," he says.

In the big data world, capacity management is critical, says Linder. He advises making sure to build a great capacity management team, because "typically, your infrastructure engineers do not appreciate sudden deviations to their models," he says. "If you come along and ask for another 100 virtual machines without warning, they are not going to like that. Do not be fooled if you are in the public cloud either—getting surprised with a big, unexpected bill is almost as painful."

Linder also advised finding key people in every functional area in the organization, and including them in the project from the beginning of the journey. "Because at that point, you have to begin nurturing that collaborative culture, while giving people the opportunity to contribute from day one so they do not feel that they're being dictated to once you bring your final toolset to market," he says.

This approach has paid off a hundred times over on both the cloud-native and the big data sides of Sling TV. "Both teams are encouraged to speak up," says Linder. "We ask each other, 'here is what we got. What do you think? What is missing? What else can we do? Does something in here seem absolutely crazy to you?' And if the other side says, 'no, that is broken,' make sure you dig in to understand why. The team we have built is amazing and I am lucky to be a part of it, really."

"We have to build a collaborative world because change is a given," says Linder. "We have seen just crazy changes in the last 18 months. And with the competition that is coming, we have to continue to innovate and let our customer experience and usage of our service drive where we go."

"Get ready and hold on, it's going to be a fun ride," Linder says.

CHAPTER 4

Turner Broadcasting Company: Dedicated to the Cloud for its Data-Driven Journey

Vikram Marathe is technical director of development and architecture for the Turner Data Cloud (TDC) at Turner Broadcasting. He's been at Turner for slightly more than four years. At the beginning of that time, Turner was just exploring how it could use data to better align with audiences. Then, in 2015, Turner created the TDC, which brings together numerous first- and third-party data sources with advanced analytics so that Turner can better anticipate consumer behavior.

Today, everybody in the industry now understands that the more data you have, the more useful it becomes, says Marathe. "And to gather more data, you need more ways to reach out to more consumers. That's why we're seeing so much consolidation in the media and entertainment industry right now."

The impetus for Turner to create the TDC was realizing that it needed to invest in technologies to build an integrated enterprise-wide big data platform. "The ultimate goal is to get a 360-degree view of consumers that encompasses understanding their past, present, and future behaviors," says Marathe. TDC supports Turner's initiatives in enabling Turner to tailor content, marketing, and advertisements based on its understanding of consumers' preferences. By bringing together numerous data sources, the TDC enables

Turner to deliver more relevant consumer experiences while offering advertisers more effective ways to reach target audiences.

Turner can now utilize this data to acquire or create new shows that are likely to appeal to the audiences that watch its portfolio of channels, which includes TBS, TNT, TCM, Adult Swim, CNN, HLN, and others. Turner has also been able to take advantage of this data to identify the right audiences for new Time Warner movies. These marketing campaigns are omni-channel and can target prospective audiences wherever they are, whether they are watching broadcast television or cable, or perusing digital channels like Facebook, Twitter, and Instagram.

"When we started leveraging TDC data it was quite uncommon for media and entertainment companies to be data driven. However, most media companies today either have similar initiatives or are embarking on them," says Marathe.

What Made Turner Turn Toward Data

Previously, like most cable channel operators, Turner didn't use data to make decisions. However, consumers today have many more options for entertainment beyond linear TV channels. Because of the nature of their businesses, video-on-demand services and digital-channel operators have tremendous amounts of data about each consumer's behavior at their fingertips. It thus became an imperative for Turner to better understand its customers. This was not an easy task, however, given that channel operators do not own the relationships with their customers. Their consumers happen to be customers of cable companies or telecom companies.

On the digital side of Turner's operations—that is, on its websites and apps—data was always important. Yet use of this data was limited compared to what Turner does today. "We considered certain metrics, but the bulk of advertising revenue at Turner originated from linear sources, and hence digital channels were used more to increase consumers' engagement with our linear TV operations," says Marathe.

Today, cable operators must compete with digital channels such as Facebook and Google for advertising dollars. These digital providers can target advertisements to specific segments of consumers, whereas advertising in linear television is still mostly based on dem-

ographics. Turner would sell advertising time slots to advertisers based on guaranteeing "a certain number of males or females, between certain age ranges." When those were the only options available for advertisers, they were not good enough.

As more and more consumers "cut the cord"—getting rid of their cable services—Turner and its competitors have to make sure that their channels are must-see propositions, and something that consumers can't do without. This shift has been going on for some time, but it has reached a point where the industry has to take notice and retool how it has done things in the past.

"Data is rapidly becoming central to everything Turner does," says Marathe, who added that the transition to a data-driven culture started approximately four years ago, and was one of the reasons he was recruited by Turner.

At that time, Turner was using Netezza as the underlying technology for its data platform. It performed some analytics for the digital side of the house, but that was more for research teams to publish stats for advertisers than for general business users to employ. For example, Turner would track how many visitors came to its sites, how many of them were unique, how many page views accrued, and other like metrics. "It really did not go too much beyond that," says Marathe.

But the week he arrived at Turner, his group began learning about Hadoop and figuring out how to create a data lake using the firm's new onsite Hadoop cluster. This platform became the firm's mainstay for the next two years. Then, Marathe began looking at other options, particularly moving to the cloud. "That's when I started looking at various vendors," he says. "Last year we moved to the cloud and began utilizing the Qubole platform."

Marathe notes that the data management industry has also been changing rapidly, during the same timeframe that data at Turner has become so prominent. Cloud adoption and cloud tools for data management have vastly improved in the past four years. "If you are an on-premises company, you can only store so much data. You have to keep adding more and more servers to your Hadoop platform," says Marathe. "That's not always easy to do, and it doesn't necessarily scale when you need it to scale. So, moving to the cloud was—and is—essential."

He adds that cloud technology has been sharply focused on data management, beyond its roots in just running apps and websites. Turner must use the knowledge it is getting from its new cloud-based big data analytics to provide more engaging content along with more relevant ads and marketing.

Turner currently has about 150 internal users for its TDC. The biggest advantage for these users is time to market of new shows, initiatives, and services. Previously, transforming data and creating new data structures was very time consuming. But now, because of the data lake with schema-on-read and the power of new big data processing technologies, users can just begin asking questions of the data. They can change what they're asking for, and Turner's team can quickly answer them without having to retransform that data into a different data structure.

"Happily, the days when it would take three months before we could get to a request, and another three months before we could deliver it to them, are long gone," he says.

Moving up the Big Data Maturity Model

Depending on the group and their use cases, data usage at Turner straddles the three middle stages of the big data maturity model: experimentation, expansion, and inversion (see Figure 4-1).

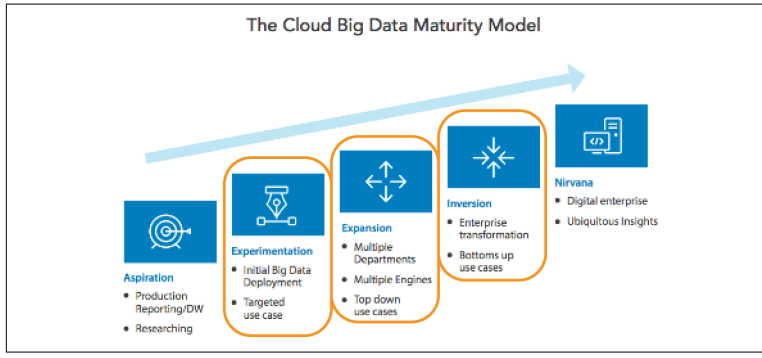

Figure 4-1. Turner Broadcasting straddles the three middle stages of the Data-Driven Maturity Model: experimentation, expansion, and inversion

"In the experimentation stage, we are bringing in new data sets all the time in to the Turner Data Cloud," says Marathe. "In these cases,

we need to figure out, 'how do we use these datasets?' How do we create new initiatives and services based on them?"

Also in the experimentation stage, Turner has datasets that multiple departments are already using. But these datasets are very large and very diverse, so business users are still figuring out how to use them, what types of analytics can be run on them, and how to build new business strategies around them.

Then, Turner has some groups in the expansion stage of the Data-Driven Maturity Model. For example, business users from its revenue operations use data from the TDC to forecast ad inventory. Turner also uses data to determine which audiences to target—because more and more ads are now being sold using knowledge about audience preferences rather than demographics. "By doing this, we get a much higher CPM [cost per thousand impressions] than if the ads we sold were just based on demographics," says Maranthe.

This same data is being used by brands to figure out what content is best suitable for the audiences that Turner already has. The data is also used by Turner's research team to create reports and forecasts of the business for upper management. "Multiple groups are leveraging the same data, just slicing and dicing it in different ways to suit their requirements," says Marathe.

Finally, Turner does have a few teams in the inversion stage of the Data-Driven Maturity Model. "This is where we brought in extra datasets and started building more centers of excellence around data, or data COEs [Centers of Excellence]," says Marathe. Until recently, the users of Turner's data were not operating like data COEs. But today it has data COEs within its various brands.

The Evolution of the Turner Data Team

Four teams are involved in managing the TDC: Marathe's IT group (called the data engineering group), a data governance group, a business engagement group, and a data scientist group.

Before Turner created the TDC, Marathe's group was called Audience Insights Reporting, and was quite limited in what it did. "Whereas the mission of the Turner Data Cloud is to get the best possible alignment of consumers and content, our earlier mission

was simply reporting certain basic statistics," he says. "The new mission is reporting plus analytics plus everything in between."

Marathe has dedicated two of his 30 engineers to R&D and innovation. "I need this because the data field is changing so fast," he says. "Every day there is some new tool, whether it's a new AWS [Amazon Web Services] product, or enhancements to our Qubole platform or a Spark innovation. Something is always changing. And we want to make sure that we keep abreast of it all."

Marathe says he considers Turner's architecture well structured to benefit from all the innovations coming from the cloud providers, the open source community, and from their Hadoop/Spark platform —Qubole. "As new tools come up or as new data processing platforms become available, we have the ability to shift to them fairly quickly. Our users will always have access to the latest and the greatest. That's the cornerstone of the architecture and that's what cloud affords us, which we couldn't have done if we were based on an on-premises architecture."

Within his IT group, Marathe has a development and quality assurance team responsible for developing processes to bring in new data sources, by first extracting data through various means such as APIs and data feeds, and then transforming and loading that data into a data lake or, for some sources, into a data warehouse. In addition, Marathe has a small development/maintenance and operations team —a team of data architects who are also data analysts, and who analyze new data sources or use cases and figure out what kinds of transformations and models would serve those use cases.

Though the team's goal is always to do as little data transformation as possible, Marathe says, "we try to keep data in a data lake format, and do schema-on-read. So, we try not to do extensive ETL [extract, transform, and load] activities."

Moving Toward User Self-Service

When it comes to putting self-service in place for data users, Turner is currently "all over the map." According to Marathe, "For some business groups, we are already doing that. We have given users a business intelligence platform, and they are quite self-sufficient." But then there are many users who are not independent—yet. Turner's

goal is to democratize data, and to make most of its users self-sufficient in using it.

However, says Marathe, there will always be a data science team within the TDC to perform tasks for smaller business groups within Turner that will never hire data scientists of their own.

"We'll always have a hybrid model for that reason. Self-service will mostly be the domain of the larger business units," he says.

Challenges and Next Steps

The biggest challenge Turner has faced thus far is *data wrangling*. Just the day-to-day chore of bringing in quality data has turned out to be extraordinarily difficult. "We bring in and ingest so much data and so many social media files that it's an ongoing challenge," says Marathe. "But it's smaller today compared to the previous model, where we had to do a lot of ETL and a slight variation in data would throw it off."

Marathe doesn't consider the big data infrastructure challenges facing the media industry to be unique. "Most of the challenges we face are common to all industries," he says. "In a way, we have less concerns when compared to industries like finance and banking and healthcare where they own, store, and process a lot of PII [personally identifiable information] data. We don't have that much PII. But we still have to make sure that we are abiding by all the different laws and regulations—especially now that Europe is implementing GDPR [General Data Protection Regulation]." Complying with everything that is in the European GDPR is going to keep Turner busy for next few months, Marathe admits. However, that's something all businesses in all industries need to do, and it will ensure that no industry misuses people's personal information. Marathe believes this measure will be a good thing in the long term for the data technologies.

One challenge that *is* unique to the media industry is that the data it gets can have many quality issues compared to other industries. Take financial services—core financial industry data has to be very clean. When you go to an ATM and withdraw money, the data says that the $100 you withdrew is a $100 debit transaction from your account. "But when you visit a broadcaster's web page, not all the data is always collected. This is because the web page needs to be

rendered quickly and time is important. "We can't spend a lot of time collecting and sending data, so if there are any hiccups we may not get all the data fields that we want to collect" Marathe says. "So, data quality is perhaps one aspect of data that is particularly challenging to media companies."

Marathe says Turner completely believes in the DataOps way of working. "In the old world, where we did a lot of ETL, we were always challenged because by the time we finished doing ETL, our business users would have new questions or they would have moved on. Putting data and ops teams together has increased velocity," he says. "It's less of a problem now because we are constantly working with them, collaborating with them, and doing as little transformation to the data as possible. We are now always in sync."

Marathe stays in close touch with users. He schedules regular meetings with Turner's business units, which include data scientists as well as users. And, as previously mentioned, Turner also has a group within the TDC devoted exclusively to what it calls "business engagement." "For example, within that group, we have a person dedicated to CNN. A person is dedicated to Turner Sports, and so on," says Marathe. "All this ensures that we are very close to our users."

With his team, which is more on the technical side, Marathe will typically have biweekly meetings with users—mainly the data scientists from the various business units, to see what challenges they are currently facing. "They often have questions about the data because it is so large," says Marathe, who says Turner has datasets consisting of thousands of columns, and it's not easy to understand them. "We try to help them with that," he says.

As Turner's Data Cloud team grows and matures, Marathe expects even more engagement from the business users' side. "After all, we don't really know everything that this data can be used for. That is why the self-service model is important because we just don't know what users' issues are and what problems they're trying to solve," he says. "Users will always know best."

Lessons Learned

The main lesson Turner has absorbed from its data journey thus far is: *build an infrastructure that is as flexible as possible*. This means

going to the cloud. Marathe also recommends keeping compute and storage completely separate, to improve your ability to manage and utilize data. "We are no longer pushing data into a traditional RDBMS [relational database management system] database and thus only able to access it from the ecosystem that supports that RDBMS," he says. "Our data sits in an object storage such as S3 [Amazon's Simple Storage Service], technologically speaking, which gives us the ability to move quickly to newer data access products and does not keep us tied to, for example, one DBMS."

Marathe is a cautiously enthusiastic advocate of open source technology. "Even though we are using open source technology, we are not using it directly. Our cloud platform vendor—Qubole—is baking open source technology into its unified product, and then exposing it to us," he says. The issue with doing things that way, he points out, is that you might not be able to move as quickly on what is out there in the open source world because you need to wait for your vendor to release its product. But the advantage is that Turner is not in the Wild Wild West. "We don't have to figure everything out. The open source technology gets baked in and it works together as a product. And, we don't want to necessarily move that fast anyway for security reasons," says Marathe. Turner is therefore a "conservative" user of open source technology.

Looking back on his data journey thus far, Marathe says Turner is progressing "bit by bit." "We've gone beyond simple reporting to analytics, and we do a certain amount of machine learning and artificial intelligence, but we're not doing predictive or prescriptive analytics as much as I would have liked" Marathe anticipates doing so in the next few years.

Turner will be able to tell, for example, based on the content a customer has consumed in the past, what that customer will be interested in consuming in the future. "We already do that, of course, because we know your areas of interest. But that is still very preliminary. It's not really deep learning at this point," he says. After all, content has many different aspects. A particular viewer could be attracted by the theme of the movie or by the actor playing the main hero or heroine of the plot. So, many different variables exist. Figuring out what is relevant and what is not is where deep learning will come into play, Marathe says.

Turner's use of data will continue to evolve. "We definitely need to keep on top of it. If we don't do it, our competitors will," says Marathe. "So, there are no other options other than being a data-driven company. If we don't do this, we perish."

CHAPTER 5

Comcast: How a Focus on Customer Experience Led to a Focus on Data Science

Comcast is on a mission to improve and enhance the customer experience. In search of what some at the company have called a "superior end-to-end customer experience," Comcast recently made its customer experience team a part of its Technology and Product (T&P) group, aligning its next-generation technology initiatives—which include a focus on data science—with a customer-centric strategy.

Since the day Barbara Eckman, principal data architect in the T&P group, started at Comcast, she's been working to move Comcast forward on its data-driven journey.

The T&P team uses data analytics to improve products and deliver better, faster, more reliable experiences to customers. To do that, the company collects large amounts of nonpersonal telemetry data about network utilization, latency, and throughput, as well as any technology issues that might be affecting performance.

And T&P is just one of a number of groups within Comcast using data science to support enhanced customer experiences.

"Our goal is to align more closely with customers," says Eckman. She is part of the team that is providing a single integrated big data platform within T&P that streams data in near real time, through transformations and enrichments, eventually storing it in a data lake.

From there it can be filtered, aggregated, and stored in auxiliary, special-purpose datastores like columnar databases. The idea is that eventually this big data platform will serve various applications. "At this point, we ingest approximately 200 data sources (Kafka topics), with more added every week," says Eckman. "These are available for relevant groups to leverage in their own product development and customer experience initiatives. We have just stood up a governed data lake in AWS [Amazon Web Services], and topics will be landed there incrementally, in the priority order given to us by the product team."

Eckman notes that when she talks about "customer experience data," it's important to point out that Comcast gathers and processes only that sort of data in the aggregate. "We don't look at individual customers' data. That's private. We're very strict about that," she says.

The T&P group, led by President Tony Werner, is charged with designing, developing, and supporting the products and services that are needed to deliver internet, TV, and voice to Comcast's more than 28 million customers. T&P also supports Comcast employees by building and operating mission-critical systems, including those to monitor its network—one of the world's largest—and its associated cloud infrastructure.

"We like to think of ourselves as the cool kids because we work with the exciting new technologies," says Eckman. In fact, one of her colleagues was on the team that worked on Comcast's X1 Entertainment Operating System, which just won an Emmy for *Outstanding Achievement in Interactive Media Program*. "So, he has a picture of himself holding the Emmy in his office. That's fun," she says. Comcast also has people who monitor network traffic and ensure that the network runs reliably and performs as needed.

Eckman says she was astounded by the amount of network telemetry data Comcast collects. "The data I'm working with here at Comcast is the biggest big data I've ever seen in my career," she says. "It's enormous. So, we're building the central platform that will consolidate data and make it available to all big data analytics activities. And, again, all of this is focused on transforming the customer experience."

Why a Single Platform?

The T&P organization's initiative to build a single, integrated platform came about because almost every department within T&P that routinely did analytics was asking for its own Kafka (ingest) cluster and its own data lake.

"That's crazy on the face of it—and also needlessly expensive," says Eckman. T&P's response: to design a single platform for the data needs that all departments have in common, while recognizing that the kinds of data analytics people will want to do within those departments could vary considerably.

Thus, this is not a one-size-fits-all solution. Eckman's team provides the data in a format in which users from different departments can easily get at it, and easily understand what it means. They can then do whatever custom analytics they need for their particular part of the business.

"In other words, the real power of what we're doing comes from integrating data, and the platform it resides on, across different aspects of the business," says Eckman.

How Data Is Used to Solve Business Challenges

One of the business challenges that data helps Comcast solve is the uncertainty about which network topology to use in a particular geographic area. What would be an ideal network topology when it rolls out a new product that uses the network differently? What would serve customers best?

"For example, we've been moving away from the QAM [quadrature amplitude modulation] standard toward a new digital standard that will transport all content over IP [internet protocol]. That's better for us and for our customers," says Eckman. "But not everyone has made that conversion yet." Consequently, Comcast is analyzing the data to determine the right approach. If a certain region, say Boston, converts, how should Comcast change the network topology in that geographic area to meet customers' needs better?

"We simulate what that looks like using data," says Eckman. "If 20 percent of the people are using X1 now, and then, 80 percent use it in six months, what will the traffic look like? How can different networks' topologies optimize that?"

Comcast also uses data to study situations when error codes are transmitted from a customer's set-top box. What should be said and done in its customer-care discussions? What should Comcast recommend that customers *do*? If possible, can the problem be fixed remotely? "Can we just send a signal to the box, instead of asking customers to do something?" asks Eckman. "Or even better, can we use data to try and figure out what's wrong before the customer even knows it's a problem? There's that possibility."

As another example, Eckman works with the team that supports the technology for Comcast's digital video recorder (DVR) product. The video team uses data to analyze how well it's working. The goal is to understand immediately if a problem occurs on a set-top box. Rather than just giving the customer a blue screen, Comcast would remotely push the "off" button on the set-top box for 20 seconds and release. This would happen automatically, of course, with a message saying, "We're doing this because…"

"In other words, we're putting in self-healing devices by automatically doing things that the people on the first line of customer care would tell you to do if you called," says Eckman. "In that way, we're dramatically streamlining support." Sometimes that's all it needs. There's no need to disrupt the customer. "In T&P we're doing this in a number of areas," she says. "For many things having to do with customer satisfaction, we gather the data, analyze it, and make sure the customer is having the best possible experience."

Comcast also relies on data to find ways to improve latency when content is being downloaded. All networks experience issues from time to time, whether from physical causes like downed or cut lines, or configuration issues. But data coming from sensors in Comcast's network help it detect what's gone wrong. "The idea's always, in these efforts, to save the customer from either having to call or wait for a technician to come in a truck to fix something," says Eckman. "We certainly do that when necessary, but it's much preferable to not have to. Nobody wants to sit around waiting for the cable guy to show up."

This is good for Comcast, too, because it helps it send technicians only when they're actually needed.

Why Governance Is Essential

Data analysts have traditionally spent 70 percent of their time data wrangling. Just finding out where data is and what it was took up most of their time. Only 30 percent of their time was spent doing actual analysis.

Comcast's T&P organization wanted to change this. It decided that from the very beginning it would put all data under strict governance. "From the minute it enters our system, we have a fully documented schema for it," says Eckman.

Eckman says that the T&P integrated big data platform that her team is building is a relatively small initiative to prove that DataOps works, creating the justification for the next, bigger step. This means that it had the luxury of starting from scratch, and not having to clean up after aging legacy databases.

But before the team established the current data governance protocol, Comcast had no easy way to perform data discovery.

"That approach to data is crazy," says Eckman. "It's the old adage, GIGO [garbage in, garbage out]. If you throw things in and you have no idea about the quality of that data, you have to chase around to find out who owns it. You have to chase around to make sure that if they say "deviceID," it means the same as the deviceID that you get from some other project, because it doesn't always align. You can't merge data if your IDs don't use the same semantics. You can't join data if that's the case. You have a mess on your hands."

Comcast uses Apache Avro schemas for governance. Avro is an open source data serialization system developed under the auspices of the Apache Software Foundation. It is well structured to begin with, but Comcast put conventions on it that make it even more useful.

For instance, every attribute must have a documentation string associated with it, and that documentation must be informative. For example, no acronyms. The owner of the data must spell out everything. For example, if an attribute is called "Event Type," the documentation can't say "the type of the event." It must be more useful. "You have to be thinking of who will be seeking the attribute beyond just you and your team," says Eckman. "You have to think of others.

We—yes, humans—manually review all the schemas and frequently ask, 'Can you please explain this a little better?'"

Comcast performs automated validation, but Comcast does human reviews, as well, just to make sure that the data is as clean and understandable as it can be, using—where possible—common language. Eckman and her team have developed "core schemas" that they encourage people to use so that everybody is using the same schema for the same thing.

"If you're talking about geolocation, you want to reference latitude and longitude, state and city, and postal code. Those are all standard across Comcast today," says Eckman. "There's no reason to have your own field called Lat when we provide a field called Latitude that's for the same thing."

Additionally, Comcast is careful to track data transformations and lineage. "If data comes in and someone transforms it or enriches it with some extra data that comes from somewhere else, and then stores it in the data lake, we have to capture the new schema, document it, and link it to the old schema and the enrichment process," says Eckman.

Comcast uses Apache Atlas to keep track of data lineage. And, cautions Eckman, it's important to track it from two directions: both coming and going. "The people who are publishing their data to Kakfa need to know what happens to it in its journey through the system," she says. "Perhaps even more importantly, the people reading this data in the data lake need to know where it came from, who produced it, and how it was transformed along the way."

Comcast, of course, can easily do this for data that's coming into the system for the first time. For data that's already in the system, it must infer its lineage, which is a little more difficult. "But that's the state that most organizations, I think, find themselves in," says Eckman. "There are tools to do that, including some light machine learning solutions that we are going to use in the near future, when we begin to integrate our legacy, on-pre-Hadoop data."

But Comcast's master plan is to nip data chaos in the bud before it becomes a problem. "After all, you don't want your analysts wasting their time. You want them to be doing analyses, not data wrangling," says Eckman.

Team Interactions at Comcast T&P

The Comcast Integrated Data Platform team's purpose is to provide a foundation for many analytics groups, including its own in-house analytics team as well as a wide variety of user-facing products and solutions (Figure 5-1).

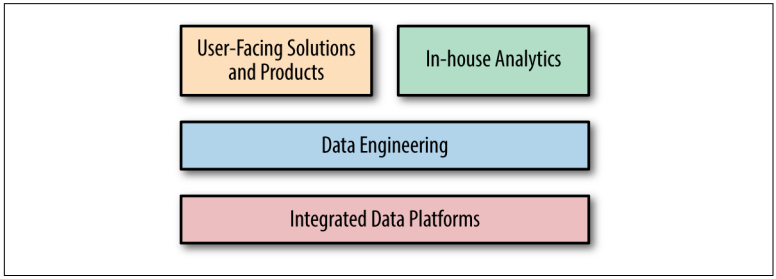

Figure 5-1. Team interactions at T&P (Source: Comcast)

The Integrated Data Platforms all work together to govern, ingest, transform, and store data of the highest quality possible (Figure 5-2).

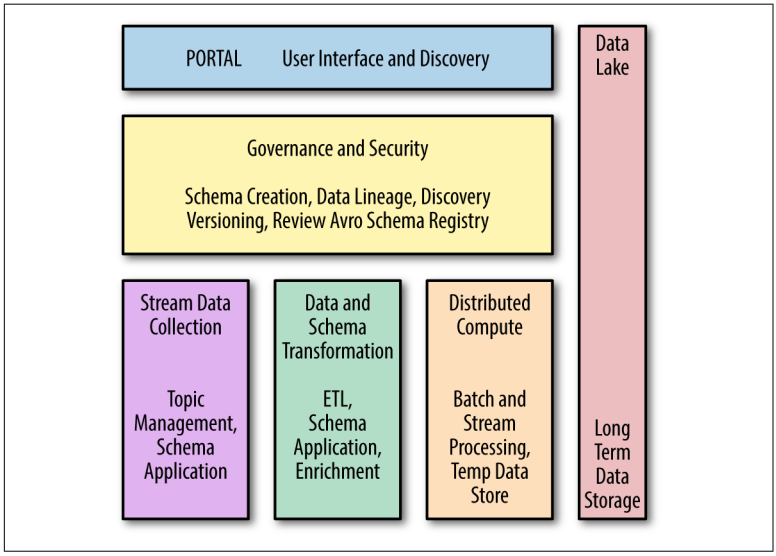

Figure 5-2. Detail of Integrated Data Platforms

The data enters the system through the stream data collection platform. The data and schema transformation platform enriches the

data and acts as the connector between the streaming data and the data lake.

Comcast also has a team responsible for the data lake platform, and then one that manages its distributed compute platform, both of which spring into action when users need clusters to do things such as hive queries. "That's all self-service," says Eckman. "Users tell us 'I want to do this, and I want these tools.' They give us a sense of the size of their projects, and then we spin it up." Everything of course is done in the cloud, Eckman says.

There's a portal that will sit on top of everything. "We're only at the beginning stages of that," says Eckman. "The idea is that the portal will be the first point of contact where people find out what the various platforms offer, and where they can put data into the system as well as take it out."

In the middle of everything is the data governance piece, which is what Eckman leads. "We make sure that at every point, we know what the data looks like. If it's enriched, it has to have another schema. We trace the lineage precisely."

Comcast also has in-house analytics resources. Data engineers look at the data in the data lake, and aggregate the network traffic data. They can aggregate it by 15-minute intervals or half-hour intervals, or aggregate it by region. They can slice and dice it any way they want.

"Everyone has his or her own role to play," says Eckman. "My team makes sure the data works. The data engineering team makes sure that what they built on top of it works. Then there are the outside analysts who use the various platforms to do their own analyses. That's how it's structured."

All of this is being done in the cloud. Comcast uses Amazon Web Services (AWS). "One of the wonderful things about Apache Atlas is that it's extensible, so we extended it to include new data types that are specific to AWS," says Eckman. "We're maintaining Hive tables and saving lineages in a highly heterogeneous environment. We use AWS object stores rather than HDFS [Hadoop data file system], and AWS lambda functions to capture end-to-end lineage, not limited to the data lake. We think this is pretty darn cool."

Comcast's vision is to offer a single platform. "I don't know if we'll ever get to be the platform for Comcast as a whole," says Eckman.

"We can hope for that. The path would be to build our system as solidly and reliably as possible, using best practices and governance. Then, we have to evangelize it, so people know it's there, and that they could be using it."

The governance part tends to be a little more challenging because nobody wants to be governed, Eckman says. But after she talks to people, they generally see the point. "Doing what we ask means a little more trouble for them upfront, but they see the value in it."

"Our objective: we are so good at what we do that all departments will ask us to take care of their data—so they don't have to. That's what I mean by evangelizing. You've got to make them want to join in," she says.

DataOps as a Way of Work

Getting to a DataOps culture is important. Collaboration is key to it, says Eckman.

Within T&P, members of the data team tend to wear a lot of hats. "One of my favorite aspects of working with data, is that it's like being on a softball team," says Eckman. "I'm like the second baseman, an important part of a bigger whole. But being part of this, and on a scale I've never seen before, is really nice."

For her especially, as the governance person, she touches everything. She has to work well with everyone. "We're building a massive data pipeline, and I have to do my job well so it can be handed off to the person who comes next," she says.

Would Comcast be able to stay competitive without what the data team is doing? Not for long, Eckman says. "Everyone else is doing this now, too."

Because the T&P team started its big data initiative from scratch, the platform is still in the early stages—currently version 1.5. "It is a viable solution at this point," says Eckman. "It works. We may add more platforms, but the idea is to keep hardening and optimizing and adding features. Like with any software project, you're never done."

As far as where Comcast is on the Data-Driven Maturity Model scale, Eckman thinks Comcast is in the "expansion" stage because everything is so new. This initiative is just two years old, after all.

Some media companies have been on their data journeys for much longer. "But they might not be doing it optimally," says Eckman. "I talked to somebody from another large media company, and he said, 'Gee, I wish that I could do the governance you're doing. Our culture just won't stand for it.'"

So, in some ways, Comcast is ahead of others. But one thing is certain: Comcast has no plans to go back to the days of having data silos and letting people do everything by themselves. "This is the future of data science," says Eckman. "We can't look back."

CHAPTER 6

The Changing Data Landscape for Media, and Next Steps Toward Becoming Data Driven

The market has dramatically changed for media companies in recent years. Everything now depends on data.

Media companies are finding they must discard all of their previous assumptions as they enter this new information-centric age. Their customers today consume media on a plethora of disparate devices. They access it from an array of different channels. And personalization is now critical to winning their hearts and minds. For all these reasons and more, media companies must become data-driven or they won't survive.

By now, you've read about overall industry trends as well as the data-driven journeys of SlingTV, Turner Broadcasting, and Comcast. In this chapter, we go over the most significant industry changes that are driving these businesses' new data-driven reality as well as provide a five-point checklist to help other media companies on their data-driven journeys.

Three Industry-Wide Changes Compelling Media Companies to Become Data Driven

Here are the three biggest changes that are transforming the media industry today:

- Changes in the ways content is delivered
- Proliferation of devices and connectivity leading to a data explosion
- Availability of highly scalable and cost-effective tools to manage it all

The Changing Pace and Face of Content Distribution

First, we are moving from linear (as in broadcast TV programming) to *over-the-top delivery* (OTT) models created by the Netflixes of the world.

OTT, which we introduced in Chapter 3 in our discussion of Sling TV, is a content distribution practice in which a content provider sells or rents audio, video, and other media services directly to the consumer over the internet via streaming media. These services are generally sold as standalone products, bypassing telecommunications, cable, or broadcast television service providers that have traditionally distributed such content.

With OTT content providers like Sling TV, Hulu, HBO Now, YouTube, and, of course, Netflix and Amazon, consumers don't need to watch content in a linear way. They can chose when—and how—they want to consume it. The global OTT market is predicted to grow at a more than 17% compound annual growth rate (CAGR) between 2017 and 2025, to reach approximately $3.49 billion by 2025.

And the sheer volume of content choices is accelerating. According to YouTube statistics, more than 300 hours of video are uploaded to YouTube every minute, and nearly five billion videos are watched on YouTube every day by more than 30 million unique visitors. This represents a whopping one billion hours' worth of YouTube video watched each day.

Netflix alone intends to produce 1,000 hours of new creative content in 2018—and will spend $6 billion doing so. A full 70% of Netflix customers confess to "binge watching" shows on a wide variety of platforms: Smart TVs, laptops, smartphones, and tablets.

And it's not just the volume of content, but the variety. People are looking to consume TV shows, movies, music, independently produced videos, blogs, news, articles, even books as part of these OTT services.

A final point that media companies must consider when planning their distribution strategies is that their customers' attention spans are becoming shorter all the time. Smaller attention spans mean that optimizing and targeting content that is delivered to individuals becomes even more important. After all, things move very fast in today's media world. A recent Microsoft consumer study found that the human attention span today is eight seconds, down from 12 seconds in 2000. Just to put this in perspective, a goldfish has an attention span of nine seconds.

Proliferation of Devices and Connectivity

The second change driving media companies to become more data driven is the availability of infrastructure to collect more and more data about consumers. This has led to an explosion in the variety as well as the velocity of the consumer data collected today.

The two fundamental drivers here are ubiquitous connectivity along with the proliferation of powerful end devices. The later are usually devices on which the content is consumed by the consumers. In the past decade and a half, we have seen not only the rise of smartphones but also smart TVs, home entertainment systems, voice recognition and home automation devices, and many others. These devices come with high amounts of processing power and rich software capabilities to collect data. As a comparison, a smart phone today packs more processing power than the computers used by NASA for the moon missions in the 1960s. As a result, it has become progressively easier to generate data about user behavior, viewing habits, and user interactions with media. At the same time, the communications infrastructure has also progressed to provide high-speed, ubiquitous connectivity to most of the world population. According to GSMA Intelligence, five billion people world over now have mobile phone connectivity. What this means is that not only can a lot of data be generated, but also that data can be connected and centralized at one place for media companies to analyze for many different purposes (see Figure 6-1).

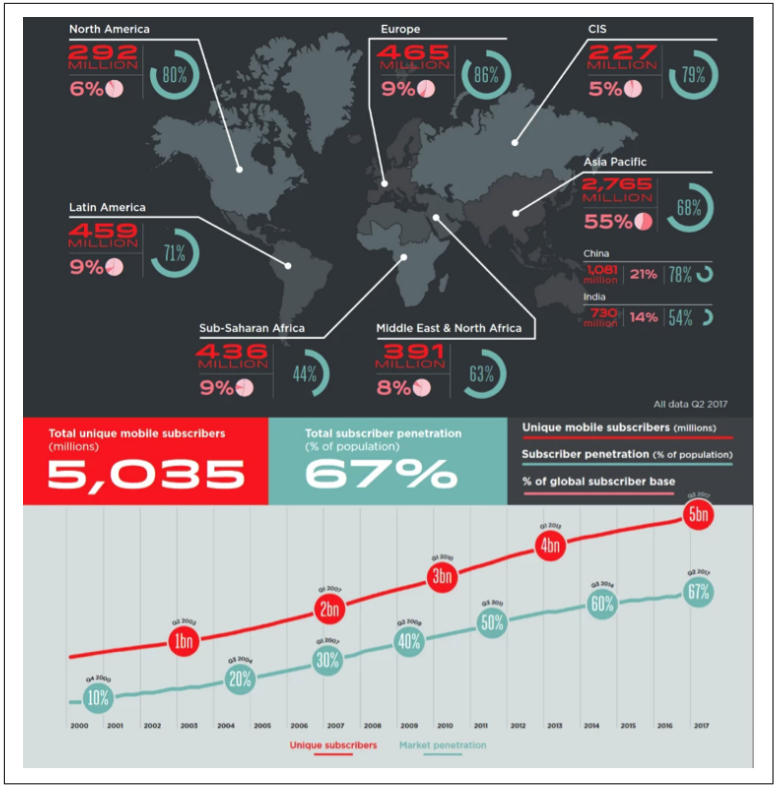

Figure 6-1. According to GSMA Intelligence, five billion people world over now have mobile phone connectivity (Source: © GSMA Intelligence 2018)

In addition, social media has also played a role in the way data has changed. There's a lot more data generated by interactions between people. Although this is unstructured data, and therefore more difficult to process than semi-structured or structured data, capturing and analyzing it can help media companies understand their users at an individual level.

Moreover, the number of data sources has also increased. Many third-party services now exist that can provide deeper-level demographics and psychographics data that can be joined with a content provider's own information about its customers to come up with a detailed—and fairly accurate—understanding of individual customers' likes and dislikes. Companies such as Acxiom, Oracle Marketing

Cloud, and Epsilon have built billion-dollar businesses based on providing this data to marketers.

Availability of Scalable and Cost-Effective Tools to Manage Big Data

Finally, there have been breathtaking changes in technology used for collecting, storing, and analyzing data that makes becoming data driven much easier.

Cloud computing has become more and more mainstream. Cloud technologies over the past decade have shown a tremendous advancement in scalability and data security and at the same time have continued to drive the costs down. According to Tariff Consulting's Pricing the Cloud 2—2016 to 2020 report, the average entry-level cloud computing instance now costs approximately $0.12 per hour. That's a 66% drop since 2013. On the storage side, the decreases in the costs have been even more dramatic.

Along with the cloud, we've also seen the emergence of big data technologies that scale linearly and provide a rich set of analytical frameworks for different types of analysis on the growing volume of data. Big data technologies are built as distributed systems that work on commodity hardware and can be conveniently and incrementally scaled to the increasing needs of data. As a result, the scale of data processing that was once available to companies like Google is now available in a very cost-effective way to every enterprise. The open source Apache Software Foundation has an ever-growing list of projects that have brought these categories to the mainstream—chief among them Apache Hadoop, Apache Spark, Apache Hive, and many others.

Many of these analytical frameworks have made rich types of analytics available to an ever-increasing number of enterprises—not just SQL but frameworks that harness data science, deep learning, and other emerging cognitive technologies. Previously, people used to work primarily on SQL, and it was widely considered the only analytical framework that companies could cost-effectively utilize.

Now graph-processing, data-science, and other analytical frameworks are widely available through open source projects. In general, open source has played a tremendous role in democratizing big data—moving it from the purview of the Googles and Facebooks of the

world to being available to any enterprise that wants to become data driven.

The combination of cloud and big data technologies has even simplified the operations of these technologies to a degree that many are now available as a service in the cloud. The agility, flexibility, and ease of adoption that this combination has led to has tremendously simplified how all industries in general and media industry in particular can adopt big data and put their rich data assets to use.

Adopting an Agile, Data-First Mentality

Because of these three major industry trends, it's critical for media businesses to adopt data-first rather than media-first mindsets.

We used to call it "mass media" in the last century because it reached so many people—that is, the masses. But today, broadcasting content to the masses doesn't work. The audience is too fractured. They're tuning in at different times, on different devices, through different channels. Now that media companies have the ability to identify individual users and their preferences, they can be much more effective at targeting the right content to the right person at the right time through the right channel.

With the new measurement tools that are available, media firms are embarking on programs of intensive data collection, deep analyses, and experimentation to capture what used to be called "eyeballs," in earlier days of the media industry. Today, many of the internet services companies routinely run thousands of A/B tests to understand the particular tastes of individual customers. Given the latest technologies, they can move away from having purely "creative" but limited ideas to implement, to scientifically testing thousands of ideas simultaneously. Rather than depending on individual employees to come up with what would be considered a good theme for a campaign, piece of content, or ad, today's media companies have teams that generate thousands of ideas—some good, some perhaps not so good. But they can all be tested out quickly because of the new tools.

Media companies also need to focus on technologies that can allow them to experiment quickly enough with data to become extremely agile. In other words, you don't want to adopt technology for which

an experiment takes months to set up. You want to be able to do this kind of experiment in minutes.

For example, Netflix does continual A/B testing so that it can constantly improve its quality of experience (QoE) for viewers.

The goal of Netflix's experiments is to find out if a new algorithm, change to an existing algorithm, or a configuration parameter change improves QoE metrics. For example, it measures metrics related to video quality, rebuffers, play delay (time between initiating playback and playback start), and playback errors, among other ones.

An A/B system test might last a few hours or might take a few days, depending on the type of change being made and to account for daily or weekly patterns in usage and traffic. Because Netflix has such a large member base—consisting of more than 100 million members, they can collect millions of "samples" very quickly, enabling it to iterate and perform multiple system experiments sequentially. For example, Netflix can find out such things as whether members would watch more Netflix if they have better video quality or lower rebuffers or faster playback start, or whether they remain customers after the free trial month ends and in subsequent months.

Five Steps to Becoming Data Driven

For media companies who have yet to reach the "nirvana" of the Data-Driven Maturity Model (see "The Data-Driven Maturity Model" on page 15), here are five steps to help you along your way.

Hire Data Visionaries

You need people who see the "big picture" and understand all the ways that employees can use data to improve their businesses. Although this certainly includes analyzing marketing, sales, and customer data, it doesn't end there. Data-driven decisions can help with internal operations, such as making customer service and support more efficient, and cutting costs from inventory, for example. And it all begins by hiring people who are open minded about what the data will tell them regarding the way forward—people who have a vision.

Consolidate Data into Cloud Data Lakes with Enterprise-Wide Access

All of the data in the universe won't help if that data is inaccessible to the people who need it to make business decisions. A data-driven company consolidates its data while keeping it continuously up to date so that employees have access to the most accurate information at any given point in time. This means eliminating data silos and effectively democratizing data access. There are, of course, always data security and compliance issues. But making data available to everyone within the framework of the company security and compliance policies is an important feature of a self-service data culture. Always allow employees to see the data that affects their work. They need to see this not only at a granular level, but also in a holistic way that helps them to understand the bigger picture. Doing this will make your employees more informed, skilled, and enthusiastic about using data to improve the business.

A common best practice in this area is to embrace data lake architectures and technologies. These technologies help consolidate all types of data—structured, unstructured, and semi-structured—in one place so that these datasets can be easily correlated. In addition, using the cloud to create this data lake is imperative to embrace an Agile infrastructure, which would be aligned with the need for fast iteration and constant experimentation that media companies need to continuously refine their understanding of their audiences. In addition, operationalizing data lakes in the cloud is much easier and cost effective because of the automation capabilities available. With automation and the use of platforms like Qubole, continuous operations of a data lake are simplified and are achieved at a fraction of the cost.

Empower All Employees

All employees should feel comfortable taking initiative when it comes to suggesting ways that data can be used. This kind of mentality goes well beyond just using data, of course. If you build a company where all employees feel free to give opinions—as long as they are backed up by data—even if those opinions contradict senior executives' assumptions, you are building an organization where the best ideas will naturally gravitate to the top and keep you competitive in even the fastest-moving markets.

Invest in the Right Self-Service Data Tools for Each Type of User

Your data, even if readily accessible, won't help your business much if most of your employees can't understand it or don't have the right tools and analytical frameworks that they can use and understand. The answer to this is not to look for one tool for all personas, but to invest in the right data tools for the different data users within the organization. As an example, although tools such as Tableau might be right for the nontechnical business users, the new generation notebooks interfaces such as Zeppelin and Jupyter might be the right integrated development environments (IDEs) for data scientists.

It is equally important that common definitions and terms describe the data set and are standardized within the organization and that these definitions are published and available. In addition, the chosen tools should not just make access easy and intuitive for the data users, but should also make it easy for them to share the analysis and collaborate with their colleagues.

In addition to the plethora of new tools and new types of analytical frameworks emerging, make sure to invest in training for these tools and analytical frameworks. Having an "intuitive interface" isn't enough. Do your users understand basic principles of data analysis, transformation, statistics, and visualization? To achieve return on investment on your tools, they must understand exactly what capabilities each tool offers. Training can be live, video-based, or online, and should use a shared data store so that employees can compare their data discoveries and explorations with one another.

Hold Employees Accountable

Technology will take you only so far. You also need to put incentives in place to encourage employees to use the technology and tools. You also should have a way to measure and grade progress toward a self-service data culture. This means holding employees accountable for their actions and give them recognition when they effectively use data to drive business decisions. Only when you recognize employees for actions based on data will you achieve true cultural transformation.

The collaborative, social dimension of a self-service, data-driven culture is also not to be underestimated. Without it, you will fail, and your investments in software, data processing tools, and platforms will be wasted. Although many organizations pay lip service to this notion of collaboration and openness, not all follow through with the appropriate actions. Keep in mind that data doesn't belong to IT, data scientists, or analysts. It belongs to everyone in the business. So, your tools need to allow all employees to create their own analyses and visualizations and share their discoveries with their colleagues.

In Conclusion

Data is the fuel of the digital age, just as power and steel fueled the industrial age. Media is among the first industries that became digitized, and not adopting a data-first approach is akin to not using steel and power to drive your factories during industrialization. Media "factories" have to be run on data.

We've seen the effects of the digital age on media companies. Take print news media, for example. Many newspapers have closed, and readership is down in part because many were not adopting new technologies and not using data effectively (of course, much of this was due to readers' reading habits changing to the massive number of new digital competitors). According to the Pew Research Center, estimated total US daily newspaper circulation (print and digital combined) in 2016 was 35 million for weekday and 38 million for Sunday. This represented an 8% drop over 2015. Newspapers such as the *Guardian*, the *New York Times*, and the *Wall Street Journal* are estimated to be leading the pack in digital readership, but many local dailies or alt weeklies have closed their doors permanently.

The trend is remarkably different in purely digital news media world. In the United States, a full 93 percent of adults get their news online. Digital advertising revenue continues to grow year to year, but instead of newspapers, it's now the data-driven behemoths like Google and Facebook that have a huge influence on how news travels as well as the revenue news generates.

In summary, to be one of the leaders in the media world, data is critical.

About the Authors

Ashish Thusoo and **Joydeep Sen Sarma** were part of building and leading the original Facebook Data Service Team from 2007–2011 during which they authored many prominent data industry tools, including the Apache Hive Project. Their goal was not only to enable massive speed and scale to the data platform, but also to provide better self-service access to the data for business users. With the lessons learned from successes at Facebook, Qubole was launched in 2013 with these very same product principles: speed, scale, and accessibility in analytics. The company is headquartered in Santa Clara, CA, with offices in Bangalore, India.