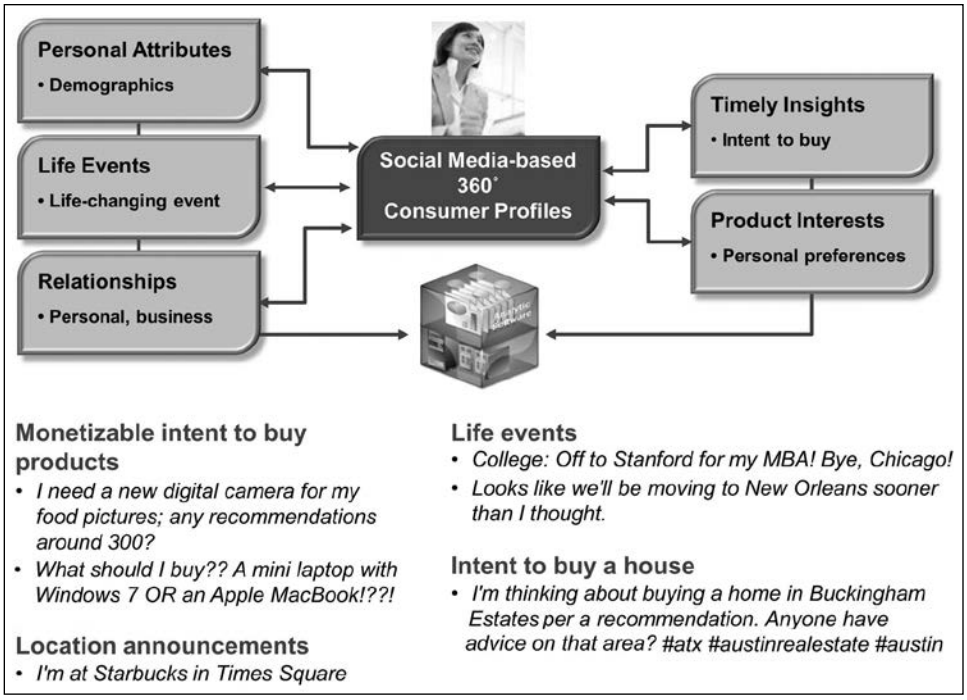


Figure 4.4: Big Data view of the customer

capabilities for using customer and product hierarchies to normalize data across systems. However, in most cases, the format for the data is known and the content is primarily structured. What does source data for Big Data's single view of a customer look like?

Blogs and tweets posted by consumers on social media sites provide a wealth of information for sentiment analysis; however, this data is not structured. Consumers do not always use proper company or product names. The data contains a fair amount of slang words, and there is a mix of languages in a multi-ethnic, multi-language environment. They may use a variety of words to convey positive or negative sentiments. The link to the author is not very well articulated. We start with scant information, such as Twitter handles and unstructured references, and filter and link this data to decipher demographics, location, and other important characteristics required in making this data meaningful to a marketer.

How do we now link the Twitter handle to the customer ID? Obviously, the customer would be the best person to link them together, and customers can sometimes be incentivized to do so with product promotions or information

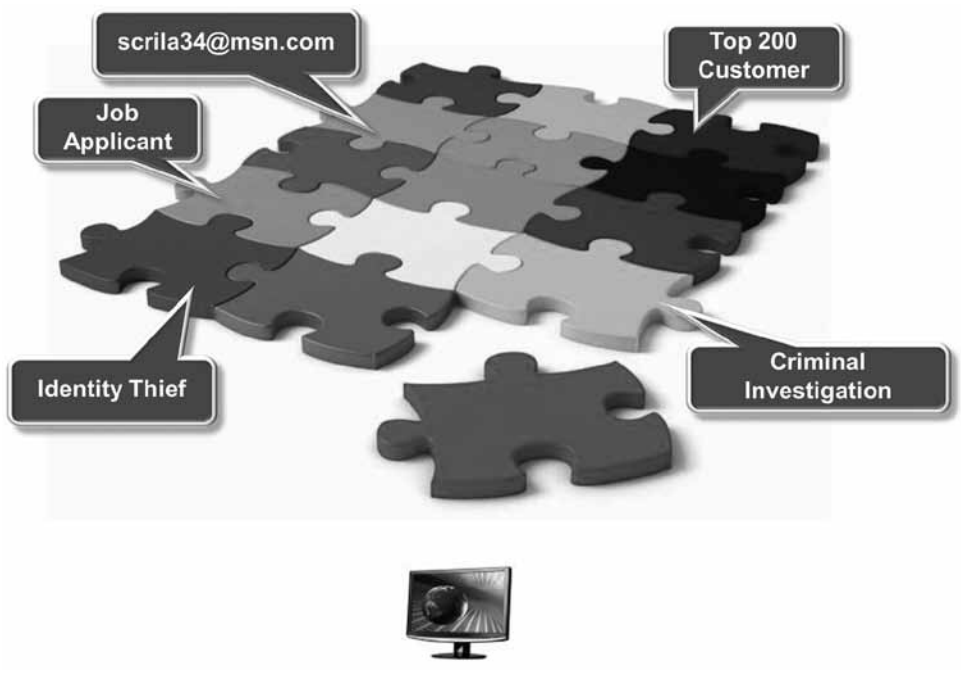


Figure 4.5: Identity resolution

exchange. When such direct means are not available, entity resolution technologies are providing ways to discover and resolve identities (see Figure 4.5).

Identity resolution is the next step in the evolution of matching technologies for MDM. Initially developed by Jeff Jonas for the casino industry, this is a powerful technology that takes into account both normal as well as deceptive data from customers. The technology is based on a set of rules that place the probability of a match on a set of seemingly unrelated facts. As hard facts match, the probabilities are altered to reflect newly discovered information. Customer-initiated actions, such as accepting a promotion, can be hard evidence added to customer handles or user IDs, connecting them to device IDs, product IDs, or customer account information.

Whenever I have made this idea part of a presentation, I have seen several raised eyebrows and questions about customer privacy. Customer privacy is always an area of major concern. For years, corporations collected all types of privacy information and matched it from a variety of sources to obtain a single view of customer. However, most of that information collection was transparent to the customer and happened with full disclosure. Now, however, Big Data has

the potential to correlate data across industries and across sources far more extensively than in the past. As a result, privacy is a major issue that I address in the next section.

4.4 Data Privacy Protection

I have banking/investment accounts with five major financial institutions. A major bank recently approached me to consolidate all my banking accounts with them. As we were going through the details, I was being asked to share a fair amount of private information. I wondered how much the bank already knew about me, since I have dealt with them for over a decade and have given them access to credit reports and mortgage applications. Also, a data scientist at the bank could correlate information authorized by me, information publicly available, and self-provided personal information. How is this full and complete view of my customer profile stored and accessed at the bank?

We have heard about data security breaches. Recently, the *Wall Street Journal* published an article about a Yahoo! security breach that exposed 453,000 unencrypted user names and passwords.[24] Is all this data that the bank is collecting about me safe? Often, we assume a large global brand is safe; however, the recent data breaches include a long list of famous brand names.

While an outright data breach is catastrophic and feared by most corporations, predictive models may uncover private facts not yet shared openly and could also lead to privacy loss. Let me illustrate with a famous news story from consumer marketing. Charles Duhigg, a staff writer for the *New York Times*, dug up the process for predicting consumer attributes used by the retailer Target. In his article, he describes how statisticians at Target created a sophisticated customer segmentation model that analyzed customer purchase behavior to predict customer life-cycle stages and related micro-segments. One of the predictive models was a "pregnancy-prediction" model that could predict with reasonable accuracy whether the customer making the purchase was expecting a baby. Unfortunately, the resulting Target campaign reached a house with a girl in high school, and her father decided to make a visit to the Target store. "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" The manager at Target promptly initiated an investigation to understand how the campaign was mailed to this girl. However, in a later communication, the girl's father apologized, stating, "I had a talk with my daughter. It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."[25]

The data privacy for Big Data is a serious business and is causing regulators around the globe to set up a variety of policies and procedures. Recently, the U.S. Federal Trade Commission settled a case with Facebook that now requires Facebook to conduct regular audits. Facebook, Inc., agreed to submit to the government audits of its privacy practices every other year for the next two decades. The company also agreed to obtain explicit approval from users before changing the type of content it makes public.[26] Similar processes have been put in place at MySpace and Google. In many cases, consumers trade their privacy for favors. For example, my cable/satellite provider sought to have my channel click information shared with a search engine provider. They offered me a discount of $10 if I would "opt-in" and let them monetize my channel surfing behavior.

This leads us to several interesting possibilities. Let us say that a data scientist uses the channel surfing information to characterize a household as interested in sports cars (for example, through the number of hours logged watching Nascar). The search engine then places a number of sports car advertisements on the web browser used by the desktop in that household and places a web cookie on the desktop to remind them of this segmentation. Next, a couple of car dealers pick up this "semi-public" web cookie from the web browser and manage to link this information to a home phone number. It would be catastrophic if these dealers were to start calling the home phone to offer car promotions. When I originally opted in, what did I agree to opt-in to, and is my cable/satellite provider protecting me from the misuse of that data? As we move from free search engines to free emails to discounted phones to discounted installation services, all based on monetization of data and advertising revenue, there is money for everyone, if the data is properly protected against unauthorized use.

The first part of the solution is a data obfuscation process. Most of the time, marketers are interested in customer characteristics that can be provided without Privately Identifiable Information (PII)—that is, uniquely identifiable information about the individual that can be used to identify, locate, and contact an individual. We can possibly destroy all PII information, which may still provide useful information to a marketer about a group of individuals. Now, under "opt-in," the PII can be released to a selected few, as long as it is protected from the rest. In the preceding example, by collecting $10, I may give permission to a web search engine to increase sports car advertisements to everyone in my Zip+4 while at the same time expecting protection from dealer calls, which require a household-level granularity. We can provide this level of obfuscation by destroying PII for house number and street name while leaving Zip+4 information in the monetized data.

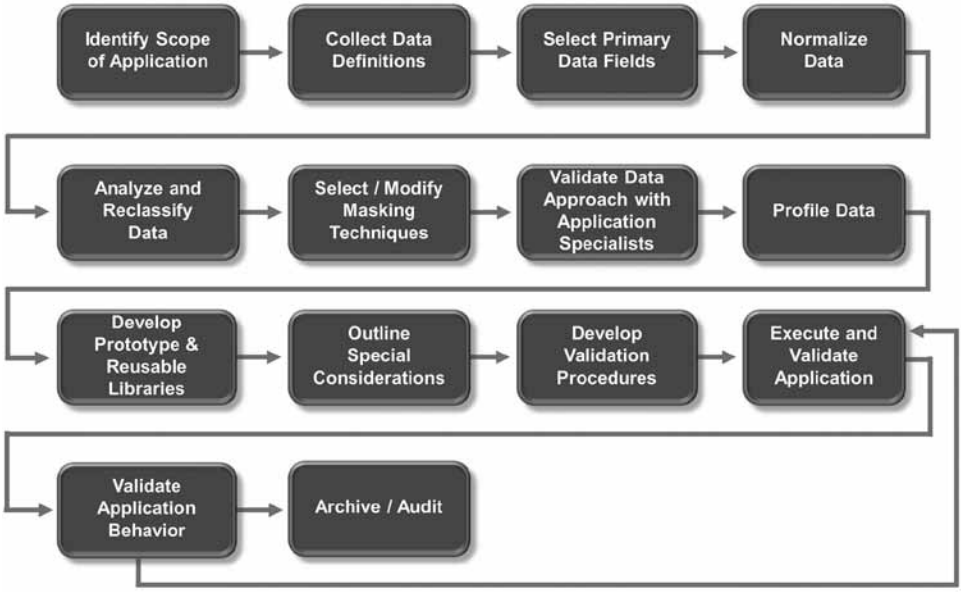


Figure 4.6: Data masking process

As we worked on the data obfuscation process, we found that this process is a lot more complex than expected. While PII data is destroyed, we cannot leave related information that, if joined with obfuscated data, might lead to the individual. For example, if we destroyed the address and phone number but left location information, someone could use the location information to establish the residential address. Also, there are grades of PII information. Zip+4 or county designation may be an appropriate locater unless we are dealing with home addresses of billionaires. Also, small samples are a problem. The non-PII information could uniquely identify an individual if only one individual meets the profile. IBM has been investing in data masking products and processes, which allow us to systematically identify PII information in a data set, tag it, select masking algorithms, test the masking process, and establish the effectiveness of the masking solution (see Figure 4.6 and details on the USPTO site[27]).

Data masking algorithms are equally interesting. The algorithm should remove or randomize PII but not destroy statistical patterns required by a data scientist. For example, if we take a set of real addresses and replace them with XXX, anyone looking for statistical patterns along geographic boundaries would not be able to use the obfuscated data. We have developed a number of algorithms that preserve the uniqueness of the IDs or the statistical patterns while obfuscating the PII data in groups.[28]

A privacy infrastructure provides the capability to store information about "opt-in" and to use it for granting access. Anyone with proper access can obtain the PII information, as granted by the user, while others see only obfuscated data. This solution provides us with enormous capability to use statistical data for a group of individuals while selectively offering "one-to-one" marketing wherever the consumer is willing to accept the offers.

An audit can test whether the obfuscation process, algorithms, and privacy access are working properly in a multi-partner environment where third parties may also have access to this data. If properly managed, the data privacy framework provides gated access to marketers based on permission granted by the consumer and can significantly boost consumer confidence and ability to finance data monetization.

4.5 Real-Time Adaptive Analytics and Decision Engines

As I was watching a movie online recently, the website displayed an advertisement every 10 minutes. Since I had not paid anyone for watching the movie and am used to watching commercials on television, it should not have been a big deal to see a commercial every 10 minutes or so. The website, however, decided to show me the same commercial over and over. After about the fifth time, I felt sorry for the poor advertiser (one of the two U.S. presidential candidates) because the effectiveness of the ad had long since dissipated and, instead, an annoyance factor had crept in. I was facing a real-time decision engine that was rigid and was placing an advertisement without any count or analysis of saturation factor. (As an aside, I live in a "swing state" for the fall 2012 U.S. presidential elections, so it is possible the advertising agency for the candidate had decided to saturate the Internet advertisements to my location.)

How do we build real-time decision engines that are based not on static rules but on real-time analytics and are adaptive, introducing changes as they are executed? In real-time ad placement, a number of factors could have been used as input using information already available to the website that was offering me the movie. For example, the site was well aware of the movie genre I had accessed during several visits to the site. An analysis of this genre could have placed me in several viewing segments. In fact, the same website offers me movie recommendations, which are based on prior viewing habits. This recommendation engine could be offered to the marketers in placing advertisements that match my viewing habits.

Advertisements saturate over a given number of times, after which any additional viewing is ineffective. The website could count the number of times

an ad was displayed and decrement the likeness score for the specific ad each time it was shown, thereby favoring a different advertisement to be shown after a certain number of views. A number of sophisticated marketing experiments can be run to effectively control the saturation effect.

After watching the movie for a while and repeatedly viewing the same advertisement, I decided to take a break to search for a food processor as a gift to my son. When I returned to watching the movie, I again faced the same advertisement. I had secretly hoped that my food processor search would conveniently trigger an advertisement for a good food processor to help me in my purchase. By sensing and analyzing my previous web searches, marketers could have offered me appropriate information or deals, thereby increasing the advertisement relevance for me.

It seems there are two sets of input variables that constantly impact the success of advertising. The first includes search context, saturation, and response to advertisement and is fast-moving and must be tracked in real-time. The second set includes viewing habits, shopping behaviors, and other micro-segmentation-related variables and is either static or changes gradually, accumulating over a long time period.

How would real-time adaptive analytics and decision engines acquire the background and accommodate changes to the models while at the same time rapidly executing the engine and providing a context-dependent response? There are four components of a real-time Adaptive Analytics and Decision Engine (see Figure 4.7).

A sensor identifies an incoming event with a known entity. If we do not identify this identity, we can create a new one. However, if this is a known entity that we have been tracking, we will use the identifiers in the event to connect it to previous history for this entity. The entity can be an individual, a device (e.g., a smartphone), or a known web browser identified via a previously placed cookie

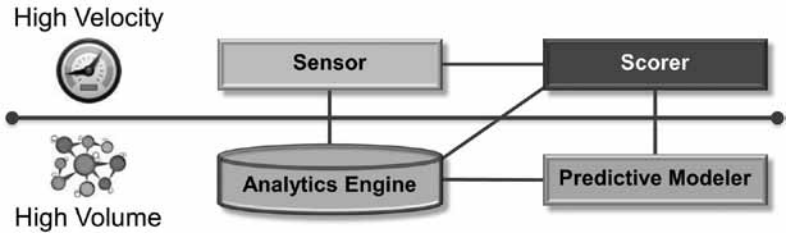


Figure 4.7: Real-time adaptive analytics

or pixel (see note[29] for web tracking technologies). Under opt-in, if we placed a coupon on a smartphone and the person opted-in by accepting the coupon, we may have a fair amount of history about the individual. The analytics engine maintains a detailed customer profile based on past-identified history about the entity. The predictive modeler uses predictive analytics to create a cause-effect model, including impact of frequency (e.g., saturation in advertisement placement), offer acceptance, and micro-segmentation. The scorer component uses the models to score an entity for a prospective offer.

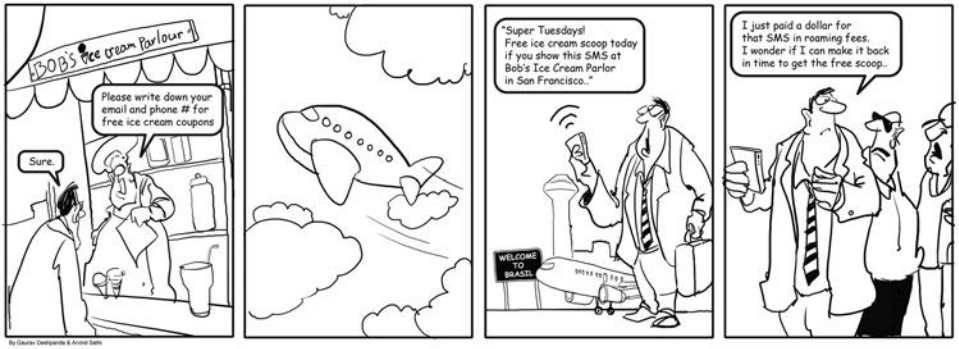
While sensor and scorer components may operate in real-time, the analytics engine and predictive modeler do not need to operate in real-time but work with historical information to change the models. Returning to our example of online advertising, a cookie placed on the desktop identifies me as the movie watcher and can count the number of times an ad has been shown to me. The scorer decrements an advertisement based on past viewership for that advertisement. The analytics engine maintains my profile and identifies me as someone searching for a food processor. The predictive modeler provides a model that increases the score for an advertisement based on past web searches. The scorer picks up my context for web search and places a food processor advertisement in the next advertisement placement opportunity. The sensor and scorer work in milliseconds while the analytics engine and the modeler work in seconds or minutes.

Without a proper architecture, integration of these components could be challenging. If we place all of these components in the same software, the divergent requirements for volume and velocity may choke the software. The real-time components require rapid capabilities to identify an entity and use a number of models to score the opportunity. The task is extremely memory and CPU intensive and should be as lean as possible. On the other hand, the analytics engine and predictive modeler may carry as much information as possible to conduct accurate modeling, including three to six months of past history and the ability to selectively decay or lower the data priority as time passes by or subsequent events confirm purchases against previously known events. I may be interested in purchasing a food processor this week, and would be interested in a couple of well-placed advertisements, but the need will diminish over time as I either purchase one or lose interest.

As we engage with consumers, we have a number of methods to sense their actions, and a number of stages of engagement. A typical online engagement process may track the following stages:

- *Anonymous customer*—We do not know anything about the customer and do not have permission to withhold information.
- *Named customer*—We have identified the customer and correlated to identification information such as device, IP address, name, Twitter handle, or phone number. At this stage, specific personal information cannot be used for individual offers because of lack of opt-in.
- *Engaged customer*—The customer has responded to an information request or advertisement and is beginning to shop based on offers.
- *Opted-in customer*—The customer has given us permission to send offers or track information. At this stage, specific offers can be individualized and sent out.
- *Buyer*—The customer has purchased a merchandise or service.
- *Advocate*—The customer has started to "Like" the product or is posting favorably for a campaign.

A real-time Adaptive Analytics and Decision Engine can help us track a customer through these stages and engage in a conversation to progress a customer from one stage to the next.


BOB's Ice cream Parlour
Please write down your email and phone # for free ice cream coupons
Sure.
"Super Tuesdays! Free ice cream scoop today if you show this SMS at Bob's Ice Cream Parlor in San Francisco.."
WELCOME TO BRASIL
I just paid a dollar for that SMS in roaming fees. I wonder if I can make it back in time to get the free scoop..
By Gaurav Deshpande & Arvind Sathi

Chapter 5
Advanced Analytics Platform

The architecture components described in the previous chapter must be placed in an integrated architecture where they can all coexist and provide overall functionality and performance consistent with our requirements. However, the requirements are at odds with each other. On one hand, we are dealing with unstructured data discovery over very large data sets that may have very high latency. On the other hand, the adaptive analytics activities are bringing the analytics to a conversation level requiring very low latency. How do we establish an overall architecture that respects both of these components equally while establishing a formalized process for data integration? This chapter describes an integrated architecture that responds to these challenges and establishes a role for each component that is consistent with its capabilities. The architecture outlined in this chapter is *Advanced Analytics Platform (AAP)*. We have been experimenting with the integration of architecture components in IBM's Dallas Global Solution Center using a physical implementation of AAP.

I will use an analogy from sports television coverage to demonstrate how this architecture closely follows the working behavior of highly productive teams. I have always been fascinated with how a sports television production is able to cover a live event and keep us engaged as an audience using a combination of real-time and batch processing. The entire session proceeds like clockwork. It is almost like watching a movie, except that the movie is playing live with just a small time buffer to deal with catastrophic events (like wardrobe malfunctions!).

As the game progresses, the commentators use their subject knowledge to observe the game, prioritize areas of focus, and make judgments about good or bad plays. The role of the director is to align a large volume of data, synthesize the events into meaningful insight, and direct the commentators to specific focus areas. This includes replays of moves to focus on something we may have

missed, statistics about the pace of the game, or details about the players. At the same time, statisticians and editors are working to discover and organize past information, some of which is structured (e.g., the number of double faults in tennis or how much time the ball was controlled by one side in American football). However, other information being organized is unstructured, such as an instant replay, where the person editing the information has to make decisions about when to start, how much to replay, and where to make annotations on the screen to provide focus for the audience. The commentators have the experience and expertise to observe the replays and statistics, analyze them in real-time, and explain them as they do for the game itself.

The commentators process and react to information in real-time. There cannot be any major gaps in their performance. Most of the data arrives in real-time and is processed and responded to in real time as well. The director has access to all the data that the commentators are processing, as well as the commentators' responses. The director then has to script the next couple of minutes, weighing whether to replay the last great tennis shot or football catch, focus on the cheering audience, or display some statistics. In the course of these decisions, the director scans through many camera views, statistics, and replay collections and synthesizes the next scenes of this live "movie." Behind the scenes, the statisticians and editors are working in a batch mode. They have all the history, including decades' worth of statistics and stock footage of past game coverage. They must discover and prioritize what information to bring to the director's attention.

Let us now apply this analogy to the Big Data Analytics architecture, which consists of three analytics layers. The first is a real-time architecture for conversations; this layer closely follows the working environment of the commentators. The second is the orchestration layer that synthesizes and directs the analysis process. Last, the discovery layer uses a series of structured and unstructured tools to discover patterns and then passes them along to the orchestration layer for synthesis and modeling.

Figure 5.1 maps the Big Data Analytics architecture to the sports television roles.

5.1 Real-Time Architecture for Conversations

Let us start with the low-latency environment first. What are the characteristics of the information alignment and assessment process during a conversation, and how does it differ from the other two layers? How do we ensure that the conversation layer makes rapid inferences and leverages all the hard work by the lower layers?