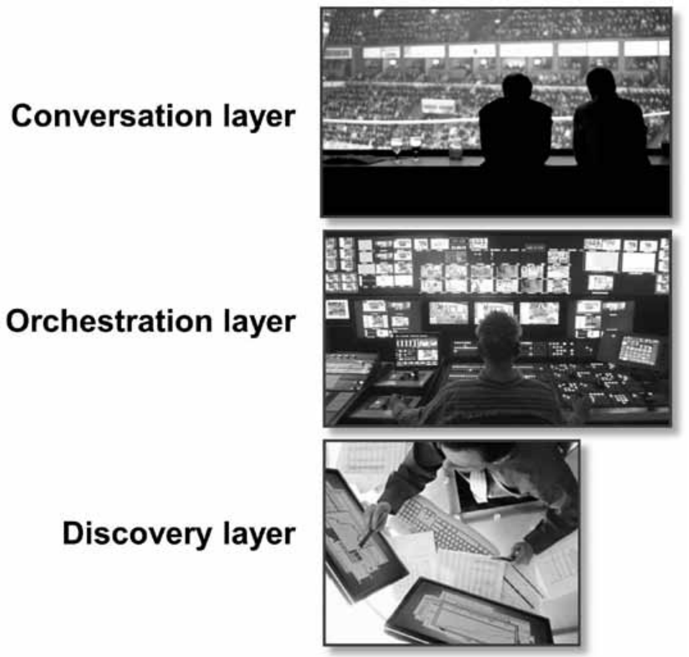

Figure 5.1: Mapping of sports television to Big Data Analytics architecture

Also, how does the orchestration layer guide the conversation layer to make the conversation more relevant and insightful without slowing it down? This section will continue to make use of the sports television analogy and drive it toward characteristics of the architecture and supporting infrastructure of analytics.

The first task at hand is to *identify a situation or an entity*. An analytics system working in conversation mode must use simple selection criteria in the form of counters and filters to rapidly reduce the search space and focus on its identification within a window of time. A television commentator is able to follow the ball, identify the players in the focus area, and differentiate between good and bad performances. Similarly, a real-time campaign system filters all the available data to identify a customer who meets certain campaign-dependent criteria. A customer who has been conducting online searches for smartphones is a good candidate for a smartphone product offer. A number of real-time parameters, such as customer location and recent web searches, would play major roles.

The second task is to *assemble all associated facts*. At this point, the sports television director may offer data previously collected about the players to the commentator, who can combine the data with his or her personal experience to narrate a story. In the Big Data Analytics architecture, the moment a customer walks into a retail store or connects with a call center, the orchestrator uses identifying information to pull all the relevant information about this customer.

The third task is to *score and prioritize alternatives* to establish the focus area. It always fascinates me in U.S. football when half a dozen players wrestle with each other to stop the ball. The commentator has the tough task of watching the ball and the significant players while ignoring the rest. Similarly, in the Big Data Analytics architecture, we may be dealing with hundreds of predictive models. In a relatively very short time (less than one second in most cases), the analysis system must score these models on available data to compare the most important alternative and pass it on for further action. In online advertising, the bidding process may conclude in less than 100 milliseconds. The Demand Side Platform (DSP) must view a number of competing advertisement candidates and select the one that is most likely to be clicked by the customer.

The last task is to *package all the real-time evidence*. The information is turned over to the orchestration layer for storage and future discovery. The conversation layer can now focus on the next task, while the orchestration layer annotates the data and sends it to the discovery layer.

A number of software products are emerging to provide technical capabilities for real-time identification, data synthesis, and scoring commonly referred to as *stream computing*. Stream computing is a new paradigm. In “traditional” processing, one can think of running analytic queries against historical data—for instance, calculating the distance walked last month from a data set of subscribers who transmit GPS location data while walking. With stream computing, one can identify and count, as well as filter and associate, events from a number of unrelated streams to score alternatives against previously specified predictive models. IBM’s InfoSphere Streams has been successfully applied to the conversation layer for low-latency, real-time analytics.

5.2 Orchestration and Synthesis Using Analytics Engines

Nowadays, it is impossible to imagine a live television program without orchestration. A highly productive team has replaced what used to be a “sportscaster” in the early days of sports coverage. A typical television production involves a number of cameras offering a variety of angles to the players, in addition to stock footage, commentators, commercial breaks, and more. The director provides the orchestration, assisted by a team of people who organize the resources and facilitate the live event.

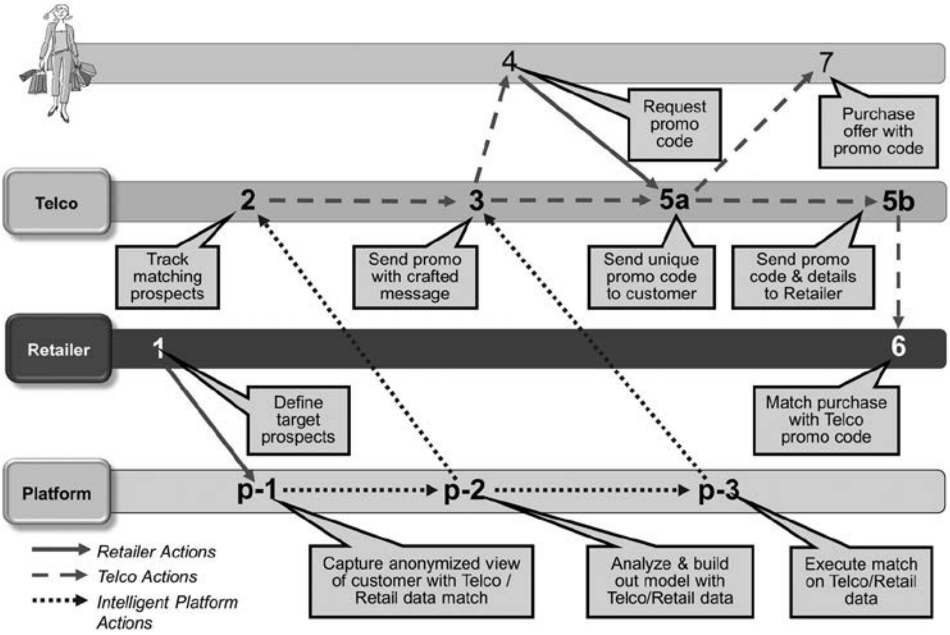
4
7
Request promo code
Purchase offer with promo code
Telco
2
3
5a
5b
Track matching prospects
Send promo with crafted message
Send unique promo code to customer
Send promo code & details to Retailer
Retailer
1
6
Define target prospects
Match purchase with Telco promo code
Platform
p-1
p-2
p-3
Retailer Actions
Telco Actions
Intelligent Platform Actions
Capture anonymized view of customer with Telco / Retail data match
Analyze & build out model with Telco/Retail data
Execute match on Telco/Retail data

Figure 5.2: Telco/Retailer orchestration in Intelligent Advisor

Similarly, Big Data Analytics is increasingly involving vast amount of components and options. The stakes are becoming increasingly higher. For example, as we discussed in Chapter 3, as new products are launched, there is a need for orchestrated campaigns, where product information is disseminated using a variety of media and customer sentiments are carefully monitored and shaped to make the launches successful.

Let us use the Intelligent Advisor scenario that was described in Section 3.6 and depicted in Figure 3.3. We repeat the steps here so we can analyze the process that the Retailer and Telco use to engage the customer:

Step 1: Lisa registers with the Retailer, gives permission to the Retailer and the Telco to use consumer information to track activities.

Step 2: Lisa follows a friend’s post on Facebook and clicks the Like button on a camera she likes.

Step 3: The Intelligent Advisor platform processes Lisa’s activity for relevant actions using Retailer and Telco information.

Step 4: Lisa receives a message with an offer reminding her to stop by if she is in the area.

Step 5: Lisa receives a promotion code for an offer while passing by the store.

Step 6: Lisa uses the promotion code to purchase the offer at a PoS device.

Figure 5.2 shows the behind-the-scenes orchestration steps. As the system interacts with Lisa, it uses an increasing amount of opt-in information to make specific offers to her. By the time we get to Step 5, we are dealing with a specific store location and related customer profile information stored in the Retailer's customer database.

These orchestrations involve a number of architectural components, possibly represented by a number of products, each performing a role. One set of components is busy matching customers to known data and finding out more data. Another set of components is performing deep reflection to find new patterns. A management information system keeps track of the overall Key Performance Indicators (KPIs) and data governance.

Figure 5.3 depicts the orchestration model. The observation space and its interactions are in real-time. IBM's InfoSphere Streams provides capabilities suited for real-time pattern matching and interaction for this space. The deep

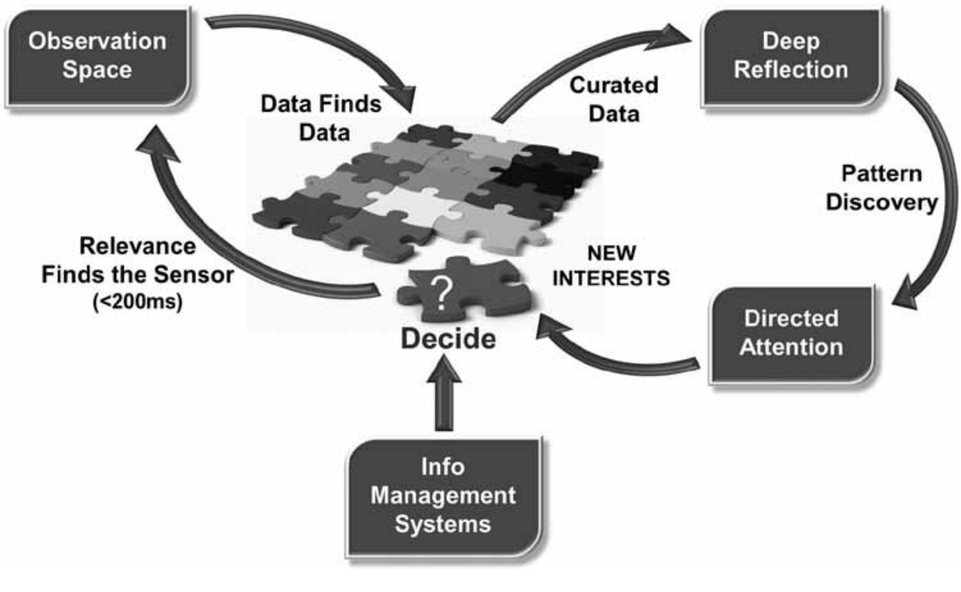

Figure 5.3: Orchestration-driven identity resolution

reflection requires predictive modeling or unstructured data correlation capabilities and can best be performed using SPSS or Big Insights. Directed attention may be provided using a set of conversation tools, such as Unica® or smartphone apps (e.g., Worklight™). Management reporting and dashboard may be provided using Cognos. Depending on the level of sophistication and latency, there are several components for the box in the middle, which decides on the orchestration focus, directs various components, and choreographs their participation for a specific cause, such as getting Lisa to buy something at the store.

Entity Resolution

Using a variety of data sources, the identity of the customer can be resolved by IBM's Entity Analytics®. During the course of the entity resolution, we may use offers and promotion codes to encourage customer participation, both to resolve identity as well as to obtain permission to make offers (as in Steps 4 and 5 above).

Model Management

IBM SPSS provides collaboration and deployment services, which are able to keep track of the performance of a set of models. Depending on the criteria, the models can be applied to different parts of the population and switched, for example, by using the champion/challenger approach.

Command Center

A product manager may set up a monitoring function to monitor progress for a new product launch or promotional campaign. Monitoring may include product sales, competitive activities, and social media feedback. Velocity from Vivisimo, a recent IBM acquisition, is capable of providing a mechanism for federated access to a variety of source data associated with a product or customer. A dashboard provides access to a set of users monitoring the progress. Alternatively, the information can be packaged in an XML message and shipped to other organizations or automated agents.

Analytics Engine

An analytics engine may provide a mechanism for accumulating all the customer profiles, insights, and matches as well as capabilities for analyzing this data using predictive modeling and reporting tools. This analytics engine becomes the central hub for all information flows and hence must be able to deal with high volumes of data. IBM's Netezza product has been successfully used as an analytics engine in Big Data architectures.

5.3 Discovery Using Data at Rest

Statisticians and video editors are the third set of team members in a sportscast. These people work with a lot of historical data and constantly work on the statistics to compare the current game with past ones. They also capture and edit replays by focusing on the game from a start time to an end time and using multiple camera angles to show the action from different viewpoints.

In the world of Big Data Analytics, statistical models perform structured data analytics, while a number of accelerators have been built for unstructured analysis. For example, query tools in Big Insights can be used to focus on data from one point in time to another and correlate a number of sources, much like the way an editor brings together multiple camera moments in sports television.

Business Intelligence has been focused primarily on structured data. MDM provides the ability to match structured customer or product data to bring a single view of customer or product. Data Warehouses provide the ability to store ingested data and build different data representations for the purpose of reporting or predictive modeling.

The presence of unstructured data brings in a set of data scientists who work with a mix of unstructured and structured data to determine insights. These insights are offered to the orchestration layer and could also be obtained at the command of the orchestration layer focused on specific entities (for example, detailed analytics about the hour before a major network outage, focusing on network entities in the vicinity of the outage).

The discovery layer provides much-needed reflection on the historical data. By nature, discovery is slower than the conversation and orchestration layers. However, the results of discovery can be integrated in the analytics engine with more recent data and used to interpret or augment the conversation. Much as a commentator may use statistics to make a point in the middle of a narration, the discovery layer in the Big Data architecture provides much-needed context and behavioral prediction to a sales conversation in order to provide an offer.

Computers have traditionally been assigned to numeric computing. As a result, structured data analytics using statistical models is a mature topic. We are now seeing the emergence of a couple of unstructured data analytics techniques, as I summarized earlier in the book (in Section 4.2). Let us turn our attention now to how we combine inferences drawn from structured and unstructured analysis and how we use them together for an overall ranking.

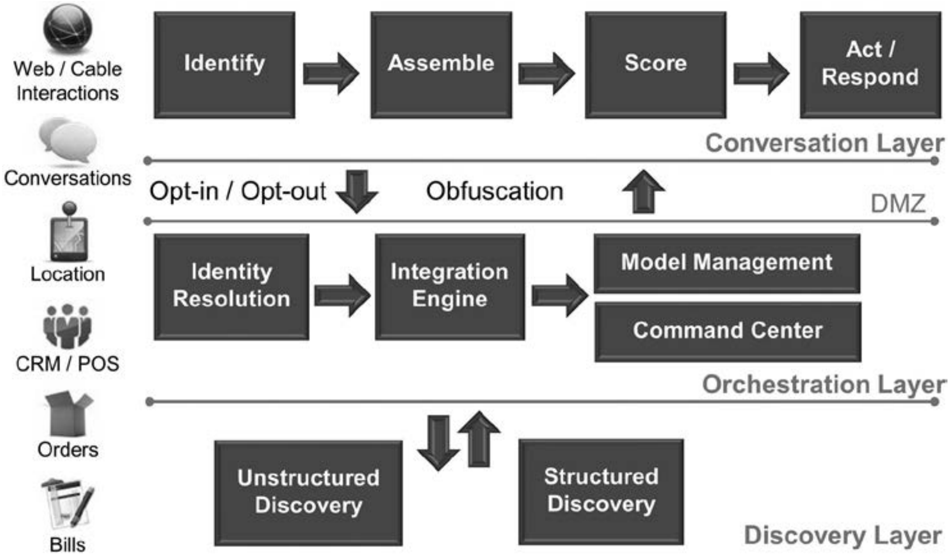

Figure 5.4: Advanced Analytics Platform (AAP)

5.4 Integration Strategies

Once connected together, the integrated Advanced Analytics Platform (AAP) is as shown in Figure 5.4. As we integrate the three layers, we can incorporate a data privacy layer to differentiate between internal and external sources or users. I have placed that layer between the conversation and orchestration layers in this figure to depict conversation on social media sites. However, the privacy layer could go anywhere in this diagram, based on the placement of external sources or users.

The current Business Intelligence environment can be integrated in one of three ways. First, in a typical greenfield Big Data Analytics implementation, we could make the current MDM or warehouse one of the sources and keep the new architecture completely out. Second, the BI environment could provide the structured data discovery component and could be connected to a Big Data Analytics engine. Third, in the most evolutionary approach, the BI environment could provide the orchestration layer and could be augmented with unstructured discovery, identity analytics, and the conversation layer. The first option provides the best architecture to meet Big Data requirements, while the third provides the best leverage of the current investment.

Those of you who may be wondering whether we can actually apply these concepts to sports television will be interested to know that IBM has been working closely with Roland Garros, one of tennis's most celebrated competitions, to build a cloud-based system that automates the three roles described above. In conjunction with the French Tennis Federation, IBM has been designing, developing, producing, and hosting the state-of-the-art *rolandgarros.com* website since 1996.[30] IBM has enhanced Roland Garros by providing an innovative and immersive online experience for millions of tennis fans worldwide. Scores, statistics, and match analysis come to life with IBM SlamTracker™ 2012. Aces, serve speed, winners, and all other key statistics are rendered in real-time, giving viewers an immediate, accurate, and visual sense of a match in progress.

IBM has mined more than seven years of Grand Slam Tennis data (approximately 39 million data points) to determine patterns and styles for players when they win. This insight is applied to determine the "keys" to the match for each player during the match:

- Prior to each match, the system runs an analysis of both competitors' historical head-to-head matchups, as well as statistics against comparable player styles, to determine what the data indicates each player must do to perform well in the match (SPSS technology).
- The system then selects the three most significant keys for each player in the match.
- As the match unfolds, the Keys to the Match dashboard updates in real-time with current game statistics.

Fans will find their voice within SlamTracker, adding comments about featured matches through Facebook and Twitter.

Chapter 6
Implementation of Big Data Analytics

This chapter provides glimpses of implementation plans and related challenges. Big Data Analytics may disrupt a major BI program. Do we use Big Data Analytics to radically transform this organization or evolve it for balanced growth? How do we establish a road map and find initial pilots? How do we evolve data governance to include considerations for Big Data?

6.1 Revolutionary, Evolutionary, or Hybrid

A typical Big Data Analytics implementation delivers three significant advancements in performance. First, it can reduce latency by an order of magnitude, providing accessibility to data in minutes or seconds as opposed to hours or days. Second, it increases the capacity to store data by an order of magnitude, moving from terabytes to petabytes. Third, it offers a much lower cost of acquisition and operation. Because the architecture is typically built on commodity hardware and requires fewer administrators, the cost, too, is reduced by an order of magnitude.

However, these implementations require a commitment to Big Data Analytics and a strong desire to migrate from the current platform. What if we have already invested a large IT budget in conventional BI? How far do we go in the first phase? Do we replace the current Data Warehouse architecture or augment it with Big Data Analytics tools? Both approaches have obvious pros and cons. In this section, I describe the three alternatives and discuss what would tilt us in one direction or another for a specific implementation.

Before I review the alternatives, let us first place the current environment in the context of the architecture described in Chapter 5 and understand how similar or dissimilar the architectures are.

In a typical “traditional” architecture, we have a set of components for ingesting data, a set of components for storing the data, and a set of components for analyzing the data and then feeding the results into a set of actions or reports. Since all the data must be routed via a storage medium using a data warehouse, the storage, organization, and retrieval of data creates a bottleneck. Typically, the traditional approach requires a reorientation of the data from the data source to a system of record and then into a set of models conducive to analytical processing—which typically requires a number of data modelers, database administrators, and ETL analysts to maintain the various data models and associated keys. Changes to the business environment require changes to models, which cascade into changes across each component and require large maintenance organizations.

Many components have already started to break off from this traditional model. Netezza as the Data Analytics engine does not strictly follow this paradigm, and it significantly reduces the model maintenance costs by reducing the need for representation and key-driven performance tuning. Use of SPSS and Cognos as user interfaces to drive modeling and reporting using Netezza's data manipulation capabilities reduces the repopulation of data in analytics tools.

The *revolutionary* approach involves creating a brand-new Big Data Analytics environment. We move all the data to the new environment, and all reporting, modeling, and integration with business processes happens in the new environment. This approach has been adopted by many greenfield analytics-driven organizations. They place their large storage in the Hadoop environment and build an analytics engine on the top of that environment to perform orchestration. The conversation layer uses the orchestration layer and integrates the results with customer-facing processes. The stored data can be analyzed using Big Data tools. This approach has provided stunning performance but has required high tooling costs and skills.

In a typical *evolutionary* approach, Big Data becomes an input to the current BI platform. The data is accumulated and analyzed using structured and unstructured tools, and the results are sent to the data warehouse. Standard modeling and reporting tools now have access to social media sentiments, usage records, and other processed Big Data items. Typically, this approach requires sampling and processing Big Data to shelve the warehouse from the massive volumes. The evolutionary approach has been adopted by mature BI organizations. The architecture has a low-cost entry threshold as well as minimal impact on the BI organization, but it is not able to provide the significant enhancements seen by the greenfield operators. In most cases, the type of analysis and the overall end-to-end velocity is limited by the BI environment.

The *hybrid* approach promoted actively by IBM's Information Agenda team places the AAP architecture on top of existing BI infrastructure. All the Big Data flows through AAP, while conventional sources continue to provide data to the data warehouse. We establish a couple of integration points to bring data from the warehouse into the analytics engine, which would be viewed by the data warehouse as a data mart. A sample of the AAP data would be directed back to the data warehouse, while most of the data would be stored using a Hadoop storage platform for discovery. The hybrid architecture provides the best of both worlds; it enables the current BI environment to function as before while siphoning the data to the AAP architecture for low-latency analytics. Depending on the transition success and the ability to evolve skills, the hybrid approach provides a valuable transition to full conversion.

Both the revolutionary and the hybrid architectures significantly challenge the data governance function. The next section describes the new set of issues and how to handle them.

6.2 Big Data Governance

Three broad categories of questions are emerging in the area of Big Data governance:

- *Single view of the customer*—We now have access to more complete data on how customers use their products for their communications, content, and commerce needs. How do we merge this newly acquired data with everything else we have been collecting to create a more comprehensive understanding of the customer?
- *Big Data veracity*—Customer data comes from a variety of "biased" samples with different levels of data quality. How do we homogenize this data, so that it can be used with confidence?
- *Information lifecycle management*—This is a lot more data than we have ever encountered before. Our current analytics systems are not capable of ingesting, storing, and analyzing these volumes at the required velocities. How do we store, analyze, and use this data in real-time or near real-time?

We will use this chapter to elaborate on these questions and will provide partial answers as they are known today.

Integrating Big Data with MDM

During the 1980s and 1990s, we created a series of departmental applications based on business cases associated with workforce automation. The result was a series of departmental databases containing customer, product, and related data. While the billing and sales views were often overlapping in these applications, it was not easy to map one to the other.

The past 10 years have seen a rapid rise in MDM for customer and product data across the enterprise. Analytics applications were the first consumers of master data to create mappings across multiple hierarchies as well as fragmented customer and product identifiers. MDM then graduated to transactional applications with much of the focus on business solutions, specifically customer relationship management (CRM) and billing systems. We can now use Big Data to build a comprehensive view of customer, network, and external data as demonstrated in the following case study.

Jim and Mary Smith have two children, Corey and Karen. The family has four phones, one for each family member. Corey and Karen are in high school and have basic phones for calls and text. Jim has an iPhone, which he uses primarily for office calls and emails. Mary has an iPhone and a WiFi-only iPad. She uses her iPad for investment research and participates in financial blogs.

When Jim received a brand-new iPhone from his employer as part of an upgrade program, he decided to give his older iPhone to Karen. Karen decided to sell her basic phone to a friend. Since they were in the last six months of their contract, the Smith family decided to keep the friend's phone on their plan until the end of the period. Karen's friend paid her for basic phone and messaging service.

The CSP providing phone service to the Smiths had done extensive household analysis to develop a customer hierarchy of their residence that tagged phones to users and connected all the users to the family account. After the changes mentioned above, the CSP's analytics applications would likely display abnormal calling patterns for the users compared with historical norms. In addition, Jim's old iPhone would show a number of web transactions that tracked to Jim's user ID but exhibited web browsing behaviors that were characteristic of a teenager. Karen's phone is now "hanging-out" in a new geohash.

Network data provides the best view of customer usage and trouble information. If this data is harnessed and offered as a strategic asset to others in the organization, it can provide a far more comprehensive understanding of the customer. In many cases, it may not even be important to connect the phone exchanges to the PII. The location and usage patterns may provide valuable insights about the user. The resulting view of customer hierarchies and households is far more accurate.

This case study would be even more dynamic if Karen were to borrow Jim's phone during a trip for a day or two. In growth markets, providers of prepaid services see massive churn in their customer base as consumers switch suppliers based on costs. The usage information can be used to track down a subscriber even as he or she switches telephone numbers. From a governance perspective, we must establish how the Big Data customer profile would be maintained, used, and integrated with the rest of MDM.

Big data brings new challenges to data quality management. If properly governed and managed, internal data quality can be measured and managed. Unfortunately, we have less control over the management of external data. However, it is even more important that we assess the value of the external data and its data quality. Merging of internal and external data should be done carefully based on an understanding of the quality of the external data and an appreciation for how the merged data will be used. Let us consider the following regarding the use of Twitter data.

A marketer launches a new product nationally and observes data relating to product sales, trouble tickets, network usage, and Twitter. A number of Tweets show consistently negative sentiments from Twitter users for the product. We are concerned that the data is an anomaly because sales of the product are brisk and there has been no significant increase in the number of trouble tickets.

Why is the Twitter data so out of whack? A closer analysis shows that older customers are relatively happy with the product and use surveys and trouble tickets to provide feedback. On the other hand, the product is not doing well with younger customers. The younger customers do not rely on traditional means of feedback and have been using Twitter to discuss the product in a negative way.

Because social media information is mostly self-reported, it is somewhat more prone to biased sampling. As a result, we must adopt a process to deal with Big Data quality during data aggregation. We must report the overall confidence level with the data, especially if it does not represent the entire population.

Big data can mean big storage, assuming all the data needs to be stored. Contrary to traditional data warehousing and analytics, we can perform Big Data Analytics at the time of data collection. As a result, we may need to maintain only a smaller subset, such as samples, filters, and aggregations in tier one storage.

Big data also provides its own tier two storage environment. Large quantities of unstructured data can be placed in Hadoop, which can be MapReduced later for any meaningful insight. A number of query tools are now available for large-scale queries on this data.

At the beginning of this chapter, we raised three questions for which we have provided partial answers as summarized below:

- *Single view of the customer*—We now have access to more complete data on how customers use their products for their communications, content, and commerce needs. As we merge this newly acquired data with everything else, we must closely monitor how the data is being used and how it is being aggregated. All this occurs as we radically change the rules on data privacy, redefine MDM, and encounter new concerns relating to data quality.
- *Big data quality*—Customer data comes from a variety of "biased" samples with different levels of data quality. As we homogenize this data, we must establish confidence levels on raw data, as well as aggregations and inferences, in order to understand and remind users of the "biases" built into the sourced data.
- *Information lifecycle management*—This is a lot more data than we have ever encountered before. Our current analytics systems are not capable of ingesting, storing, and analyzing these volumes at the required velocities. We may decide to store only samples of the data or use Hadoop for the storage and retrieval of large volumes of unstructured data.

We have explored a number of case studies, observations, and solutions in the chapter. This is a new field, and organizations are breaking new ground in terms of Big Data governance. We are sure to find new solutions to data quality, MDM, data privacy, and information lifecycle management as we deal with Big Data governance.

6.3 Journey, Milestones, and Maturity Levels

Big Data Analytics is a journey. What may be a bleeding-edge capability for one company or industry may be the base-level criteria for staying in business for another. This section describes a maturity model that allows us to measure the milestones in this journey so that we can benchmark a company in comparison with its peers. In Chapter 3, we discussed a number of business use cases. The maturity model can be applied to each of those use cases to help us measure the level of solution sophistication and the relative impact on KPIs. We can use the maturity model to represent the target state, current state, gaps, and relative maturity of the industry and competition.

Drivers are either internal or external forces that drive senior management priorities. For a commercial enterprise, factors such as revenue, cost, and customer acquisition and retention are typical drivers for its management to drive the organization's market valuation. For a government entity, the welfare and protection of citizens are typical drivers for analytics. For financial institutions, risk management is a key driver.

Capabilities represent a collection of business processes, people, and technology for a specific purpose. For example, a financial institution may have a risk management function for loan approval. The risk management would require technology components for statistical analysis and modeling, a set of trained people who can assemble risk management information from a variety of sources, and a risk management process that starts with risk data and ends with a score for a customer. Analytics supports a number of key capabilities in response to drivers. In the past five years, these capabilities have become increasingly sophisticated, as well as automated. Some of these capabilities are inter-organizational. For example, we discussed a set of business scenarios where retailers would collaborate with CSPs. As the amount of data has grown, so have the tools for faster data collection and real-time analytics. These tools have enabled a whole set of new capabilities. Let us examine a set of analytics-supported capabilities to support typical drivers.

Measurements are used to quantify the progress of a capability and its impact. With the increasing automation in products and processes, we now have many more ways to measure the effective functioning of a capability. These measurements can be visualized using a business value tree.

As we evaluate an analytics program, measurements help us visualize the capabilities required and their impact, thereby allowing management to prioritize program spending based on the capabilities that have the biggest impact to the organization. Measurements are used to link business capabilities to drivers. Value trees can also be used to identify common capabilities that impact multiple measurements and can be used to track benefits by program phases, identifying capabilities enabled by a particular phase. We can also maintain best practices for each capability to estimate the impact of a capability using past case studies.