Applications of Big Data

Hareesh Boinepelli

Abstract Over the last few decades, big business houses in various disparate areas have been accumulating data from different departments in various formats and have been struggling to correlate the datasets and make any valuable business decisions. The key stumbling block has been the inability of the available systems to process large data when the data are part structured and part unstructured. As witnessed in the previous chapters, the technology strides made over the last few years have broken the stigma of processing large datasets and have enabled mining and analysis of large data. Corporations in the data warehousing space have seen this trend as the next big opportunity to help their clients mine their historical data and help further their businesses in terms of adding strategic and tactical value based on the insights gained from their accumulated data over decades. In this chapter, we will see typical examples of how different businesses analyze their data and enhance their business objectives. We will present some examples in the fields of financial services, retail, manufacturing, telecommunications, social media, and health care.

Keywords Basket analysis · Fraud detection · Customer churn · Path analysis · Predictive analysis · Sentiment analysis · Social networking analysis · Sessionization · Graph analysis · Data visualization · K-means clustering

1 Introduction

All the major corporations face a highly competitive environment with constant pressure to increase the profitability by identifying avenues for operational efficiencies and at the same time keeping the business risk to a minimum. All of the big businesses have realized the importance of analyzing loads of historical data that

H. Boinepelli (✉)
Teradata India Pvt. Ltd., Hyderabad, India
e-mail: hareeshkb@gmail.com

H. Mohanty et al. (eds.), *Big Data*, Studies in Big Data 11,
DOI 10.1007/978-81-322-2494-5_7

they have collected over the years, and analysis of this data has become an integral part of making strategic business decisions for these corporations. There is a big push to setup an integrated data management systems and utilize the business intelligence and analytic techniques for improving their businesses.

Over the recent past, big data analytics has found its ways into multiple applications in diverse areas with widespread interest from both the academia and industry. Although this area has made significant progress over the last decade or so, many more challenging problems still exist and finding avenues to new and complex problems in this growing market is ongoing. Various techniques in modeling, statistical analysis, data mining, and machine learning are used to forecast the future events and predict customer behaviors and then proactively act on them to safeguard and enhance business objectives.

In the following sections, we will give a high-level overview of the challenges faced by different industries and how big data are being used to solve them in their respective market segments. Even though big data analytics has the potential in multiple industry domains, we will restrict ourselves to only a few, namely banking and finance (Sect. 2), retail (Sect. 3), manufacturing (Sect. 4), telecommunications (Sect. 5), social media (Sect. 6), and health care (Sect. 7).

2 Big Data Reference Architecture

Figure 1 shows the high-level architecture framework [1] of a typical big data analytics system that includes the following components

1. Acquisition of data from various sources,
2. Infrastructure to do data transformations,
3. Store the data into multiple repositories,
4. Running through high-performance analytic engines, and
5. Reporting and visualization toolset.

Sources of data could be from operational systems which have a nice structure to it (schema/tables/columns/etc.) or can be unstructured such as social media data, click stream data, event logs, and multimedia data. Most of the structured data are stored in the traditional data warehousing environments and the semi-structured and non-structured data on Hadoop clusters. Data are distributed to downstream systems, such as data marts and analytic engines of various types, where the end users can query using SQL-based reporting and analysis tools. Depending on the application, various analytic techniques such as correlation analysis, pattern and trend analysis, collaborating filtering, time series analysis, graph analysis, path analysis, and text analysis are performed before presenting the data on the dashboards using various visualization techniques. In-depth coverage of these components is presented in the previous chapters.

Teradata and IBM are two of the many vendor companies that provide solutions based on the above reference architecture. Figure 2 shows the big data analytics

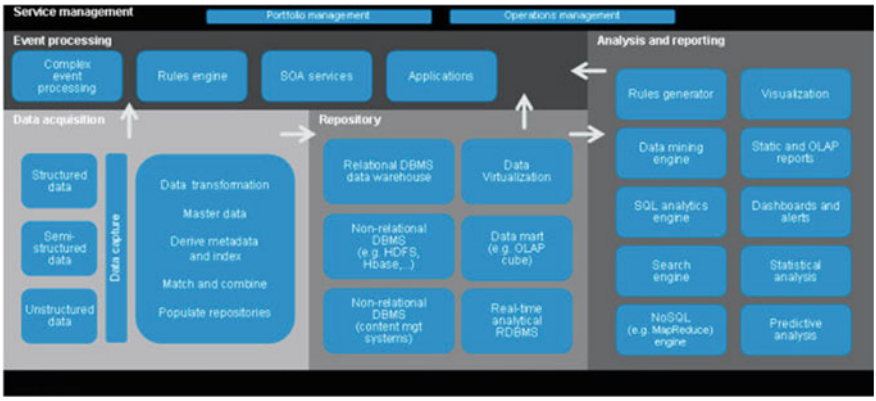

Fig. 1 Big data infrastructure architecture

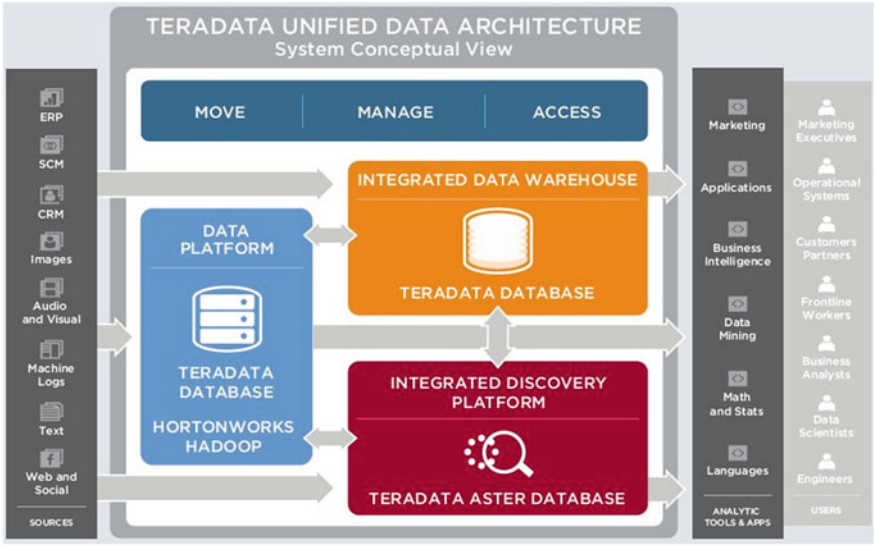

Fig. 2 Teradata unified data architecture

platform from Teradata called the Unified Data Architecture platform [2] with the capabilities to

1. Capture and transform data from variety of sources that are structured, semi-structured, or unstructured data.
2. Ability to process huge data volumes through the use of Hadoop with the data discovery and integrated data warehouse.

3. Support for a number of prepackaged analytic functions in categories such as path analysis, cluster analysis, statistical analysis, predictive analysis, text analysis, relational analysis, and graph analysis.
4. High scalability and performance.

More details on the big data analytics solution can be gathered from the Ref. [2] provided.

Although the reference architecture in Fig. 1 captures the complete set of capabilities required for any big data application, it needs to be noted that not all subsystems represented are required for every application. In the following sections, we will present the frameworks and its components for industry-specific application.

3 Applications in Banking and Financial Industries

Massive amounts of data are being generated by the banking and financial industries through their various service offerings such as checking/savings accounts, mobile banking, credit and debit cards, loans, insurance, and investment services. Most of these data are structured data. Also, most of these organizations have set up their presence online for better serviceability and marketing through which lots of data are collected. As indicated in Fig. 3, some of the channels include

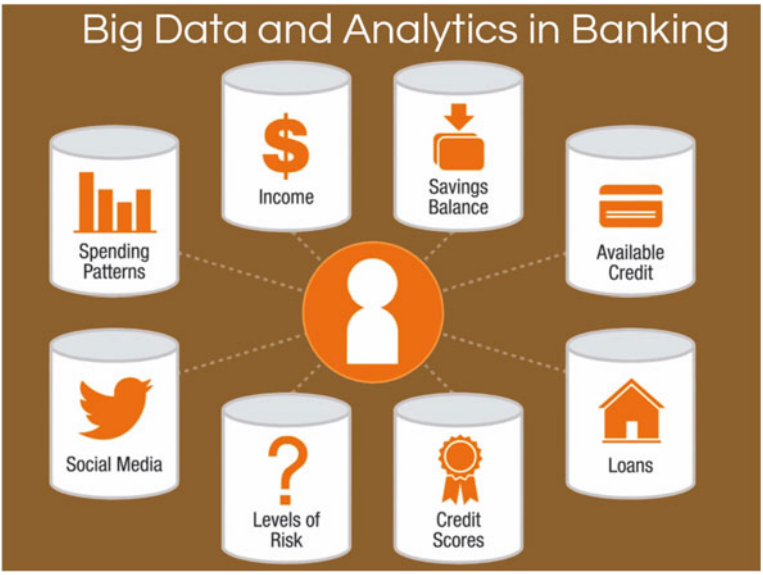

Fig. 3 Big data analytics in banking industry [3]

- Customer interaction through e-mails and chat logs;
- Social networks through tweets and Facebook feeds/posts; and
- Semi-structured data through Web logs and customer reviews.

Most of the data collected are unused, and the industry is looking to various new technologies in data mining and business analytics to help understand and identify customer needs and offer new services which will enhance their business opportunities and increase their margins and profitability. The industry is also looking for solutions in risk management and fraud detection which will help minimize the business exposure. Another area of interest for the industry is on the strategies of retaining customers.

In the following sections, we will cover how big data analytics is applied to the few of the most important areas in more detail.

3.1 Fraud Detection

Various surveys and studies [4] indicate that banking and financial services industry is the victim of the most of the fraud cases among various industries. Following are some of the widely known frauds in the banking industry:

1. Online Banking Fraud: Involves fraudsters taking over the access to the victim's account and performing transactions which siphon the funds out of the accounts.
2. Card Fraud: Involves fraudsters stealing the card information and transact fraudulent transactions.
3. Insider Fraud: Involves fraud by bank's employees.
4. Money Laundering: Crime involving transactions with mainly foreign banks to conceal the origins of illegally obtained wealth.

The traditional approach of sifting through the reports manually and applying various rules is only useful for compliance process and not for detecting fraud and stopping losses. The financial industry requires real-time fraud detection to effectively identify the fraudulent transactions real time and stop them from executing [5].

The key element of fraud detection is the use of analytics to detect patterns of fraudulent behavior. This requires clear understanding of the customer's past behavior in terms of the nature of the transactions so that the distinction of fraudulent versus non-fraudulent transactions can be made effectively by analyzing the transaction against the customer profile which may also include a risk score. This process of scoring the transactions needs to account for the unpredictable nature of transaction activity with varied customer base which includes normal customers and criminals.

Hence, the fraud detection involves a 2-step process which includes

1. Creating the customer profiles based on the historical transactions and identifying the patterns of transactions that lead to fraud
2. Using these customer profiles to catch any outliers or map the transaction sequences/events to the pre-defined fraud patterns and flag any probable fraudulent transactions

Building of customer profiles involves usage of statistical techniques through calculation of statistical averages, min/max values, standard deviations, etc., on the historical transactions to capture the typical transactions mix. Another aspect of the customer profile is capturing the relationships with whom the transactions take place. Graphing techniques [6] are used to capture the network of relationships by mapping transactions between customers along with the modes of payment. Figure 4 captures the flow of creating these customer patterns and profiles.

Figure 5 presents the flow of real-time fraud detection and isolation during the execution of a transaction. If the submitted transaction does not fit the profile of the customer in terms of the transaction amount, transaction connection, etc., it is flagged off for next level of investigation. Statistical outlier detection based on the historical statistics in the customer profile is one technique to detect the suspect transaction.

Pattern analysis [7] on the transaction events and comparing against the pre-recordings of patterns of fraudulent activity is the popular technique used to catch any fraudulent customer activity real-time. Time series analysis techniques [8] are also employed to identify whether the customer activity fits the well-defined business rules of fraudulent activity.

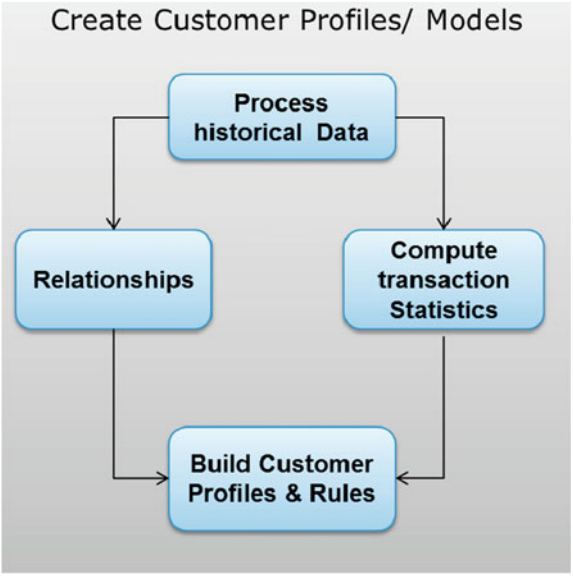

Fig. 4 Building customer profile for fraud detection

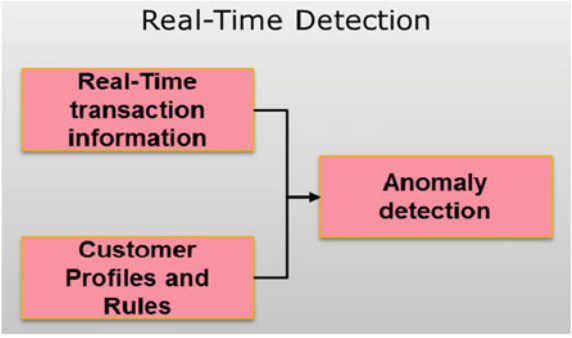

Fig. 5 Real-time fraud detection using customer profile

3.2 Money Laundering

Money laundering is a more sophisticated fraud, and detecting it requires setup of more complex and integrated multi-dimensional database system with data from different sources such as bank transactions, and law enforcement databases.

Complex networks of relationships are identified by linking data gathered over sources such as phone, e-mail, Web browsing, and travel records thereby identifying the connections between known and unknown players. Graphs of interconnected banking institutions, customer accounts, and transactions at certain times using certain devices are used to help identify potential money laundering schemes. Data analysis techniques such as clustering, classification, outlier identification, and data visualization tools [9] can be used to detect patterns in transactions involving large amounts between specific set of accounts. These techniques have the potential to identify key associations and patterns of activities that help identify suspicious cases for further investigation.

3.3 Risk Analysis

In general, banks and financial institutions have mechanisms for quantifying risk and mitigating it. Various market forces play their part in various types of risk, and a clear understanding of the potential losses for all the possible situations is needed.

Out of the various types of risks in the financial industries [10], prediction of default on loan accounts and credit card accounts is one of the important areas due to the enormity of these accounts and mitigating losses from these accounts becomes fundamental to the business. Prediction of various factors that are responsible for defaults is done using data mining techniques related to attribute selection and attribute relevance (Fig. 6). Based on the outcomes of the analysis, banks can identify customers who belong to low-risk category or offer favorable payment plans to the customer.

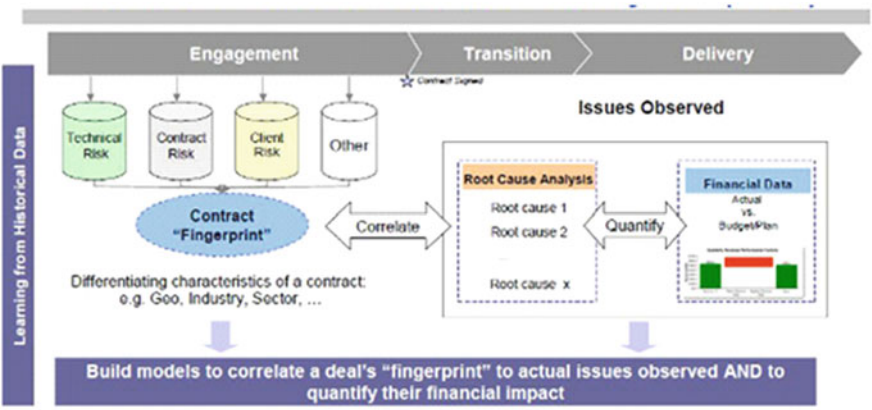

Fig. 6 Financial risk analytics framework

4 Applications in Retail Industry

Most of the big name retailers such as Costco, Wal-Mart, Target, and Amazon use big data for various operations including inventory management, product recommendations, tracking customer demographics, and tracking and managing the side effects of product recalls. Some retailers have used this customer data to improve the quality of service and enhance the customer loyalty.

4.1 Recommendation of Products

One of the well-known strategies that the retail companies employ to increase their revenues is to recommend products to the customers that they might be interested in based on what the customer is currently purchasing. This is typical of an e-retailer whose back-end systems run product recommendation engines by cross-referencing the items among sales records from various customers who may have purchased the same item earlier.

The retailers who have presence online and offline (brick and mortar) can use the data collected across multiple channels and come up with the purchase patterns and recommend products online. Path and pattern analytics are used on the historical customer purchasing behavior across multiple channels to generate high-quality recommendations. Collaborative filtering techniques are used on a customer's historical purchases and searching patterns, and comparing against other customers to predict the next recommendation.

Collaborative filtering technique is used in the recommendation systems by the e-retailers [11] such as amazon for recommending products, and the same techniques are used by the movie recommendation engine that Netflix uses. These same

techniques are employed offline in coming up with the weekend fliers, advertise on sales receipts, or coming up with the promotions by bundling items in order to promote sales.

4.2 Predicting Trends

Retailers collect huge amount of data about customers including location, gender, and age from their various transactions. Mining of retail data can help identify customer buying patterns and trends which will in turn help identify customer needs for effectively plan for product promotions and attract more customers and increase revenues/profits [12]. Multi-dimensional analysis and visualization tools of the dataset can be used for the prediction which could help with the company planning of the logistics/transportation of the needed goods.

5 Applications in Manufacturing

Manufacturing companies have become highly competitive across the world with the margins of doing business going down every day. The manufactures are always on the lookout for optimizing costs in running factories thereby increasing the margins. Big data analytics is helping in a couple of areas as discussed below [13].

5.1 Preventative Maintenance

In the automated world of manufacturing, sensors are used everywhere in monitoring the assembly line so that the failures can be quickly identified and fixed to minimize the downtime. The root cause of plant failure could be due to one or more of the numerous possible parameters spread across different subsystems linking the assembly line. Huge amount of sensor data, all unstructured data, is accumulated over the running of the manufacturing plant. Historical maintenance records for the various subsystems are also gathered in the semi-structured format. And logs related to the productivity relative to the peak capacity are also gathered along with the maintenance records and sensor data.

Time series analysis of the various subsystems based on their respective sensor data and performing pattern matching against the failure case is used for catching the potential failures. Also, path analysis and sessionization techniques are used to capture the critical events based on correlations between the sensor readings, historical maintenance records, and logs to predict the probable failures. This helps take preventative measure to keep the line running for extended period of time without interruptions and also help with improving the safety of running the operations.

5.2 Demand Forecasting

The most important factor in businesses which are tied to manufacturing industry is to optimally use the resources where the day-to-day orders keep changing dynamically. Forecasting sales and the time frame when they happen will help plan for timely acquisition of raw materials, ramping up/down production, manage warehousing, and shipping logistics. In the short term, overestimating demand leaves the manufacturer with unsold inventory which can be a financial drain and underestimating implies missed opportunities. In the long term, demand forecasting is required to plan for strategic investments and business growth. Hence, for effective running of a business with maximum profitability requires a solid forecasting system.

Time series analysis is a popular forecasting technique used to predict future demand and is based on the historical sales data. This simplistic method in generating future forecast is inaccurate when the environment is dynamic with factors such as changing customer requirements and impact of competition.

Predictive modeling [14] is a more advanced and accurate forecasting technique which has the capability to factor in all the variables impacting future demand. The model also facilitates with testing various scenarios and helps understand the relationship between the influencing factors and how they affect the end demand.

6 Applications in Telecommunications

With the expansion of telecommunications services across the globe, the telecommunications industry is trying to penetrate various markets with diverse service offerings in voice, video, and data. With the development of new technologies and services across multiple countries, the market is growing rapidly and has become highly competitive between various service providers.

Figure 7 shows the big data analytics framework for telecom domain that is used as the basis for formulating the strategies for better business. Business insights for different departments are mined based on the data collected across various platforms. Some of these include

1. Customer/subscriber data: Personal information and the historical relationship with the provider.
2. Usage patterns.
3. Customer service records: Service-related complaints or request for additional services and feedback.
4. Comments on social media.

In the following sections, we will review a couple of areas where the industry is trying to identify avenues for revenue preservation and generation using big data.

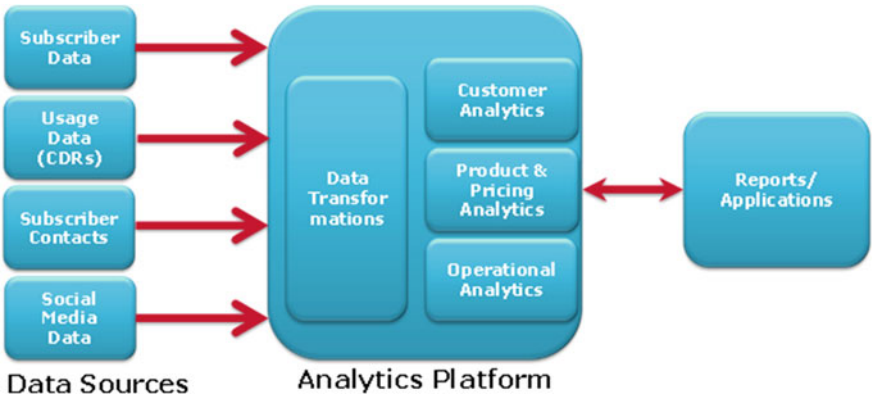

Fig. 7 Big data framework: telecommunications domain

6.1 Customer Churn

It is well known that the customer churn is a big headache for the all the telecom service providers. Customers leaving the existing service provider and signing up with a competitor cause revenue/profit/loss. It is a costly affair to acquire new customers with new promotions and has an effect of increased marketing costs which in turn has the effect on profitability.

Studies have shown that proactively identifying the key drivers for churn and developing strategies in retaining customers help minimize the revenue and profit erosion. The service provider can then focus on upgrading the underlying network infrastructure for better quality of service and better support services to retain and grow the customer base.

Various statistical data analysis techniques [15] are traditionally used to identify the triggers to customer defections and apply these triggers to the existing subscribers and evaluate chances of canceling their service and moving to another provider. Using customer behavior data collected on different channels such as calling profiles, customer complaint calls to the call centers, comments over e-mail, and feedback surveys, better churn prediction can be done to identify high-risk customers. In order to figure out the patterns of events leading to the churn, path analysis techniques are used. Using the Naive Bayes classifier for text analysis, a model is built to identify the high-risk customers.

Another popular technique used is graph engines [16] to represent connections between users based on the call detail records and then identify communities and influencers within the user communities. One of the remedial actions is to engage the high probable churn customers and offer incentives and extend the contracts for additional time period.

6.2 Promotion of Services

Telecom service providers are constantly looking to increase their revenues by recommending auxiliary services to customers that they might be interested in based on the current subscription plan. This is done either through cross-referencing with the customers with similar profiles. Another strategy is to promote the next best plan for a small incremental price. The data analytic techniques used for these recommendation engines [17] are fundamentally same as used in e-tailing business.

7 Applications in Social Media

Online social media is growing leaps and bounds as witnessed by the growth in the active user base and the amount of data that it generates. Sites such as Facebook, Twitter, Google+, LinkedIn, Reddit, and Pinterest are some of the most popular online hangout places these days. Even big corporations have started using social media as a business channel by having their presence through Facebook accounts, Twitter accounts, YouTube channels, and company blogs to name a few. The inherent openness of the social media to everyone to hear and voice their opinions and build new relationships has paved way to the creation of wealth of data. This has caught the attention of data scientists in exploring the use of social media in various areas.

Figure 8 illustrates a typical framework for applications involving social media analysis [18] with the various components highlighted. Social media analysis involves gathering and analyzing huge data that the social media generate to make

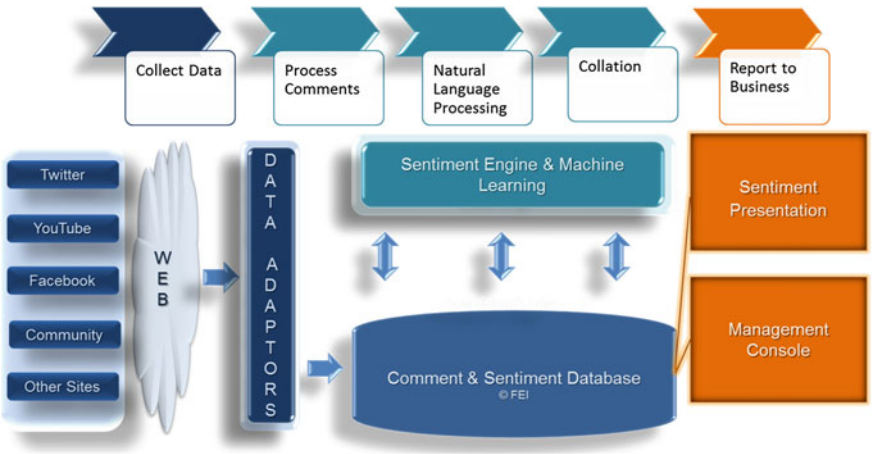

Fig. 8 Typical framework for social media analysis

business decisions. The goals of this analysis include strategies on product marketing, brand promotion, identifying new sales leads, customer care, predicting future events, foster new businesses, etc.

The work flow includes the phases of data collection, data transformation, analysis, and presentation dashboard. The social media data consist of mostly unstructured data ranging from blog posts, and its comments link to Facebook friends, tweets/retweets, etc. Based on the specific objective of the analysis, the data filtering is performed on the raw data which are then analyzed for the understanding and predictions on the structure and dynamics of community interactions. Sentiment analysis [19] and reputation management are few of the applications where natural language processing is applied to mine blogs/comments. Graph analysis techniques are applied to identify the communities and influencers within the communities.

7.1 Social Media Marketing

In social media, it is well known that different people have different levels of influencing others based on various factors, the prime being the number of connections he/she has. Representing the user-to-user connections in a graph as shown in Fig. 9 helps identify the key influencers [20] who then can be targeted with the product advertisements. This has been shown to help in creating brand awareness and facilitate viral marketing of new products.

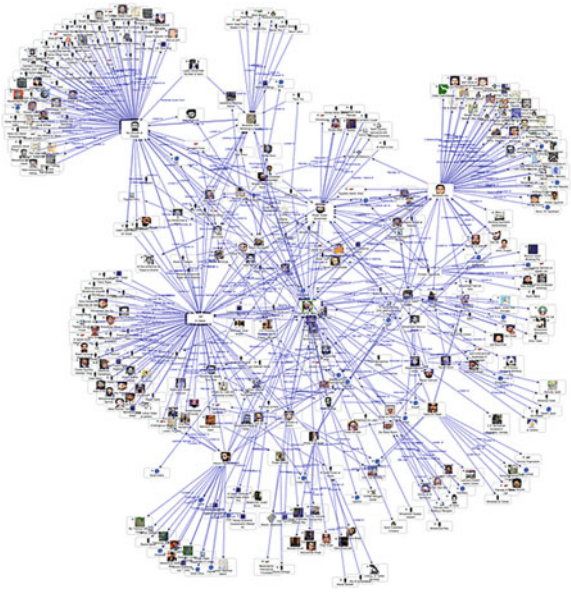

Fig. 9 Identifying influencers using K-means clustering techniques

Also, by offering incentives to the customers with most influence in a community, and leveraging his/her influence, customer churn can be contained.

7.2 Social Recommendations

Graph and link analysis is used extensively in professional social networking sites such as LinkedIn to identify and recommend other professionals that a user may be interested in establishing connection based on the existing connection mix.

Reddit site uses similar analysis of graphs built using the articles/posts and the interests of the users reading them to recommend new articles/posts to users with similar interests. List of articles in the database, user profiles, and profile of user interests are analyzed to come up with the recommendations across multiple users. Analytic technique used to organize data into groups or clusters based on the shared user attributes is K-means clustering algorithm.

8 Applications in Health care

Application of big data analytics is gaining importance in the health care industry due to the characteristics of the business involving huge dataset of customer electronic health records, the goal to deliver service at minimum cost, need for critical decision support, etc.

Figure 10 shows a typical framework for applications in health care industry capturing various components of the typical platform. Huge amounts of health care data are collected that includes clinical data such as laboratory records, doctor's notes, medical correspondence, electronic medical records (EMRs), claims, and finance. Advanced analytics on this data is used to improve customer care and

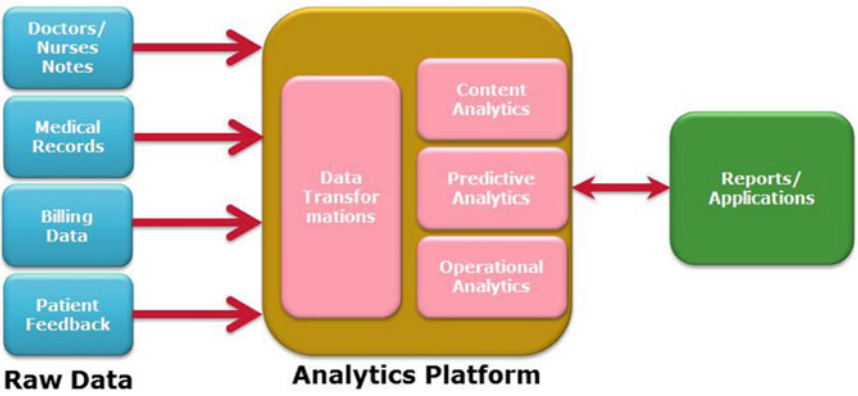

Fig. 10 Analytics framework for health care

results, drive efficiencies, and keep the costs to minimum. Analytics is also used to do a thorough investigation and detect adverse side effects of drugs which then enable quick recall of those drugs.

Following are a few examples of big data analytics in health care industry:

Finding New Treatments
National Institutes of Health [21] in USA maintains the database of all the published medical articles on various health topics and has opened up access to all the interested researchers. This dataset of documents is huge, and mining meaningful information is a challenge.
Researchers have used the semantic searches on this database to uncover new relationships between therapies and outcomes. Graph analysis [6] is used by researchers focusing on cancer who discovered that immunotherapy performs better than chemotherapy in certain cases of cancer. Visualization techniques [22] are used to find the correlations quickly.
Multi-Event Path to Surgery
Applying path and pattern analysis techniques to the data obtained from the patient records with the different procedural codes, it is possible to identify sequence of events leading to expensive surgeries. Using this info, better preventative care can be provided to avoid surgery and help reduce the medical costs.
Reduction in Claim Review
Evaluation of medical claims involves looking at doctor notes, medical records, and billing procedural codes which is time consuming and laborious process especially in cases where the treatments were complex involving multiple procedures. In order to reduce this manual effort, text analytic techniques, namely FuzzyMatch, are employed to determine inaccurate billing practices as well as potential abusive, fraud, or wasteful activity.

9 Developing Big Data Analytics Applications

The framework for a big data analytics application is conceptually similar to that of a traditional business intelligence application with the following differences.

- The main difference lies in how the structured and unstructured data are stored and processed as demonstrated in the big data framework chapter (Chap. 2). Unlike the traditional model where the BI tool is run on the structured data on mostly a stand-alone node, the big data analytics application, in order to process the large scale of data, breaks down the processing and executes across multiple nodes accessing the locally available data.
- Unlike the classical BI tools, the big data analytics tools are complex and programming intensive and need to be able to handle data residing in multiple formats.

- A different application development framework that takes advantage of running lots of parallel tasks across multiple nodes.

Development of big data applications involves awareness to various platform-specific characteristics such as

- Computing Platform—A high-performance platform which includes multiple processing nodes connected via a high-speed network;
- Storage System—A scalable storage system to deal with massive datasets in capturing, transforming, and analyzing;
- Database Management System;
- Analytics Algorithms—Develop from scratch or use the third-party open-source or commercial software suites
- Performance and scalability needs

Other than the knowledge of general platform architecture to which the targeted applications are developed, big data application developers need to be exposed to the popular big data application frameworks supported on the platform. The most popular software suite/framework that enables big data application development is called Apache Hadoop which is a collection of multiple open-source projects. Hadoop framework comprises of various utilities on top of the Hadoop distributed file systems (HDFS) and a programming model called MapReduce as described in Chap. 2 along with various other infrastructure components supporting the framework. These include PIG, HIVE, JAQL, and HBase.

Building sophisticated analytic applications requires the expertise of the data mining techniques and algorithms on top of the platform architecture and framework for which these applications are intended for. The implementations of the popular algorithms are available as open source and some are proprietary implementations. Examples of the open-source implementations include

- R for statistical analysis,
- Lucene for text searches and analysis, and
- Mahout Library—A collection of widely used analysis algorithms implemented using the map/reduce paradigm on Hadoop platforms are used for building applications. These include collaborative filtering, clustering algorithms, categorization, text mining, and market basket analysis techniques.

Analytic function implementations provided by third-party vendors or the open source have a specific programmers interface. One of the main challenges to the application developers is the complexity involved in some of the key elements of incorporating the APIs in the application. These include the following:

- Integration of open-source packages into the overall system and how the libraries are exposed to the developers.
- Support in acquiring the required input for the functions from database tables, raw files, etc.
- Support in saving function results into tables, temporary buffers, files, etc., and

- Ability to cascade multiple analysis methods in a chain to have an output of one function as an input of the next to simplify the implementation of the overall application.

The commercial big data platform solutions offered by corporations such as IBM [23] and Teradata [2] include their own proprietary application frameworks. Integration of various open-source packages and implementation/support of proprietary packages where the open-source library lacks the functionality are the key for the sale ability of the platform. These integrated commercial solutions promote the ease of use compared to the challenges using the open-source solutions as one of the strengths when marketing their platforms.

10 Conclusions

Majority of large companies are dealing with the problem of finding value in the huge amount of data that they have collected over the years. Depending on the market segment the business is addressing, different data analytic techniques are used to identify new markets, optimize operational efficiencies, etc., so as to increase the bottom line.

In this chapter, we tried to present handful of areas in different industries where the applications of big data and analytics have been effectively used. Identifying new areas and exploring new solutions will be the area of focus for the future. Corporations have started seeing value in putting dollars in data-driven strategies, and the realization that the big data strategy is a key component of business, in order to stay competitive, is gaining ground.

Exercises

1. Write an application to recommend new music labels for users based on the historical listening profiles of the user base. The basket analysis can use the dataset that is available at http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html.
2. Yahoo! Messenger is a popular instant messaging application used by many users to communicate to their friends. A sample dataset of so-called friends graph or the social network is available at http://webscope.sandbox.yahoo.com/catalog.php?datatype=g titled "Yahoo! Instant Messenger Friends Connectivity Graph." Write an application to identify the top 5 users who have most influence in the given social network.
3. Visualize the social network of users for the dataset indicated in Exercise 2 above. Use the open-source graph analysis tool called Gephi for this

visualization (available at http://gephi.github.io/users/download/. Use the quick start tutorial at http://gephi.github.io/tutorials/ to render and identify the communities for the above dataset.

4. Microsoft Corp. has published a dataset which captures the areas of www.microsoft.com that users have visited over a one-week time frame. This dataset is freely available to users at http://kdd.ics.uci.edu/databases/msweb/msweb.html.
   Write an application to predict the areas of www.microsoft.com that a user can visit based on data on what other areas he or she visited.
5. Using sentiment analysis concepts/algorithms gained in the earlier chapters, analyze the movie reviews/feedback data available at http://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data to build a model to predict the positive, negative, and neutral sentiment of the reviewers. Use 75 % of the data for the model and the remaining 25 % of the data to validate the model.
6. Using R open-source statistical and data analysis tools, write an application to predict the movement of a stock belonging to DOW Jones. The sample dataset is provided at https://archive.ics.uci.edu/ml/machine-learning-databases/00312/
7. Demonstrate with an example how to build a prediction model based on Naive Bayes for text. And then demonstrate with an example using the built model to do text prediction.

References

1. Big Data Calls for New Architecture, Approaches. http://tdwi.org/articles/2012/10/24/big-data-architecture-approaches.aspx
2. Big Data: Teradata Unified Data Architecture in Action. http://www.teradata.com/white-papers/Teradata-Unified-Data-Architecture-in-Action/
3. How Bigdata can help the banking industry: A video post. http://www.bigdata-startups.com/BigData-startup/big-data-analytics-banking-industry-video/
4. Global Fraud Study: Report to the Nations on Occupational Fraud and Abuse, Association of Certified Fraud Examiners http://www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/rttn-2010.pdf
5. Whitepaper from ACL: Fraud detection using Data Analytics in the Banking Industry. http://www.acl.com/pdfs/DP_Fraud_detection_BANKING.pdf
6. Diane, J.C., Holder, L.B.: Mining Graph Data. Wiley, New York (2007)
7. Jean-Marc, A.: Data Mining for Association Rules and Sequential Patterns. Springer, Berlin (2001)
8. Nettleton, D., Kaufmann, M.: Commercial Data Mining Processing, Analysis and Modeling for Predictive Analytics Projects, Elsevier, North Holland (2014)
9. Analytics in Banking Services. http://www.ibmbigdatahub.com/blog/analytics-banking-services
10. IDC White Paper: Advanced Business Analytics Enable Better Decisions in Banking. http://www.informationweek.com/whitepaper/Customer-Insight-Business-Intelligence/Analytics/idc-white-paper-advanced-business-analytics-enabl-wp1302120869
11. Su, X., Taghi, M.K.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 421425 (2009)

12. Das, K., Vidyashankar, G.S.: Competitive advantage in retail through analytics: developing insights, creating value. Inf. Manage. http://www.information-management.com/infodirect/20060707/1057744-1.html (2006)
13. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
14. Makridakis, S., Wheelwright, S., Hyndman, R.J.: Forecasting: Methods and Applications. Wiley, New York (1998).
15. Analytics in Preventing Customer Churn. http://www.teradata.com/Resources/Videos/Prevent-Customer-Churn-Demonstration/
16. Richter, Y., Yom-Tov, E., Slonim, N.: Predicting customer churn in mobile networks through analysis of social groups. In: SDM, Columbus (2010)
17. Big Data Analytics Use Cases—Ravi Kalakota. http://practicalanalytics.wordpress.com/2011/12/12/big-data-analytics-use-cases/
18. Foundations Edge: Media Analysis Framework. http://www.foundations-edge.com/media_analytics.html
19. Ogneva, M: How companies can use sentiment analysis to improve their business (2010). http://mashable.com/2010/04/19/sentiment-analysis/
20. Blog postings on Social Media Marketing. http://practicalanalytics.wordpress.com/category/analytics/social-media-analytics/
21. Big Data Disease Breakthroughs—William Jackson: Information Week. http://www.informationweek.com/government/big-data-analytics/big-data-disease-breakthroughs/d/d-id/1316310?
22. Periodic Table of Data Visualization Methods. http://www.visual-literacy.org/periodic_table/periodic_table.html
23. Big Data Technology. http://www.ibm.com/big-data/us/en/technology/

Author Biography

Hareesh Boinepelli is presently an engineering manager at Teradata Corp. Past work experience includes Kicfire Inc., McData Corp., Sanera Systems, and Nortel Networks over the last 15+ years. Experience covers the areas of telecommunications networks, storage networks, data warehousing, data analytics, etc. Academic qualifications include Ph.D. from Telecommunication Research Center, ASU, Arizona, M.E (ECE) from IISc, Bangalore, and B.E. (ECE) from College of Engineering, Osmania University.