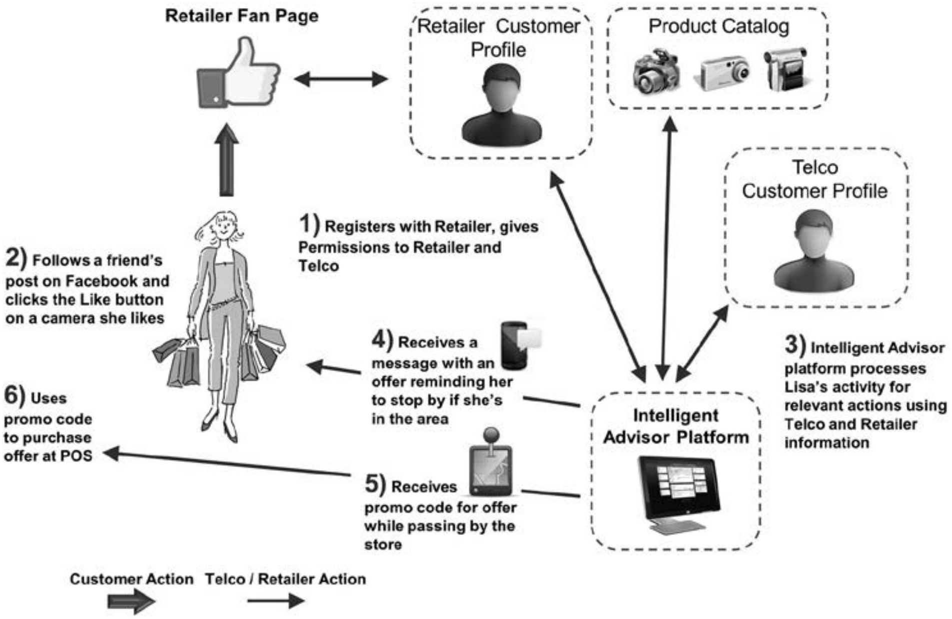
Retailer Fan Page
Retailer Customer Profile
Product Catalog
Telco Customer Profile
1) Registers with Retailer, gives Permissions to Retailer and Telco
2) Follows a friend's post on Facebook and clicks the Like button on a camera she likes
3) Intelligent Advisor platform processes Lisa's activity for relevant actions using Telco and Retailer information
4) Receives a message with an offer reminding her to stop by if she's in the area
5) Receives promo code for offer while passing by the store
6) Uses promo code to purchase offer at POS
Intelligent Advisor Platform
Customer Action
Telco / Retailer Action

Figure 3.3: Intelligent Advisor

purchase, or air commercials based on what a specific consumer is buying. It requires our ability to connect retail and cable advertising data as well as an ecosystem where the two analytics systems (retail and cable) can collaborate.

A triple-play CSP (providing cable, broadband, and wireless services) could use its customer database to correlate customer activities across these three screens. Many consumers are viewing media using Internet over their desktops or tablets. We can now start to correlate media viewing, location-based micro-segments, and customer purchase intentions as known through social media to make retail offers. Figure 3.3 shows a scenario where consumer profiles from CSP and retail segments are used for creating context- and micro-segment-based offers to consumers. The consumer registers on the retailer's website, giving permission to the retailer to use profile data. The retailer uses consumer context and location to tailor a specific promotion.

3.7 Online Advertising

Television and radio have used advertising as their funding model for decades. As online content distribution becomes popular, advertising has followed the content distribution with increasing volumes and acceptance in the marketplace. The recently concluded Olympics in London provided a testament to the popularity of mobile and other online media distribution channels as compared with television. Almost half of the Internet video delivered during the Olympics went to mobile phones and tablets. That's a watershed for portable TV. Nearly 28 million people visited *NBCOlympics.com*, eight percent higher as compared with the Beijing Olympics four years ago. Sixty four million video streams were served across all platforms, a 182 percent increase over Beijing. Nearly 6.4 million people used mobile devices.[18]

Online advertising is also becoming increasingly sophisticated. I discussed the supply chain for digital advertising with a number of specialized players in Section 2.3. The biggest focus is the advertisement bidding managed for a publisher, such as Google, by either a Supply Side Platform (SSP) or Advertising Exchange. Online advertising provides tremendous opportunity for advertising to a micro-segment and also for context-based advertising. How do we deliver these products, and how do they differ from traditional advertising?

The advertiser's main goal is to reach the most receptive online audience in the right context, who will then engage with the displayed ad and eventually take the desired action identified by the type of campaign.[19] Big Data provides us with an opportunity to collect myriads of behavioral information. This information can be collated and analyzed to build two sets of insights about the customers, both of which are very relevant to online advertising. First, the micro-segmentation information and associated purchase history described in Section 3.6 allows us to establish buyer patterns for each micro-segment. Second, we can use the context of an online interaction to drive context-specific advertising. For example, for someone searching and shopping for a product, a number of related products can be offered in the advertisements placed on the web page.

Over the past year, I found an opportunity to study these capabilities with the help of Turn Advertising. Turn's Demand Side Platform (DSP) delivers over 500,000 advertisements per second using ad bidding platforms at most major platforms, including Google, Yahoo, and Facebook. A DSP manages online advertising campaigns for a number of advertisers through real-time auctions or bidding. Unlike a direct buy market (e.g., print or television), where the price is decided in advance based on reach and opportunities to see, the real-time Ad Exchange accept bids for each impression opportunity, and the impression is sold to the highest bidder in a public auction. DSPs are the platforms where all the information about users, pages, ads, and campaign constraints come together to make the best decision for advertisers.

Let us consider an example to understand the flow of information and collaboration between publisher, Ad Exchange, DSP, and advertiser to deliver online advertisements. If a user initiates a web search for food in a particular zip code on a search engine, the search engine will take the request, parse it, and start to deliver the search result. While the search results are being delivered, the search engine decides to place a couple of advertisements on the screen. The search engine seeks bids for those spots, which are accumulated via Ad Exchange and offered to a number of DSPs competing for the opportunities to place advertisements for their advertisers. In seeking the bid, the publisher may supply some contextual information that can be matched with any additional information known to the DSP about the user. The DSP decides whether to participate in this specific bid and makes an offer to place an ad. The highest bidder is chosen, and their advertisement is delivered to the user in response to the search. Typically, this entire process may take 80 milliseconds.

A Data Management Platform (DMP) may collect valuable statistics about the advertisement and the advertising process. The key performance indicators (KPIs) include the number of times a user clicked the advertisement, which provides a measure of success. If a user has received a single advertisement many times, it may cause saturation and reduce the probability that the user will click the advertisement.

As online advertising is integrated with online purchasing, the value of placing an advertisement in the right context may go up. If the placement of the ad results in the immediate purchase of the product, the advertiser is very likely to offer a higher price to the publisher. DSP and DMP success depends directly on their ability to track and match consumers based on their perceived information need and their ability to find advertising opportunities related closely to an online sale of associated goods or services.

3.8 Improved Risk Management

A credit-card company can use cell phone location data to differentiate an authentic user from a fraudulent one. As the credit card is used in a location, the credit-card transaction location can be matched with the cell phone location for the customer to reduce risk of fraudulent transactions.

My work requires me to travel often, almost once a week. Because I travel to a variety of international destinations frequently but use my personal credit card rarely, any purchase with it is very likely to be tagged as unusual activity. This behavior places me under the close scrutiny of the credit-card company's fraud engine because the usage is sporadic and geographically diverse. Invariably, my credit card is occasionally denied at the time of purchase, requiring me to telephone the call center for security verification. I remember talking to a support line three times from India, with each call taking ten minutes or longer. The overall cost of such a call, including telephone charges, the call center agent's time, and my time, adds up significantly.

While I am thankful to the credit-card company for taking my card security seriously, I was curious whether there was an easier way for them to deal with this situation. I asked the credit-card call center agent how I could make the credit-card company's monitoring easier, and the response was to call them before each trip. This solution might reduce the number of times my credit card is denied; however, it would significantly increase the call-center costs. Plus, I would have to make a call every time I traveled, which could be a lot more calls than the number of times my personal credit card is used.

The premise for credit-card fraud is that someone could steal my credit card and use it. A typical fraud rule looks for an unusual purchase initiated in a new international location. Unfortunately, for frequent travelers like me, irregular personal credit-card use can easily mimic these fraudulent transactions. However, I carry a smartphone all the time when I travel. Although my credit-card company may not know of my travel to distant geographies, my smartphone has full awareness of my location. Also, the chances of my losing both my credit card and my phone are significantly lower, and even if someone picked up both, it is highly unlikely they would travel with both credit card and smartphone to make fraudulent purchases. If only I could authorize my credit-card company to check my phone location each time there is a concern about the credit-card usage, and even download an app to my phone that could ask me to authorize the charges using a secure login or password to eliminate the possibility of my phone being stolen at the same time.

Financial institutions are rapidly using smartphones for banking transactions. Today, Chase offers mobile check deposit using the Apple iPhone (see *https://www.chase.com/online/services/check-deposit.htm*). Using my iPhone camera, I can take a picture of both sides of the check and then use the Chase Mobile app on my iPhone to log into my account with a special authorization ID supplied by Chase. Now that my phone and bank are aware of each other, they can use this information for a variety of applications to improve my customer experience.

Chapter 4
Architecture Components

Big Data requires technical capabilities in dealing with velocity, variety, veracity, and volume. A number of emerging applications are scaling to velocity in milliseconds; a variety of unstructured text, sounds, videos, and semi-structured machine-to-machine data; veracity-based weighting; and volumes ranging up to petabytes. This is a tall order and unthinkable in the legacy analytics environment. How do we build these applications, and how do we integrate them with the current environment, which may only be dealing with terabytes of data at "D-1" velocity? A number of architectural components are evolving to deal with these extreme levels of velocity, variety, veracity, and volume. This chapter samples some of the most significant technical components required for Big Data Analytics to work.

I have taken standard components from traditional analytics systems: data ingestion and storage, reporting, master data management (MDM), predictive modeling, and data privacy. For each of these areas, I describe the challenges faced in this brave new world of Big Data. In most cases, we need significant new technical capabilities for extending the current architecture to include Big Data. It is often an easy decision to evolve the current capabilities to include some aspects of Big Data and call it a success. However, as the data tsunami hits the shores, the key question is whether these evolutionary approaches will suffice for ongoing tides of Big Data, or will they get buried in the tide leading to crisis and catastrophic competitive losses. In many industries, business leaders are aggressively turning to Big Data analytics in each of these capability areas to pilot, integrate, or replace the current environment.

4.1 Massively Parallel Processing (MPP) Platforms

Big Data usually shows up with a data tsunami that can easily overwhelm a traditional analytics platform designed to ingest, analyze, and report on typical customer and product data from structured internal sources. In order to meet the volume challenge, we must understand the size of data streams, the level of processing, and related storage issues. The entire analytics environment must have the capabilities to deal with this data tsunami and should be prepared to scale up as the data streams get bigger.

The use of massively parallel computing for tackling data scalability is showing up everywhere. In each case, the underlying principle is a distribution of workload across many processors as well as storage and transportation of underlying data across a set of parallel storage units and streams. In each case, the manipulation of the parallel platform requires a programming environment and an architecture, which may or may not be transparent to the applications.

Let us start with the platform for large-scale data integration. Any environment facing massive data volumes should seriously consider the advantages of High Performance Shared Service Grid/HA computing as a means to host their data integration infrastructures. Today, the maturity of High Performance Shared Services Grid/Stream/HA computing is such that it is now common, with most companies including it in their strategic planning for architectures and enterprise data centers. The core tenets of Grid/Stream/HA computing are the same as traditional clusters and massively parallel processing (MPP) solutions in terms of the desire to maximize the use of available hardware to complete a processing task. However, what is new about the IBM Information Server Grid/HA and InfoSphere Streams environments is their ease of setup and use, the unlimited linear scalability to thousands of nodes, the fully dynamic load node/pod balancing and execution, the ability to achieve automatic high availability/disaster recovery (HA/DR), and the much lower price points at which you can achieve comparable performance of traditional symmetric multiprocessing (SMP) shared memory server configurations. The overall adoption of grid computing is now becoming commonplace, and it is accelerating, driven by the price-to-performance statistics, flexibility, and economics. The key to the success of the Grid/Stream/HA Shared Service implementations lies in the entire solution working for the business providing straight through processing (STP) with dynamic process allocation, flexibility, and scale-up. A typical parallel data integration platform:

- Designs an integration process without concern for data volumes or time constraints
- Leverages database partitioning schemes for optimal load performance
- Simplifies steps to define partitions within each process if needed
- Uses a single configuration file to add processors and hardware

- Requires no hand coding of programs to enable more processors
- Supports SMP, clustered, grid, and MPP platforms

In InfoSphere Streams, continuous applications are composed of individual operators, which interconnect and operate on one or more data streams. Data streams normally come from outside the system or can be produced internally as part of an application. The operators may be used on the data to have it filtered, classified, transformed, correlated, and/or fused to make decisions using business rules. Depending on the need, the streams can be subdivided and processed by a large number of nodes, thereby reducing the latency and improving the processing volumes.

The Netezza Performance Server (NPS®) system's architecture is a two-tiered system designed to handle very large queries from multiple users. The first tier is a high-performance Linux® symmetric multiprocessing host. The host compiles queries received from Business Intelligence applications and generates query execution plans. It then divides a query into a sequence of subtasks, or snippets, which can be executed in parallel, and it distributes the snippets to the second tier for execution. The host returns the final results to the requesting application, thus providing the programming advantages while appearing to be a traditional database server. The second tier consists of dozens to hundreds to thousands of Snippet Processing Units (SPUs) operating in parallel. Each SPU is an intelligent query processing and storage node and consists of a powerful commodity processor, dedicated memory, disk drive, and field-programmable disk controller with hard-wired logic to manage data flows and process queries at the disk level.

The massively parallel, shared-nothing SPU blades provide the performance advantages of massively parallel processors. Nearly all query processing is done at the SPU level, with each SPU operating on its portion of the database. All operations that easily lend themselves to parallel processing (including record operations, parsing, filtering, projecting, interlocking, and logging) are performed by the SPU nodes, which significantly reduces the amount of data moved within the system. Operations on sets of intermediate results, such as sorts, joins, and aggregates, are executed primarily on the SPUs but can also be done on the host, depending on the processing cost of that operation.

A recent development in the scalability for databases is evident from IBM's pureScale offering. Designed for organizations that run online transaction processing (OLTP) applications on distributed systems, IBM® DB2® pureScale® offers clustering technology that helps deliver high availability and exceptional scalability transparent to applications. Based on technology from IBM DB2 for z/OS®, DB2 pureScale is available as an option on IBM DB2 Enterprise Server Edition and Advanced Enterprise Server Edition, offering continuous availability, application cluster transparency, and extreme capacity.

Hadoop owes its genesis to the search engines, as Google and Yahoo required massive search capabilities across the Internet and addressed the capability of searching in parallel with data stored in a number of storage devices. Hadoop offers the Hadoop Distributed File System (HDFS) for setting up a replicated data storage environment and MapReduce, a programming model that abstracts the problem from disk reads and writes and then transforms it into a computation over a set of keys and values.[20] With the open source availability, Hadoop has rapidly gained popularity, especially in Silicon Valley.

When dealing with high volumes and velocity, we cannot leave any bottlenecks. All the processes, starting with data ingestion, data storage, analytics, and its use, must meet velocity and volume requirements. Some of these systems are designed to be massively parallel and do not require configuration or programming to enable massively parallel activities. In some cases, such as Hadoop, the parallel processing requires programming using special tools, which exploit the parallel nature of the underlying environment (in this case, HDFS). The Hadoop development environment includes Oozie, an open-source workflow/coordination service to manage data processing jobs; HBase for random, realtime read/write access to Big Data; Apache Pig for analyzing large data sets; Apache Lucene for search; and Jaql for query using JavaScript® Object Notation (JSON). Each component leverages Hadoop's MapReduce for parallelism; however, this elevates the skill level required for building applications. To make the environment more user friendly, IBM is introducing a series of tools, such as Big Sheets, that help to visualize the unstructured data.

4.2 Unstructured Data Analytics and Reporting

Traditional analytics has been focused primarily on structured data. Big Data, however, is primarily unstructured, so we now have a couple of combinations available. We can perform quantitative analysis on structured data as before. We can extract structure out of unstructured data and perform quantitative analysis on the extract quantifications. Last, but not least, there is a fair amount of non-quantitative analysis now available for unstructured data. This section explores a couple of techniques rapidly becoming popular with the vast amount of unstructured data and looks at how these techniques are becoming mainstream with their powerful capabilities for organizing, categorizing, and analyzing Big Data.

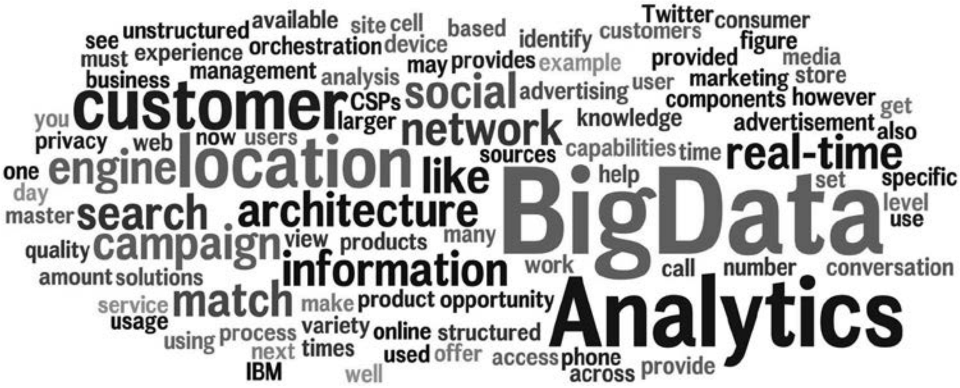

Figure 4.1: Wordle™ Word Cloud

Search and Count

Google and Yahoo rapidly became household terms because of their ability to search the web for specific topics. A typical search engine offers the ability to search documents using a set of search terms and may find a large number of candidate documents. It prioritizes the results based on preset criteria that can be influenced by how we choose the documents.

If I have a lot of unstructured data, I can count words to find the most commonly used words. Wordle™ (*www.wordle.net*) provides word clouds for the unstructured data provided to it. For example, Figure 4.1 shows a word cloud for the text used in this book. The font size represents the number of times a word was used in the text.

This data can be laid out against other known dimensions. For example, this summer we were working on unstructured data analytics for a CSP in India. We received a large quantity of unstructured text. Our first exercise was to use the Text Analytics capabilities in Cognos® Consumer Insight (CCI) to study key words being used as plotted against time. Figure 4.2 shows the results of this word count plotted against time.

Context-Sensitive and Domain-Specific Searches

Anyone with telecommunications knowledge can easily understand what “3g” and “4g” in Figure 4.2 refer to. Context-sensitive search engines can differentiate between “gold medal” (Olympics) and “gold bullion” (commodity trading). Also, some of the search engines are fine tuned for industry or corporate terms.

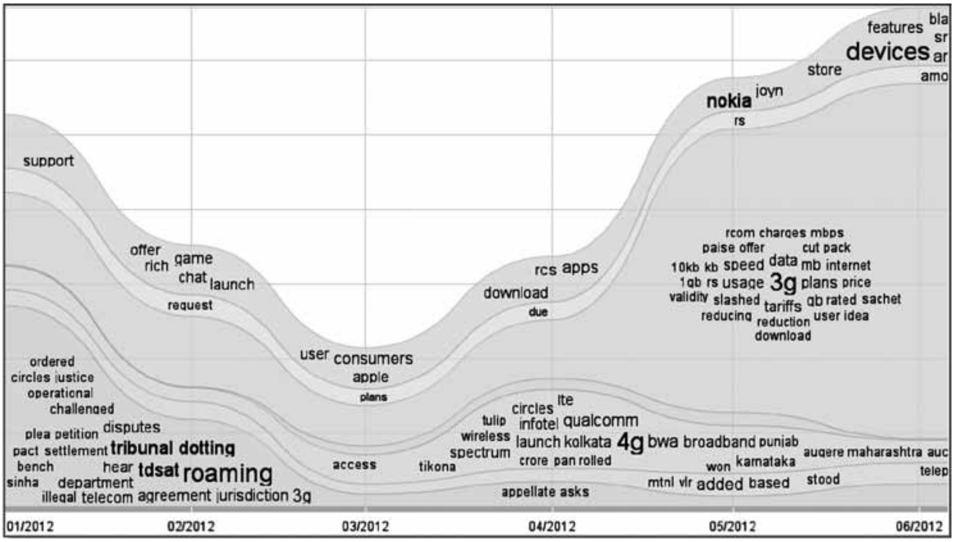

Figure 4.2: Word count graphical display plotted against time

Vivisimo offers the capability to specialize a search engine for a specific purpose, thereby fine tuning it for corporate terms when used inside a corporate intranet.

Categories and Ontology

Often, we like to classify unstructured data into categories. This gives us an understanding of the relative distribution across a known classification scheme. Let me use an example from online purchasing. I use Slice (*www.slice.com*) to keep track of my online purchases. Slice scans my email for any online purchases and extracts relevant information so I can track shipments, order numbers, purchase dates, and so on. Slice also lets me “slice and dice” the orders. That is, it analyzes my purchases against a set of categories to report the number of items and money spent in each category. Figure 4.3 shows Slice’s category analysis: Travel & Entertainment, Music, Electronics & Accessories, and so on. Slice must be doing rigorous unstructured analytics to understand what is considered “Movies & TV” and how that is different from “Music.”

The classic product categories originated from the Yellow Pages. We remember the classic Yellow Pages books that we received so often and are nowadays getting incorporated into online Yellow Pages and other shopping and ordering tools. However, categories are typically tree structured, where each node is a sub-class of the node above and can be further sub-classified into further specialized nodes. For example, a scooter is a sub-class of two-wheeler, while an electric scooter is a sub-class of scooter. A node can be a sub-class of

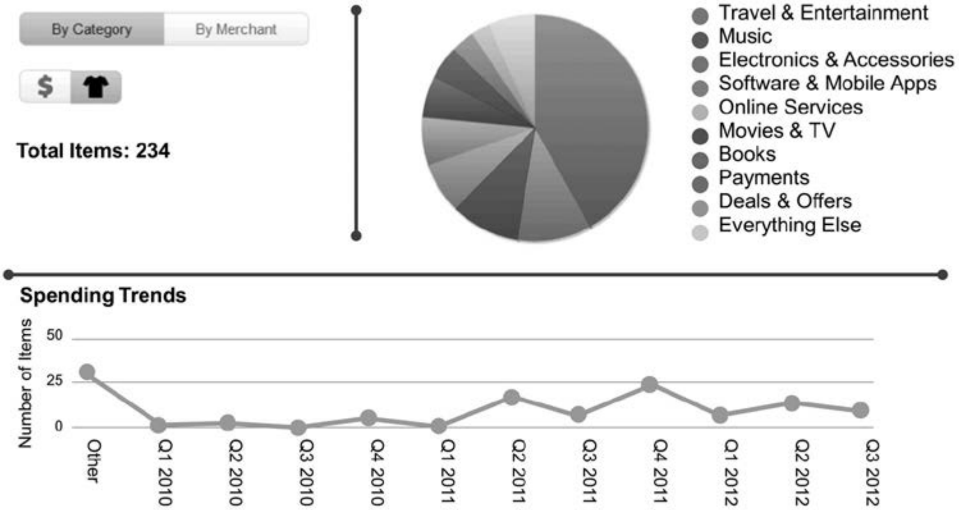

Figure 4.3: "Slice and dice" using product categories

more than one entity. A sub-class shares the attributes of its super-class. Therefore, both scooters and electric scooters should have two wheels. While the classic product catalogs were static and were managed by administrators without organized feedback, the unstructured analytics provides the ability to make a dynamic hierarchy, which can be adjusted based on usage and search criteria.

A more general representation of conceptual entities is found in ontology, which is an abstract view of the world for some purpose.[21] Ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information (a domain is a specific subject area or area of knowledge, like medicine, tool manufacturing, real estate, automobile repair, financial management, and so on) and may include classifications, relationships, and properties.[22] With formal ontology, we can create a "Semantic Web," which can provide structural extracts to machines, thereby providing them with ability to extract, analyze, and manipulate the data.

Qualitative Comparisons

Once we start to categorize and count unstructured text, we can begin to extract information that can be used for qualitative analytics. Qualitative analytics can work with the available data and perform operations based on the characteristics of the data.

If we can classify the data into a set of hierarchies, we can determine whether a particular data belongs to a set or not. This would be considered a nominal analysis. If we have an established hierarchy, we can deduce the set membership for higher levels of the hierarchy.

In ordinal analysis, we can compare two data items. We can deduce whether a data is better, higher, or smaller than another based on comparative algebra available to ordinal analytics. Sentiment analysis is one such comparison. For example, let us take a statement we analyzed from a customer complaint.

> "Before 12 days, I was recharged my Data Card with XXX Plan. But i am still not able to connect via internet. I have made twise complain. But all was in vain. The contact number on Contact Us page is wrong, no one is picking up. I have made call to customer care but every guy telling me..."

As humans, it is obvious to us that the sentiment of this sentence is negative. However, Big Data requires sentiment analysis on terabytes of data, which means we need to assign a positive or negative sentiment using a computer program. Use of words or phrases such as "I am not able," "complain," "in vain," and "no one is picking up" are examples of negative sentiments. A sentiment lexicon can be used as a library to compare words against known "positive" or "negative" sentiment. A count of the number of negative sentiments is qualitative analytics that can be performed on sentiment data, as we can differentiate between positive and negative sentiment and conclude that positive sentiment is better than negative sentiment. We can also create qualifiers such as "strong" sentiment and "weak" sentiment and compare the two sets of comments.

In typical interval scaled data, we can assign relative values to data but may not have a point of origin. As a result, we can compute differences and deduce that the difference between two data items is higher than another set of data items. For example, a strong positive sentiment may be better than weak positive sentiment. However, these two data items are more similar than the pair of a strong positive and a strong negative sentiment.

Focus on Specific Time Slice or Using Other Dimensions

Data Warehouses are at the receiving end of a large number of transactions. The source data is typically created by a series of (mainframe) applications, which are connected together in a food chain where the output of one program became the input of the next. I studied a data accumulation process where the financial organization was the recipient of these cascading transactions, and the organization needed to balance the books in a short time. If the transactions failed in source systems, the financial reports were likely to be delayed. They were tasked with reducing the number of situations that would lead to delays in data collection. The most common denominators were the system logs from all of these systems.

If we know the past failures in this cascading set of transactions, can we use Big Data Analytics to isolate the failure conditions? IBM's Big Insight includes a query tool that allows us to study a slice of data on a chosen dimension such as time. I can set the time for analysis to be the 24 hours preceding a failure and look for system log error conditions, thereby helping to isolate the combination of error conditions that happened together. If we have located a pattern of error conditions leading to a failure, we can use the query tool to check for all the time slices within system failure and look for any systematic failure patterns.

There are many other use cases where this analysis can be applied. For example, a new pricing model may lead to strong negative sentiment. A new feature release may cause disruption in consumer use. A new competitive offering may reduce interest among shoppers. As long as we have time slices with unstructured data representing independent variables and consumer sentiment as a dependent variable, the data can be analyzed to discover causal chains. This is the most powerful aspect of unstructured data analytics.

4.3 Big Data and Single View of Customer/Product

In any enterprise, there are likely to be many views of customers and products. Most of the fragmentation comes from divergent views of customers and products. Customer and product MDM solutions are popular ways of bridging and bringing together a single unified view. However, over the past decade or two, this integration has been focused primarily on intra-organization sources of traditional "structured" data.

Automation and data collection technologies have opened up new sources of data from the product itself, processes supporting the customers, and third parties. For example, the web interface offers a significant amount of information that can be used for additional customer insight. Tealeaf®, a recent IBM acquisition, specializes in improving multi-channel customer experience by analyzing customer behavior across channels and making that information centrally available. This information about customer behavior at the web interface can now be used for a variety of purposes. It can be used by the contact centers to improve their response to the customers, by product management to improve products, and by IT to improve customer touch points.[23] If the customer has logged into the website, the customer identity is known and can be used to connect customer behavior to the rest of the customer MDM. Product usage can also be tracked and collected, summarized, and categorized. For example, call detail records collect calling information for cell phones. These records carry fairly detailed information, such as cell towers used to make the call, which can be used to pinpoint customer location at the time of the call. The CDR data can be used to understand customer behavior. The device ID is often already connected with the rest of the customer MDM, and the behavior collected from the CDR data can augment the rest of the customer MDM.

With the wide availability of external data, we have opened up the customer and product masters to also include external sources, including Twitter, Facebook, Yelp, You Tube, other blogs, and in general any information publicly available. The information published externally could include intent to buy, product preferences, complaints, endorsements, usage and other useful segmentation data. This data can be collected, collated, identified with individual customers or segments, and connected with the rest of the customer view.

How do we merge internal and external views to create what we may call a Big Data view of the customer or product? This integrated view is a far more holistic understanding of the customer or product. By analyzing and integrating this data with the rest of the customer master, we can now do a far more extensive household analysis. This data may reveal that while the cell phone is being paid for by the parent, the actual user resides in a college dormitory in a different city and should not be offered promotions for regional stores in the city where the parents live unless the student is visiting the parents for a fall break. Figure 4.4 shows some sample elements of a Big Data customer master. It includes personal attributes, life events, relationships, timely insights, and product interests. These elements provide an enormous opportunity to marketers to target products and improve customer service.

Product data is equally interesting. In section 3.2, I described Product Knowledge Hub as the driving application, and in section 4.2 I discussed product ontology, which can be used to organize product data. If we were to gather information from outside sources regarding a set of products, a product MDM allows us to organize this data for a variety of users, including call centers and the web.

The central technical capability in any MDM product is its ability to match identities across diverse data sources. How do we integrate Big Data with the matching capabilities of the MDM solution? Most MDM solutions offer powerful matching capabilities for structured data. I can use MDM software to match customers based on fuzzy logic and to create new IDs that combine customer data from a variety of sources. These solutions are also providing significant