Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

UNIT –I

1. Introduction

Big Data Analytics is a popular topic. While everyone has heard stories of new Silicon Valley valuation bubbles and critical shortages of data scientists, there are an equal number of concerns: Will it take away my current investment in Business Intelligence or replace my organization? How do I integrate my Data Warehouse and Business Intelligence with Big Data? How do I get started, so I can show some results? What are the skills required? What happens to data governance? How do we deal with data privacy?

Over the past 9 to 12 months, I have conducted many workshops with practitioners in this field. I am always fascinated with the two views that so often clash in the same room—the bright-eyed explorers ready to share their data and the worriers identifying ways this can lead to trouble. A similar divide exists among consumers. As in any new field, implementation of Big Data requires a delicate balance between the two views and a robust architecture that can accommodate divergent concerns.

Unlike many other Big Data Analytics blogs and books that cover the basics and technological underpinnings, this book takes a practitioner's viewpoint. It identifies the use cases for Big Data Analytics, its engineering components, and how Big Data is integrated with business processes and systems. In doing so, it respects the large investments in Data Warehouse and Business Intelligence and shows both evolutionary and revolutionary—as well as hybrid—ways of moving forward to the brave new world of Big Data. It deliberates on serious topics of data privacy and corporate governance and how we must take care in the implementation of Big Data programs to safeguard our data, our customers' privacy, and our products.

So, what is Big Data? There are two common sources of data grouped under the banner of Big Data. First, we have a fair amount of data within the corporation that, thanks to automation and access, is increasingly shared. This includes emails, mainframe logs, blogs, Adobe PDF documents, business process events, and any other structured, unstructured, or semi-structured data available inside the organization. Second, we are seeing a lot more data outside the organization— some available publicly free of cost, some based on paid subscription, and the rest available selectively for specific business partners or customers. This includes information available on social media sites, product literature freely distributed by competitors, corporate customers' organization hierarchies, helpful hints available from third parties, and customer complaints posted on regulatory sites.

Many organizations are trying to incentivize customers to create new data. For example, Foursquare (*www.foursquare.com*) encourages me to document my visits to a set of businesses advertised through Foursquare. It provides me with points for each visit and rewards me with the "Mayor" title if I am the most frequent visitor to a specific business location. For example, every time I visit Tokyo Joe's—my favorite nearby sushi place—I let Four square know about my visit and collect award points. Presumably, Foursquare, Tokyo Joe's, and all the competing sushi restaurants can use this information to attract my attention at the next meal opportunity.

Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

Sunil Soares has identified five types of Big Data: web and social media, machine-to-machine (M2M), big transaction data, biometrics, and human generated.

Why is Big Data different from any other data that we have dealt with in the past? There are "four V's" that characterize this data: Volume, Velocity, Variety, and Veracity. Some analysts have added other V's to this list, but for the purpose of this book, I will focus on the four V's described here.

1.1 Velocity

There are two aspects to velocity, one representing the throughput of data and the other representing latency. Let us start with throughput, which represents the data moving in the pipes. The amount of global mobile data is growing at a 78 percent compounded growth rate and is expected to reach 10.8 exabytes per month in 20164 as consumers share more pictures and videos. To analyze this data, the corporate analytics infrastructure is seeking bigger pipes and massively parallel processing.

Latency is the other measure of velocity. Analytics used to be a "store and report" environment where reporting typically contained data as of yesterday— popularly represented as "D-1." Now, the analytics is increasingly being embedded in business processes using data-in-motion with reduced latency. For example, Turn (*www.turn.com*) is conducting its analytics in 10 milliseconds to place advertisements in online advertising platforms.

1.2 Variety

In the 1990s, as Data Warehouse technology was rapidly introduced, the initial push was to create meta-models to represent all the data in one standard format. The data was compiled from a variety of sources and transformed using ETL (*E*xtract, *T*ransform, *L*oad) or ELT (*E*xtract the data and *L*oad it in the warehouse, then *T*ransform it inside the warehouse). The basic premise was narrow variety and structured content. Big Data has significantly expanded our horizons, enabled by new data integration and analytics technologies. A number of call center analytics solutions are seeking analysis of call center conversations and their correlation with emails, trouble tickets, and social media blogs. The source data includes unstructured text, sound, and video in addition to structured data. A number of applications are gathering data from emails, documents, or blogs. For example, Slice provides order analytics for online orders (see *www.slice.com* for details). Its raw data comes from parsing emails and looking for information from a variety of organizations—airline tickets, online bookstore purchases, music download receipts, city parking tickets, or anything you can purchase and pay for that hits your email. How do we normalize this information into a product catalog and analyze purchases?

Another example of enabling technology is IBM's InfoSphere Streams platform, which has dealt with a variety of sources for real-time analytics and decision making, including medical instruments for neonatal analysis, seismic data, CDRs, network events, RFID tags, traffic patterns, weather data, mainframe logs, voice in many languages, and video.

1.3 Veracity

Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

Unlike carefully governed internal data, most Big Data comes from sources outside our control and therefore suffers from significant correctness or accuracy problems. Veracity represents both the credibility of the data source as well as the suitability of the data for the target audience.

Let us start with source credibility. If an organization were to collect product information from third parties and offer it to their contact center employees to support customer queries, the data would have to be screened for source accuracy and credibility. Otherwise, the contact centers could end up recommending competitive offers that might marginalize offerings and reduce revenue opportunities. A lot of social media responses to campaigns could be coming from a small number of disgruntled past employees or persons employed by competition to post negative comments. For example, we assume that "like" on a product signifies satisfied customers. What if the "like" was placed by a third party?

We must also think about audience suitability and how much truth can be shared with a specific audience. The veracity of data created within an organization can be assumed to be at least well intentioned. However, some of the internal data may not be available for wider communication. For example, if customer service has provided inputs to engineering on product shortcomings as seen at the customer touch points, this data should be shared selectively, on a need-to-know basis. Other data may be shared only with customers who have valid contracts or other prerequisites.

Over the past year, the Information Agenda team has been asked to conduct a number of Big Data Analytics workshops. The three most common questions have been as follows:

1. What is Big Data and what are others doing with it?

2. How do we build a strategic plan for Big Data Analytics in response to a management request?

3. How does Big Data change our analytics organization and architecture?

Most of the material included in this book was collated in response to answering these questions.

This provides three perspectives on Big Data Analytics.

First, why is Big Data Analytics becoming so important, and what can we do with it? The book projects major trends behind the rise of Big Data and shows typical use cases tackled by Big Data Analytics, where leading organizations are already seeing major benefits. Second, the book lists major components of Big Data Analytics and introduces an integrated architecture — Advanced Analytics Platform (AAP) — that combines Big Data Analytics with the rest of the analytics infrastructures and integrates with business processes. It shows how these components work together in the AAP to provide an integrated engine that can combine Big Data with traditional Data Warehouse and Business Intelligence to provide an overall solution.

Third, the book provides a glimpse at implementation concerns and how they must be tackled. How do we establish a roadmap and implement key pilot programs to gather momentum and persist to create a game-changing vision?

How do we provide governance across this data when the originating data may have varying quality or privacy constraints?

The big elephant in the room is data privacy. I confess I have not taken a position on data privacy, nor have I predicted how the world will deal with it. It is an evolving topic, with many complications, geographical differences, and unknown consequences. However, I have outlined a number of critical areas to probe further, as well as a number of required components, irrespective of the position taken.

1.4. Drivers for Big Data

We are increasing the pace for Big Data creation. This chapter examines the forces behind this tsunami of Big Data. There are three contributing factors: consumers, automation, and monetization. More than each of these contributing factors, their interaction is speeding the creation of Big Data. With increasing automation, it is easier to offer Big Data creation and consumption opportunities to the consumers and the monetization process is increasingly providing an efficient marketplace for Big Data.

1.5 Sophisticated Consumers

The increase in information level and the associated tools has created a new breed of sophisticated consumers. These consumers are far more analytic, far savvier at using statistics, and far more connected, using social media to rapidly collect and collate opinion from others. We live in a world full of marketing messages. While most of the marketing is still broadcast using newspaper, magazine, network TV, radio, and display advertising, even in the conventional media, narrow casting is gradually becoming more prominent. This is seen in local advertisement insertions in magazines, insertion of narrow cast commercials using set-top boxes, and use of commuter information to change street display ads. The Internet world can become highly personalized. Search engines, social network sites, and electronic yellow pages insert advertisements specific to an individual or to a micro-segment. Internet cookies are increasingly used to track user behavior and to tailor content based on this behavior.

Email and text messages rapidly led toward increased interpersonal inter-actions. Communication started not only with marketers but also with third parties and friends. Communication expanded to bulletin boards, group chats, and social media, allowing us to converse about our purchase intentions, fears, expectations, and disappointments with small and large social groups. Unlike email and text, the conversations are on the Web for others to read, either now or later.

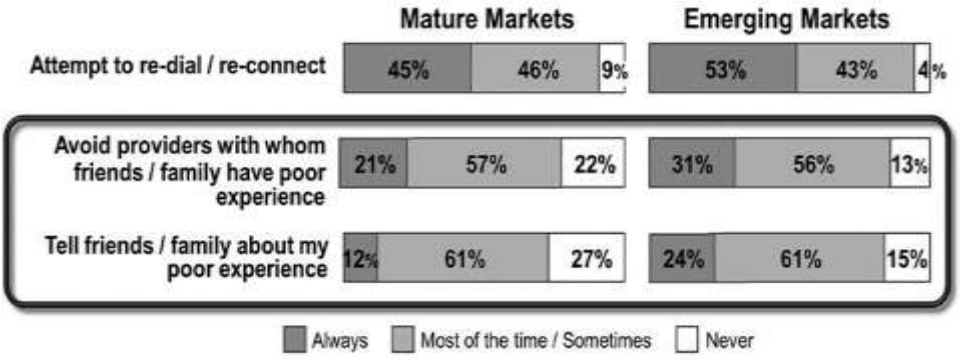
Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

So far, we have been dealing only with single forms of communication. The next sets of sources combine information from more than one media. For example, Facebook conversations involve a number of media, including text, sound clips, photos, and video. Second world and alternate reality are becoming interesting avenues for trying out product ideas in a simulated world where product usage can be experimented with.

We often need experts to help us sort out product features and how they relate to our product usage. A large variety of experts are available today to help us with usage, quality, pricing, and value-related information about products.

A number of marketers are encouraging advisor or ambassador programs using social media sites. These selected customers get a preview of new products and actively participate in evaluating and promoting new products. At the end of the day, people we know and trust sway our decisions. This is the biggest contribution of social networks. They have brought consumers together such that sharing customer experiences is now far more frequent than ever before.

How would a consumer deal with a poor service quality experience? Figure 2.1 shows typical behaviors in mature and emerging markets as studied by an IBM Global Telecom Consumer Survey conducted with a sample size of 10,177.7. In this survey, 78 percent of the consumers surveyed in the mature markets said they avoid providers with whom friends or family had bad experience. The percentage was even higher (87 percent) in growth markets. In response to a related question, survey participants said that they inform friends and family about poor experience (73 percent in mature markets and 85 percent in growth markets). These numbers together show a strong influence of social network on purchase behavior. These are highly significant percentages and are now increasingly augmented by social media sites (e.g., the "Like" button placed on Facebook). The same survey also found that the three most preferred sources for recommendation information are Internet, recommendations from family/friends, and social media.



Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

Source: 2011 IBM Global Telecom Consumer Survey, Global N = 10177; Mature Countries N = 7875

Fig.1: Behaviors in response to poor service quality experience

In any group, there are leaders. These are the people who lead a change from one brand to another. Leaders typically have a set of followers. Once a leader switches a brand, it increases the likelihood for the social group members to churn as well. Who are these leaders? Can we identify them? How can we direct our marketing to these leaders?

In any communication, the leaders are always the center of the hub (see Fig.1.1). They are often connected to a larger number of "followers," some of whom could also be leaders. In the figure, the leaders have a lot more communication arrows either originating or terminating to them compared with others.

How do we identify the leaders? IBM Research conducted a series of experiments with SPs.8 Call detail records, which carry information about

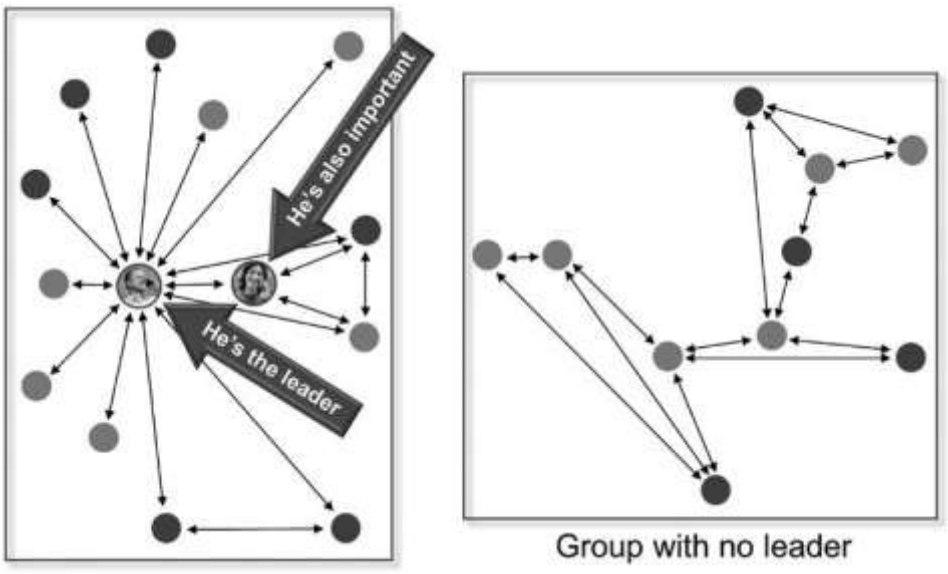


Fig.1.1: Leaders in a communications network

person A calling person B, were analyzed. By synthesizing call information and abstracting communications networks, we discovered webs of communications across individuals. We also used the customer churn information to correlate churn among leaders to subsequent churn among followers. Here are some of the highlights from one of the experiments I helped conduct:

- Leaders were 1.2 times more likely to churn compared with non-leaders.
- There were two types of leaders: disseminating leaders who were connected to their group through outgoing calls, and authority leaders who were connected through a larger proportion of incoming calls.
- When a disseminating leader churned, additional churns were 28.5 times more likely. When an authority leader churned, additional churns were 19.9 times more likely.
- Typically, there was a very limited time between leaders' churn and the followers' churn.

Social groups can be inferred from any type of communication—emails, SMS texts, calls, Facebook friendships, and so on. It is interesting to see strong statistics associated with leaders' influence on the group.

There are many ways to utilize social networks to influence purchase and reuse:

- *Studying consumer experience* — A fair amount of this data is unstructured. By analyzing the text for sentiments, intensity, readership, related blogs, referrals, and other information, we can organize the data into positive and negative influences and their impact on the customer base.
- *Organizing customer experience* — We can provide reviews to a prospective buyer, so they can gauge how others evaluated the product.
- *Influencing social networks* — We can provide marketing material, product changes, company directions, and celebrity endorsements to social networks, so that social media may influence and enhance the buzz.
- *Feedback to products, operations, or marketing* — By using information generated by social media, we can rapidly make changes in the product mix and marketing to improve the offering to customers.

Society has always played a major role in our evaluation process. However, the Internet and social networking have radically altered our access to information. I may choose to "like" a product on Facebook, and my network now has instant access to this action. If I consider a restaurant worth its money, Yelp can help me broadcast that fact worldwide. If I hate the new cell phone service from a CSP, I can blog to complain about it to everyone.

1.6 Automation

Interactive Voice Response (IVR), kiosks, mobile devices, email, chat, corporate Websites, third-party applications, and social networks have generated a fair amount of event information about the customers. In addition, customer interactions via traditional media such

as call centers can now be analyzed and organized. The biggest change is in our ability to modify the customer experience using software policies, procedures, and personalization, making self-service increasingly customer friendly.

Sales and marketing have received their biggest boost in instrumentation from Internet driven automation over the past 10 years. Browsing, shopping, ordering, and customer service on the Web not only has provided tremendous control to users but also has created an enormous flood of information to the marketing, product, and sales organization in understanding buyer behavior.

Each sequence of Web clicks can be collected, collated, and analyzed for customer delight, puzzlement, dysphoria, or outright defection. More information can also be obtained about sequence leading up to a decision.

Self-service has crept in through a variety of means: IVRs, kiosks, handheld devices, and many others. Each of these electronic means of communication acts like a gigantic pool of time-and-motion studies. We have data available on how many steps customers took, how many products they compared, and what attributes they focused on, such as price, features, brand comparisons, recommendations, defects, and so on. Suppliers have gained enormous amounts of data from self-service and electronic sensors connected to products. If I use a two-way set-top box to watch television, the supplier has instant access to my channel-surfing behavior. Did I change the channel when an advertisement started? Did I turn the volume up or down when the jingle started to play? If I use the Inter-net to shop for a product, my click stream can be analyzed and used to study shopping behavior. How many products did I look at? Did I view the product description or the price when looking at the product? This enriched set of data allows us to analyze customer experience in the minutest detail.

What are the sources of data from such self-service interactions?

- *Product* — As products become increasingly electronic, they provide a lot of valuable data to the supplier regarding product use and product quality. In many cases, suppliers can also collect information about the context in which a product was used. Products can also supply information related to frequency of use, interruptions, usage skipping, and other related aspects.
- *Electronic touch points* — A fair amount of data can be collected from the touch points used for product shopping, purchase, use, or payment. IVR tree traversals can be logged, Web click streams can be collected, and so on.
- *Components* — Sometimes, components may provide additional information. This information could include data about component failures, use, or lack thereof. For example, a wireless CSP can collect data from networks, cell towers, third parties,

and handheld devices to understand how all the components together provided a good or bad service to the customer.

1.7 Monetization

From a Big Data Analytics perspective, a "data bazaar" is the biggest enabler to create an external marketplace, where we collect, exchange, and sell customer information. We are seeing a new trend in the marketplace, in which customer experience from one industry is anonymized, packaged, and sold to other industries. Fortunately for us, Internet advertising came to our rescue in providing an incentive to customers through free services and across the board options.

Internet advertising is a remarkably complex field. With over $26 billion in 2010 revenue, the industry is feeding a fair amount of startup and initial public offering (IPO) activity. What is interesting is that this advertising money is enhancing customer experience. Take the case of Yelp, which lets consumers share their experiences regarding restaurants, shopping, nightlife, beauty spas, active life, coffee and tea, and others. Yelp obtains its revenues through advertising on its website; however, most of the traffic is from people who access Yelp to read customer experience posted by others. With all this traffic coming to the Internet, the questions that arise are how is this Internet usage experience captured and packaged and how are advertisements traded among advertisers and publishers.

Big Data Analytics is creating a new market, where customer data from one industry can be collected, categorized, anonymized, and repackaged for sale to others:

- *Location* — As we discussed earlier, location is increasingly available to suppliers. Assuming a product is consumed in conjunction with a mobile device, the location of the consumer becomes an important piece of information that may be available to the supplier.
- *Cookies* — Web browsers carry enormous information using web cookies. Some of this may be directly associated with touch points.
- *Usage data* — A number of data providers have started to collect, synthesize, categorize, and package information for reuse. This includes credit-rating agencies that rate consumers, social networks with blogs published or "Like" clicked, and cable companies with audience information. Some of this data may be available only in summary form or anonymized for the protection of customer privacy.

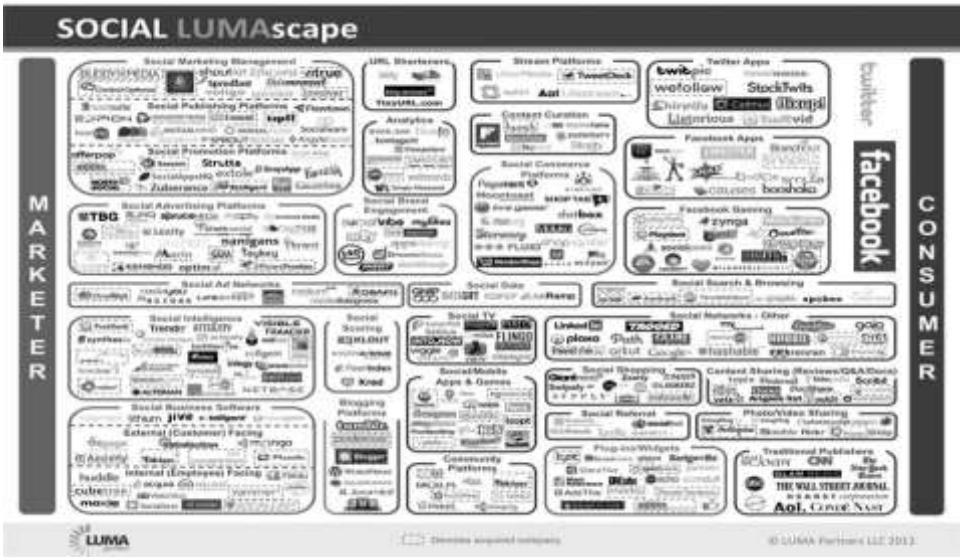


Fig. 1.2 LUMA Scape for social media (reprinted with permission)

Terence Kawaja has been studying this market for a number of years and has characterized a number of markets and associated players. "Terence Kawaja has a new way for potential investors to visualize it," says *Wall Street Journal* writer Amir Efrati. "The market involves hundreds of small and large companies that help advertisers reach consumers and help website publishers, mobile application developers, search engines, and other digital destinations generate revenue through advertising. Kawaja, who runs the investment firm LUMA Partners, spent months putting together six new graphics that show how 1,240 different companies fit into the following categories of online advertising: display, video, search engines, mobile, social, and commerce."11 I have replicated Kawaja's Social Media LUMA Scape in Figure 1.2. For the rest of the LUMA Scapes, visit Kawaja's website: *www.lumapartners.com*. A number of intermediaries play key roles in developing an advertising inventory, auctioning of the inventory to the ad servers, and facilitating the related payment process, as the advertisements are clicked and related buying decisions are tracked.

1.8 Big Data Analytics Applications

This chapter discusses a number of important use cases for Big Data Analytics. In each case, Big Data Analytics is becoming integrated with business processes and traditional analytics to provide major outcomes. In many cases, these use cases represent game changers essential to the survival and growth of an organization in an increasingly competitive marketplace. Some of these use cases are still in their infancy, while others are becoming increasingly commonplace.

1.9 Social Media Command Center

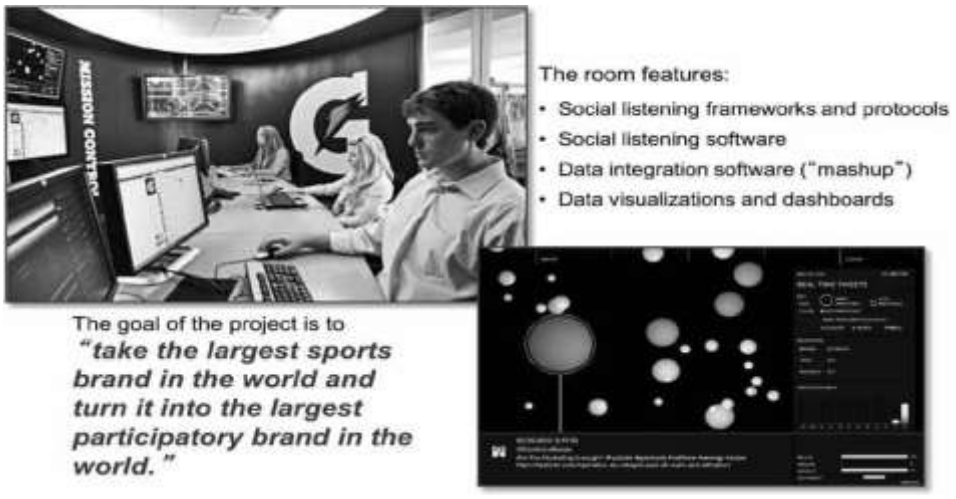
Last year, Blackberry faced a serious outage when its email servers were down for more than a day. I tried powering my Blackberry off and on because I wasn't sure whether it was my device or the CSP. It never occurred to me that the outage could be at the Blackberry server

itself. When I called the CSP, they were not aware of the problem. For a while, I was okay without receiving any emails, but then I started to become curious. So I turned to one obvious source: Twitter. Sure enough, I found information about the Blackberry outage on Twitter.

One of my clients told me that his VP of Customer Service is glued to Twitter looking for customer service problems. Often, someone discovers the problem on Twitter before the internal monitoring organization. We found that a large number of junior staffers employed by marketing, customer service, and public relations search through social media for relevant information. Does this sound like an automation opportunity?

A *Social Media Command Center* combines automated search and display of consumer feedback expressed publicly on the social media. Often, the feedback is summarized in the form of "positive" or "negative" sentiment. Once the feedback is obtained, the marketer can respond to specific comments by entering into a conversation with the affected consumers, whether to respond to questions about an outage or obtain feedback about a new product offering.

The marketing organization for Gatorade, a sports drink product, decided to create a Social Media Command Center to increase consumer dialog with Gatorade.12 Figure 3.1 shows the monitoring station with the dashboard. Big Data Analytics can be used to monitor social media for feedback on product, price, and promotions as well as to automate the actions taken in response to the feedback. This may require communication with a number of internal organizations, tracking a product or service problem, and dialog with customers as the feedback results in product or service changes. When consumers provide feedback, the dialog can only be created if the responses are provided in low latency. The automated solutions are far better at systematically finding the information, categorizing it based on available attributes, organizing it into a dashboard, and orchestrating a response at conversation speed.



Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

Fig. 1.3 Gatorade Social Media Command Center

1.10 Product Knowledge Hub

As consumers turn into sophisticated users of technology and the marketplace becomes specialized, the product knowledge seldom belongs to one organization. Take the Apple iPhone as an example. The iPhone is marketed by Apple, but its parts came from a large supply chain pool, the apps running on the iPhone come from a large community of app developers, and the communications service is provided by a CSP. Google's Android is even more diverse, as Google provides the operating system while a cell phone manufacturer makes the device. The smartphones do not work in isolation. They act as WiFi hubs for other devices. So, what happens if I want to know how to tether an iPhone to an Apple iPad? Do I call my CSP, or do I call Apple? Would either of their websites give me a simple step-by-step process I can follow?

Every time I get into these technical questions about products I am trying to use, I end up calling my son, who happens to know the answers to any such question. Recently, he decided to educate me on how he finds the answer, and so I was introduced to a myriad of third-party sites where a variety of solutions can be found. In most cases, we can find them by searching using any popular search engine. However, the solutions do not always favor the CSPs, and they are often dated, failing to take into account the latest offerings. Between the device operating system, the offerings from CSPs, and the apps, one must tread carefully through the versions to make sure the solution we discover is for the same version of software that is on the device. So now, we are facing data that is characterized by both variety and veracity. Can we use Big Data Analytics to solve this problem?

The solution involves three sets of technologies. Fortunately, Vivisimo has packaged these technologies into its Velocity product, making it easier to obtain an integrated solution. The first part of the solution is the capability to tap any sources of data. A CSP may already have pieces of the solution on its intranet, put together by product managers or customer service subject matter experts. Or, the information may reside on a device manufacturer site or a third-party site. All this data must be pulled and stripped of its control information so that the raw text is available to be reused.

The second part of the solution is to create a set of indices so that the raw information can be categorized and found when needed. Because many combinations of products exist, we would like to collect and combine information for the devices searched. The federated indexing system lets us organize the information for easy access.

Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

Fig. 1.4. Product Knowledge Hub for a CSP

What we have created is a *knowledge hub*, which can now be used directly from a website or made available to the call centers. It significantly reduces call-handling time in the call centers and also increases first call resolution. By placing the information on the web, we are now promoting the CSP's website as the source of knowledge, which increases web traffic and reduces the number of people who resort to contacting the call center. Figure 3.2 depicts the Product Knowledge Hub.

Once we have created a single source of knowledge, this source can be used to upsell other products, connecting usage knowledge to product features and using the knowledge pool to discover new product or business partnership ideas. A lot of stray, fragmented knowledge about the products may be rapidly organized and find a variety of other uses.

1.11 Infrastructure and Operations Studies

A number of industries are exploring the use of Big Data to improve their infra-structure. In many situations, the best way to improve the infrastructure is to understand its use and how bottlenecks or configurations impact performance. In the past, this data required extensive manual data collection costs. Big Data provides a natural source of data with minimal data collection costs. I will lay out examples from public services to illustrate this point.

The city of Boston decided to use Big Data to identify potholes in the streets by sponsoring a competition in the analyst community. A winner came from Sprout & Co., a nonprofit group in Somerville, Massachusetts. The solution included the use of magnitude-of-acceleration spikes along a cell phone's z-axis to spot impacts, plus additional filters to distinguish potholes from other irregularities on the road. The new algorithm made Street Bump, a free download in Apple's App Store, a winner. This analysis can save significant road survey cost. Navigation systems can also use the cell phone data to avoid traffic congestion and offer alternate routes. This type of use of Big Data is one of the best ways to gain acceptance without getting into privacy or security issues.

Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

In another example, city bus and train agencies are making their real-time transit information available to riders. This information significantly improves the user experience and reduces the uncertainty associated with both planned and unexpected delays. Transloc (*www.transloc.com*) provides this information for riders using a variety of technologies, including smartphones, web, and SMS messages. It also provides prediction capabilities on expected arrival time. Once the app is loaded on a smartphone, the rider can use it to accurately estimate travel time and also review the travel route.

IBM's Smarter Cities® initiative is using Big Data in a number of applications directed at city infrastructure and operations. Location data from cell phones can be used to provide raw material for detecting traffic patterns. These patterns can then be used to decide on new transportation projects, to change controls, or to redirect traffic in case of an emergency.

Another important application for Big Data Analytics is public safety. The New York Police Department is using Big Data for crime prevention.

1.12 Product Selection, Design, and Engineering

Product automation provides an enormous opportunity to measure customer experience. We take photos digitally and then post them on Facebook, providing an opportunity for face recognition without requiring laborious cycles in digitization. We listen to songs on Pandora, creating an opportunity to measure what we like or dislike or how often we skip a song after listening to the part of it that we like the most. We read books electronically online or on our favorite handheld devices, giving publishers an opportunity to understand what we read, how many times we read it, and which parts we look at. We watch television using a two-way set-top box that can record each channel click and correlate it to analyze whether the channel was switched right before, during, or after a commercial break. Even mechanical products such as automobiles are increasing electronic interactions. We make all of our ordering transactions electronically, giving third parties opportunities to analyze our spending habits by month, by season, by ZIP+4, and by tens of thousands of micro-segments. Usage data can be synthesized to study the quality of customer experience and can be mined for component defects, successes, or extensions. Marketing analysts can identify micro-segmentations using this data. For example, in a wireless company, we isolated problems in the use of cell phones to defective device antenna by analyzing call quality and comparing it across devices.

Products can be test marketed and changed based on feedback. They can also be customized and personalized for every consumer or micro-segment based on their needs. Analytics plays a major role in customizing, personalizing, and changing products based on customer feedback. Product engineering combines a set of independent components into a product in response to a customer need. Component quality impacts overall product performance. Can we use analytics to isolate poorly performing components and replace them with good ones? In addition, can we simplify the overall product by removing components that are rarely used and offer no real value to the customer? A lot of product engineering

analytics using customer experience data can lead to building simplified products that best meet customer requirements.

To conduct this analysis and predictive modeling, we need a good under-standing of the components used and how they participate in the customer experience. Once a good amount of data is collected, the model can be used to isolate badly performing components by isolating the observations from custom-er experience and tracing them to the poorly performing component. Complex products, such as automobiles, telecommunications networks, and engineering goods, benefit from this type of analytics around product engineering.

The first level of analysis is in identifying a product portfolio mix and its success with the customers. For example, if a marketer has a large number of products, these products can be aligned to customer segments and their usage.

We may find a number of products that were purchased and hardly used, leading to their discontinuation in six months, while other products were heavily used and sparingly discontinued.

Once we have identified less-used products, the next analysis question is whether we can isolate the cause of customer disinterest. By analyzing usage patterns, we can differentiate between successful products and unsuccessful ones. Were the unsuccessful ones never launched? Did many users get stuck with the initial security screen? Maybe the identification process was too cumbersome. How many users could use the product to perform basic functions offered by the product? What were the highest frequency functions?

The next level of analysis is to understand component failures. How many times did the product fail to perform? Where were the failures most likely? What led to the failure? What did the user do after the failure? Can we isolate the component, replace it, and repair the product online?

These analysis capabilities can now be combined with product changes to create a sophisticated test-marketing framework. We can make changes to the product, try the modified product on a test market, observe the impact, and, after repeated adjustments, offer the altered product to the marketplace.

Let us illustrate how Big Data is shaping improved product engineering and operations at the communications service providers. Major CSPs collect enormous amounts of data about the network, including network transport information coming from the routers and the switches, as well as usage information, popularly known as call detail records (CDRs), which are recorded each time we use telephones to connect with one another. As the CSP networks

grew in sophistication, the CDRs were extended to data and video signals using IPDRs. Most CSPs refer to this usage information as xDRs (where *x* is now a variable that can be substituted for "any" usage information). For larger CSPs, the usage statistics not only are high volume (in billions of transactions a day) but also require low-latency analytics for a number of applications. For example, detecting a fraudulent transaction or abusive network user in the middle of a video download or call may be more valuable than finding out this information the next day. In addition, it is always a strategic driver for CSPs to lay out all the network and usage information on their network topology and geography and use a variety of automated analytics and manual visualization techniques to connect the dots between network trouble or inefficiencies and usage. The analytics provides CSP with a valuable capability to improve the quality of the communication. If every user call is dropping in a particular area that is a popular location for premier customers, it could lead to churn of those customers to competitors.

The information about xDRs, network events, customer trouble tickets, blogs, and tweets in the social media can be correlated for a variety of business purposes. CSPs have used this analytics to detect spots with poor network performance to reorganize towers and boosters. The differences in usage can be analyzed to detect device problems such as faulty antennas on specific models. The variations can also be analyzed to find and fix network policies or routing problems. As CSPs race to implement high-volume, low-latency xDR hubs, they are finding plenty of business incentives to fund these programs and reap benefits in the form of improved product offerings to their customers.

1.13 Location-Based Services

A variety of industries have location information about their customers. Cell phone operators know customer location through the location of the phones. Credit-card companies know the location of transactions, and auto manufacturers the location of cars, while social media is trying its best to get customers to disclose their location to their friends and family. On a recent short trip to India, I decided to use Endomondo, an app on my cell phone to record my jogging activity in Mumbai, India, which was instantly posted on my Facebook page, thereby letting my friends know of my visit to Mumbai.

Let us take a wireless CSP example to study how we collect and summarize location information. A cell phone is served by a collection of cell phone towers, and its specific location can be inferred by triangulating its distance from the nearest cell towers. In addition, most smartphones can provide GPS location information that is more accurate (up to about 1 meter). The location data includes longitude and latitude and, if properly stored, could take about 26 bytes of information. If we are dealing with 50 million subscribers and would like to store 24 hours of location information at the frequency of once a minute, the data stored is about 2 terabytes of information per day. This is the amount of information stored in the location servers at a typical CSP.

Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

Customer locations can be summarized into “hang outs” at different levels of granularity. The location information can be aggregated into geohashes that draw geo boundaries and transform latitude-longitude data into geohash so that it can be counted and statistically analyzed. The presence of a person in a specific location for a certain duration is considered a space-time box and can be used to encode the hang out of an individual in a specific business or residential location for a specific time period.

Many of our smartphone apps collect location data, provided a subscriber “opts-in.”15 If a marketer is interested in increasing the traffic to a grocery store that is located in a specific geohash, they can run an effective marketing campaign by analyzing and understanding which neighborhood people are more likely to hang out or shop in that specific grocery store. Instead of blasting a promotion to all neighborhoods, the communication can now be directed to specific neighborhoods, thereby increasing the efficiency of the marketing campaign. This analysis can possibly be conducted using 6-byte location geohash over a span of one hour and finding all the cell phones that have visited the grocery store regularly. A predictive model can compute the probability of a customer visiting the grocery store based on their past hang out history, and customer residence information can be clustered to identify neighborhoods most likely to visit the shopping center.

Analysis of machine-to -machine transaction data using Big Data technolo-gies is revolutionizing how location -based services can be personalized and offered at low latency. Consider the example of Shopkick, a retail campaign tool that can be downloaded on a smartphone. Shopkick seeks and uses location data to offer campaigns. Once the app is downloaded, Shopkick seeks permission to use current location as recorded by the smartphone. In addition, Shopkick has a database of retailers and their geo-locations. It runs campaigns on behalf of the merchants and collects its revenues from merchants. Shopkick will let me know, for example, that the department store in my neighborhood would like me to visit the store. As a further incentive, Shopkick will deposit shopping points in my account for just visiting the store. As I walk through the store, Shopkick can use my current location in the smartphone to record my presence at the store and award points.

Jeff Jonas provided me tremendous motivation for playing with location data. I used *openpaths.cc*, a site that tracks cell phone location, to track my where abouts for approximately three months. Watching my movements over these months was like having a video unfold my activities event by event. I could also see how I could improve the accuracy of the location data collected by openpaths with other known information such as street maps. With the help of a business directory, it is easy to find out the number and duration of my trips to Starbucks, Tokyo Joe’s, and Sweet Tomato, my three most common eating hang outs.

Why would a customer “opt-in”? Device makers, CSPs, and retailers are beginning to offer a number of location-based services, in exchange for location “opt-in.” For example,

smartphones offer "find my phone" services, which can locate a phone. If the phone is lost, the last known location can be ascertained via a website. In exchange, the CSP or the device manufacturer may seek location data for product or service improvement. These location-based services could also be revenue generating. A CSP may decide to charge for a configuration service that switches a smartphone to silent mode every time the subscriber enters the movie theater and switches back to normal ring tone once the subscriber leaves the movie theater. Prepaid wireless providers are engaging in location-based campaigns targeted at customers who are about to run out of prepaid minutes. These customers are the most likely to churn to a competitor and could easily continue with their current wireless provider if they were to be directed to a store that sells prepaid wireless cards.

These scenarios raise the obvious data privacy concern, which is a hotly debated topic worldwide. We will spend some time in the technical sections talking about data privacy, governance, and how consumer data can be protected and used only as permitted by the customer. As expected, there are many avenues for abuse of customer data, and data privacy must be engrained in the architecture for an effective protection of customer data.

1.14 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 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.

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 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.

1.15 Improved Risk Management

Prepared By,
Dr. S. Rakoth Kandan
Prof / CSE.

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. How-ever, 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,