Chapter 12
Big Data Applications in Business Analysis

Sien Chen, Yinghua Huang, and Wenqiang Huang

Abstract How can service providers turn their big data into actionable knowledge that drives profitable business results? Using the real-world case of China Southern Airlines, this chapter illustrates how big data analytics can help airline companies to develop a comprehensive 360-degree view of the passengers. This chapter introduces a number of data mining techniques, including Weibo customer value modeling, social network analysis, website click-stream analysis, customer activity analysis, clustering analysis, Recency-Frequency-Monetary (RFM) analysis, and principle component analysis. Using the sample dataset provided by the airline company, this chapter demonstrates how to apply big data techniques to explore passengers' travel pattern and social network, predict how many times the passengers will travel in the future, and segment customer groups based on customer lifetime value. In addition, this chapter introduces a multi-channel intelligence customer marketing platform for airlines. The findings of this study provide airline companies useful insights to better understand the passenger behavior and develop effective strategies for customer relationship management.

12.1 Introduction

While many companies are aware of the significant value of big data, they are struggling in using appropriate data analytical methods to figure out useful insights from the mountains of real-time structured and unstructured data. How can service providers turn their big data into actionable knowledge that drives profitable business results? Using the real-world case of China Southern Airlines,

S. Chen (✉)
Xiamen University, Xiamen Xindeco Ltd, Xiamen, China
e-mail: sandy80@vip.sina.com

Y. Huang
San Jose State University, San Jose, CA, USA
e-mail: yinghua.huang@sjsu.edu

W. Huang
China Southern Airlines, Guangzhou, China
e-mail: hwq@csair.com

S. Yu, S. Guo (eds.), *Big Data Concepts, Theories, and Applications*,
DOI 10.1007/978-3-319-27763-9_12

this chapter illustrates how big data analytics can help airline companies to develop a comprehensive 360-degree view of the passengers, and implement consumer segmentation and relationship marketing.

Known as the "people industry", the transportation and tourism industries have been faced with the challenges of providing personalized service and managing customer relationships for a long time. As the number of international and domestic travelers keeps growing rapidly, the transportation and tourism industries generate a massive amount of consumer behavioral data through both online and offline service delivery processes, which convey important information about the customers and their value to the service providers. In this new era of business intelligence, data itself has become a critical strategic and competitive asset for every company [1].

Focusing on the passenger's travel history and webtrends data, this chapter introduces a number of data mining techniques to investigate passenger behavior. Weibo customer value modeling, social network analysis, website click-stream analysis, customer activity analysis, cluster analysis, Recency-Frequency-Monetary (RFM) analysis, and principle component analysis are applied in analyzing passenger profiles and managing customer relationships. In addition, this chapter introduces a multi-channel intelligence customer marketing platform, and explains how airlines companies can use this platform to collect passenger data and create analytical reports to inform business decision making.

Using the data mining results from the passenger database of China Southern Airlines, this chapter further discusses the marketing implications of big data analytics. The data mining results reveal passenger travel patterns, preference, travel social networks and other aspects of purchase behavior. This chapter explains how airline companies can design corresponding marketing segmentation and promotion strategies (e.g., cross-sell and up-sell) based on data mining results. This chapter sheds new light on airline precision marketing and customer relationship management.

12.2 Big Data Challenge in Airline Industry

Faced with the increasingly fierce competition in the aviation market, domestic and international airline companies have shifted their attention to customer relationship management [2]. They have come to realize that customer resource are the most valuable for competition. Nowadays, airline companies have a vast amount of data about every step that their customers take during the travel cycle. To a certain extent, airline companies take control of big data about their customers [3]. The passenger data are tremendous in the customer relationship management system. However, these data are usually only used to support specific operational procedures, rather than to develop business intelligence. The passenger databases contain many data descriptions, but data are not shared between different departments. In other words, passenger data storage and management are in a mess.

Therefore, how to utilize these unorganized data for developing business values, is the challenge faced by many airline companies. The effective big data production

Table 12.1 Travel life cycle and amount of data

Travel life cycle phase	How customer generates data?	Where is that data found?
Searching: This phase is usually applicable in case of leisure travelers. In this phase the customer is searching where he wants to travel	Browsing through travel service provider's website, traversing through OTA sites, clicking on ads on social media sites like Facebook, using marketing promotions by travel service providers, using travel search sites, online searching	Online travel agencies logs, social media sites like Facebook, travel service providers' analytical logs, travel search site logs, Google web logs
Planning: Before reaching this phase the traveler has narrowed down on the destination. Now he is planning various details of his travel, like, mode of transport to the destination, what kind of accommodation will he take, what mode of transportation will he use at the destination, places he would like to see, restaurants he would like to eat at etc.	Calling up the call centers, browsing OTA sites, surfing websites of travel service providers, reading other travelers' experiences on social media sites and blogs, surfing through travel review sites such as Tripadvisor, see pictures and videos	Call center logs, OTA web analytics logs, travel review sites database, social media sites and blog sites analytical databases, Google internet search database
Booking: After the traveler has planned his trip the next step would be to make all the necessary bookings like flight, hotel, transportation, tourist attraction sites etc	Travel service provider's website/call center/social media page/on property, travel agency, OTA	Travel service provider's web database, call center logs, social media database, OTA database, travel agency database
Experiencing: In this stage the traveler is using the travel services i.e. he is flying through the airline he booked, staying in the hotel, driving the car he rented, having dinner at the restaurant he booked etc	Traveler's on property feedback, complaints registered via call centers, his movements, time spent on the service	Feedback logs, feedback with employees, location/movement database
Sharing: After the traveler has completed his travel, he goes on and shares his experience with his own and outside network	Word of mouth, writing blogs, sharing experience on social media sites, travel review sites	Social media database, user's web profile database, travel review sites' databases

for airlines should achieve the integration of multi-channel information, and support the analysis of consumer preferences and personalized recommendation services. Table 12.1 describes how and where the customer leaves footprints during a typical travel life cycle.

Using big data techniques can help the airline companies to develop precision marketing strategies, and further increase the conversion rate of the whole network airlines [4]. For example, the likes of Delta, Emirates, British Airways, Iberia, KLM and Alaska have already armed their cabin crew with tablets and other handheld devices to ensure they are informed of passenger preferences and well equipped to push ancillary sales. In 2013, Delta Airlines equipped 19,000 flight attendants with Windows Phone 8 handheld devices, which make it easier for passengers to buy ancillary items onboard [3]. In November 2010, the international airline KLM surprised its customers: As passengers arrived at security checkpoints and gates, flight attendants were there to greet them by name and give them a personalized gift—something that the passenger could use on his or her trip, or enjoy when they returned home. In November 2010, the KLM Twitter feed was viewed more than 1 million times. What made this particular campaign stand apart from other run-of-the-mill marketing campaigns is personalization, which enabled the airline to offer customers something that held real, tangible value [5].

12.3 Passenger Data Mining Application

In order to illustrate how the massive passenger data can be utilized for business decision making, this chapter presents a real-world case of China southern Airlines. Using various data mining methods, including Weibo customer value modeling, social network analysis, website click-stream analysis, customer activity analysis, cluster analysis, Recency-Frequency-Monetary (RFM) analysis, and principle component analysis, this chapter presents how to apply big data techniques to explore passengers' travel pattern and social network, and predict how many times the passengers will travel in the future. The findings will provide airline companies the ability to become more effective in customer relationship management.

12.3.1 Weibo Customer Value Modeling

Along with the popularity of social media, the airline industry has gradually utilized this platform to further improve passenger services and enhance its own competitiveness. For example, according to a report by *Washington Post*, more and more American airlines including Southwest Airlines, United Airlines and Delta Air Lines have started to utilize social media to solve passenger complaints. KLM is the first airline that integrates social networking websites into the flight reservation process, which released the service of "meeting & sitting in the same flight", so that the passengers can link the data of Facebook or LinkedIn into the flight reservation process, and it is convenient for them to know who sits in the same flight as theirs [5]. The amount of information of social media is larger and larger, and if the airlines can utilize their resources reasonably, they can provide better services to the passengers, retain old customers and develop new customers, thus bringing better development for the airlines [6].

Currently, there are three primary models for evaluating the social media customer value: Klout scoring, Kred scoring and Weibo data mining.

12.3.1.1 Klout User Scoring Model

Klout scores website users based on a set of data mining algorithms and according to the indexes such as if a user has social media accounts or not, his/her number of fans, his/her update frequency, if his/her content is praised by others, his/her number of comments, and his/her number of forwarding, and so on. Klout score social media users with 100 as the standard. Its specific scoring mode is: if the customer has an account of Twitter or another social media website, and your messages can be seen by others, you already get certain scores. Based on this, Klout will score you according to the indexes such as your number of fans, frequency of your information update, Klout scores of your friends and fans, number of people who like your things, number of replies, number of forwarding, and so on.

The Klout Scoring Model has three main factors: true reach, amplification and network impact. True reach refers to the number of people that a user has influenced, amplification mainly refers to the number of people who forwarded your posts and frequency impact, and network impact refers to the network impact of a user.

12.3.1.2 Kred Scoring Model

The main data source of Kred is Twitter, and it established a Scoring Model based on surrounding followers, likes or shares with others. The final score of Kred consists of two parts: influence score and external connection score. The influence score reflects the breadth of a user's influence, and the external connection score reflects the influence depth. The influence score measures the ability of a user to motivate others, from 1 to 1000, which depends on the number of the user's Twitter messages forwarded, the number of new followers and the number of replies of the user. The influence depth level refers to the influence score of a certain topic posted by a user. The influence depth level does not consider the number of followers or other statistical data, but is based on what a user has already shared, said and done in a specific circle. In different circles, a user has different influence scores and influence depth scores. It is clever to be in accordance with the right methods.

The main difference between Kred and Klout is that the transparency of Kred is comparatively high, which will display how a user gets the final score, even the value score of a specific forwarding. The value of a common forwarding might be 10, but the value of a forwarding of a user with high Kred score might be 50. The score of being mentioned by others is higher than the score for being followed by others, and so forth. The score of each person is calculated in real time. In addition, Kred attempts to link virtuality with reality, which brings the achievements such as honors, medals and certificates, etc. that a user has in the realistic world into the Kred scoring calculation.

12.3.1.3 Weibo Data Mining Algorithm

The user scoring of micro data mainly inspects factors in three aspects: a user's activity, spreading capacity and reach, and its specific scoring algorithm is as below:

$$\text{Influence} = \text{activity} + \text{spreading capacity} + \text{reach} \tag{12.1}$$

In this algorithm (12.1), spreading capacity refers to the number of valid posts and number of valid people forwarding and commenting posts. Reach refers to the number of active fans. Activity refers to the number of valid posts of posting, forwarding and commenting Weibos.

12.3.1.4 Airline Passenger Weibo Value Algorithm

In the context of the airline industry, we applied the third model of Weibo data mining, and evaluated the value of airline passengers on *Sina Weibo* in terms of three aspects: influence, activity and relevance with the airline company's Weibo. The influence factor includes number of Weibos posted, number of Weibos forwarded and number of fans, and so on. Activity factor includes number of Weibos posted by the user, number of Weibos forwarded by the user and number of check-ins of the user, and so on. Relevance factor includes the Weibo interactions between a passenger and the airline company, and the influence of the user's Weibo.

We model the airline passengers' value of Weibo in the algorithm as below:

$$SNvalue = \omega_1 * IF + \omega_2 * ATC + \omega_3 * Corr \tag{12.2}$$

In this algorithm (12.2), SNvalue is the airline passenger's value on Weibo. IF is influence. ATC is activity. Corr is relevance, and $\omega_1, \omega_2, \omega_3$ are the weights of the factors.

Influence Factor (IF) IF refers to the potential spreading influence of a user. IF index is mainly related to the number of fans and quality of the fans of a user. The bigger the number of fans of the passenger, the bigger the number of people who might see the passenger's activities (posting and forwarding Weibos, etc.), and the bigger the number of people who are influenced. The higher the quality of the fans, the bigger the number of people who spread outwards and influence. Based on the idea of PageRank, we can define a PassengerRank for the passenger in the social networking websites. As to the passenger in the social networking websites, the bigger the number of "important people" in his/her friends is, the higher the rank that the passenger corresponds to is. Three indexes used for measuring PassengerRank:

- Number of fans
- If the fans have comparatively high PassengerRank values
- The number of people that the fans pay attention to

Activity Factor (ATC) At present, *Sina Weibo* uses number of fans as the basis to rank the users. User ATC index is related to number of fans, number of Weibo posted by the user and number of Weibo forwarded by the user, and so on. If we adopt the weighted approach, we can get the algorithm as below:

$$\text{ATC} = \omega_1 * \text{wb\_num} + \omega_2 * \text{trans\_num} + \omega_3 * \textit{fans\_num} \tag{12.3}$$

In this algorithm (12.3), wb_num is the number of Weibo posted. Trans_num is the number of Weibo forwarded. Fans_num is the number of fans. ω is weight factor of Weibo forwarding.

Relevance (Corr) Corr refers to the relevance degree between a passenger and the airline industry. In *Sina Weibo*, Corr is mainly related to the Weibo user properties and Weibo content. User properties mainly refer to the properties such as industry category of a user, which can directly reflect the potential possibility of the interaction of a user with the airline industry. Weibo content refers to the relevance degree of the Weibo posted by a user and the Weibo forwarded by a user with the field of airline industry. Taking one step further, we can analyze the role of a user in the public opinions related to airlines from the Weibo content, which is an important index in the relevance measurement. At present, through the analysis on the emotion of a user's Weibo towards the airline industry, we get the relevance factor as below:

$$Corr = \sigma * (\omega * SF) \tag{12.4}$$

σ refers to the relevance degree between the Weibo content and China Southern Airlines, and the value range is between 0 and 1. ω refers to the degree of impact on the result of Weibo emotion analysis. SF refers to public opinion factor, mainly considering the relevance degree of the Weibo content of a user with the public opinions on the airline industry.

12.3.2 PNR Social Network Analysis

In the airline industry, there are three important types of travel data: passenger name record (PNR), share of wallet (SOW) and webtrends. These three data types can be joined by the user's identity with the addition of user name and the order number, then be imported into a big data platform. The PNR, SOW, and webtrends data were retrieved from the passenger database of China Southern Airlines.

PNR archives the airline travel itinerary for individual passenger and a group of passengers traveling together. Usually, passengers and their accompaniers are close to each other, such as families, friends, lovers, colleagues and so on. Therefore, the social network between passengers and their accompaniers can be constructed through exploring the PNR history data. The PNR data analysis will help the airline

company to identify who are influential passengers in their social circles. The following steps can be applied to develop PNR social network:

1. Take every passenger as a node, if two passengers have travelled together, then there is a connection.
2. Add two directed connections to every pair (assume passenger A and passenger B who travelled together before are a pair), every connection strength is determined by the following factor: the proportion of "go together" times to single travel times. For example: Passenger A (a corporate executive) travelled 100 times in total. There are five times that he went with B (executive's mother), so the connection strength directed from A to B is 0.05. However, B only took five trips in all her life, then the connection strength from B to A is 1.
3. Count and show the detail information of passengers; such as gender, age and travel times.
4. Once the network is developed, the featured relationship and value of passengers can be determined. Figure 12.1 illustrates an example of PNR social network.

SOW is a marketing term representing traveler's value and contribution to a company, which refers to the amount of the customer's total spending that a business captures in the products and services that it offers. The technical measurement of SOW is a ratio of tickets purchase amount from an airline company to passenger's total travel times. With SOW data analysis, the airline identifies who are potential high-value travelers, and suggest corresponding marketing segmentation and promotion strategies based on different SOW level.

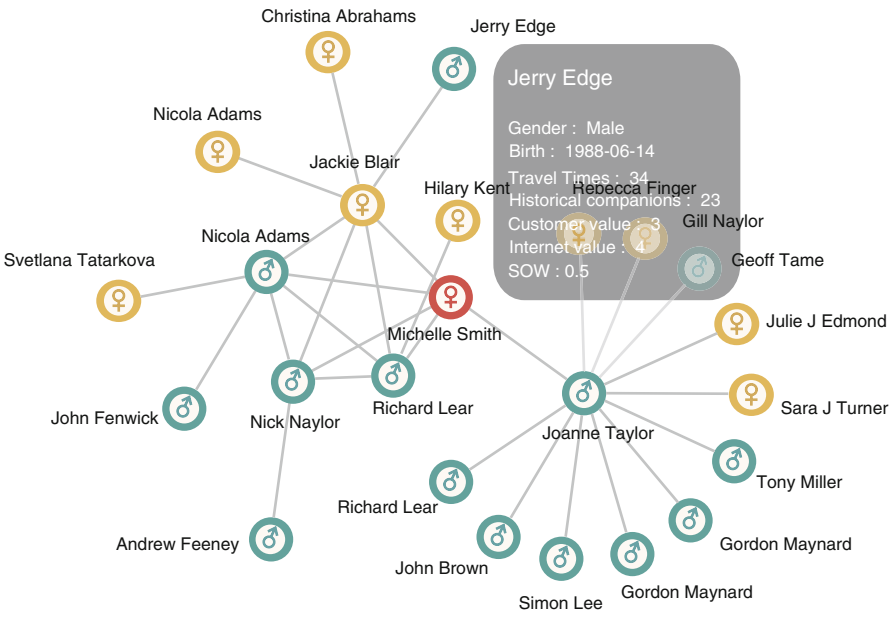

Fig. 12.1 An example of PNR social network

12.3.3 WebTrends Analysis

Airlines also can analyze webtrends data to explore passenger behavior of websites and mobile usage. Passenger's webtrends information includes mobile number, membership number, identity number, and other web browsing records. Connecting these webtrends data with other information sources provides an overview and insights on individual passenger's website and mobile usage. The following session demonstrates how the accessing event flow on WebTrends can be configured and incorporated into the sequence analysis of passenger events.

12.3.3.1 Customer Relevance

Associate the login account information (mobile number, member No., ID number, etc.) with a single view of customer (SVC) information to obtain detailed information of individual passengers. This part only targets those passengers with log data.

In addition to WebTrends, other event data of this user can be found after the users association, such as phone booking history. All these events will show in the sequence diagram of passenger events illustrated in Fig. 12.2.

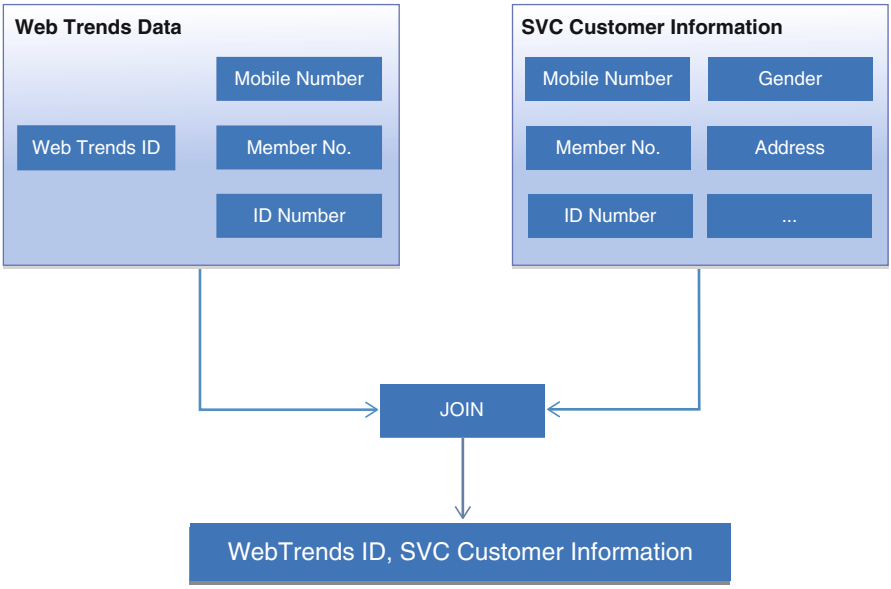

Fig. 12.2 Customer relevance

Table 12.2 Parameters in WebTrends analysis

Parameters	Description of WebTrends	Description of airline company
WT.si_n	Name of the scenario analysis	Name of procedure
WT.si_p	Identifies the step by name	Name of step pages in processing
WT.si_x	Identifies the step by position	
WT.si_cs	Identifies the step in which conversion occurs for Visitor Data Mart profiles.	

12.3.3.2 Scene Definition in WebTrends

The Parameters si_n and si_p are used in WebTrends event analysis. Because the data from the Chinese airline company are relatively rough and the scene categories defined by WebTrends are not subdivided very clearly, the Parameter ti (page title) is also used in practical applications.

Regarding the process of different events, the specific scene can be defined by client. For example, client sets up the related code on login page. Once login behavior occurred, the log will automatically classify it into the login scene. Table 12.2 shows the parameters often used to present scenes.

Decision Method About Events First, we shall decide according to the fields "Name of procedure" or "Name of step pages in processing" which are provided by Airline company. Such as a step page named "search", then this record belongs to a search event. If the page is named "flight choice", then the record belongs to the event of flight selection. Second, if these two fields show no specific events, we find "URL" to decide. Third, as the parameter "Name of step pages in processing" may not be subdivided into more specific meaning, we still will refer to the parameter WT.ti.

Parameter Extraction Parameter means the specific behavior which is equivalent to some events. For example, the parameters of flight searching maybe include departure city, arrival city, departure time . . . But, not every parameter has a value, and some fields are missing in flight inquiry information. Also not every parameter is triggered by this event, it may be triggered by the last event, just the parameter still keeping in this event.

Range of Built Event The project mainly deals with the following events and their main parameter types:

- Channels and sources (Searching, advertisement . . .)
- Login
- Users login events
- Flight searching (Extract flight information . . .)
- Flight selection (Extract flight information . . .)

- Passengers Information (Submitted information of passengers . . .)
- Orders (Order information and the Information of the finalized flight . . .)

12.3.3.3 Statistical Analysis

The following rules are made for every single passenger, and all the specifics will be decided according to practical application context.

1. How many times does the passenger login in before the ticket purchase on average (in a certain time)?
2. How many pages does the passenger browse before the purchase on average (take every page view as a unit)?
3. How many times did the passenger click advertisement before the ticket purchase?
4. How many advertisement hits after the ticket purchase?
5. Distribution of access sources (Baidu, Google, direct access etc.)
6. Distribution of passenger flight inquiries (in a certain time)

12.3.4 Display and Application of WebTrends Event Flow

The access event flow on WebTrends also is included in passengers' sequential analysis, so that all customers' event behaviors can be integrated and displayed. The event sequence diagram is based on different event types of each passenger, such as login, inquiry, order and its detail information.

The building methods are as follows:

1. Count the customer events, and show them and their frequency in time sequence.
2. After the clicks of "recent events" or "all events", the detail information in specific period can be seen, such as flight number, departure time, arrival city, and so on.

Figures 12.3 and 12.4 show examples of customer events.

12.3.5 Customer Activity Analysis Using Pareto/NBD Model

The Pareto/NBD model was originally proposed by Schmittlein et al. [7]. This model calculates customer activity level and predicts their future transactions based on their purchasing behavior. The model assumes that customers make purchases at any time with a steady rate for a period of time, and then they may drop out. The mathematical assumptions of this model are listed as below [7]:

1. While active, the repeat-buying rate λ of customer behavior follows Poisson distribution.
2. The transaction rate of different customers follows a gamma distribution $\Gamma(\gamma, \alpha)$. γ is the shape parameter and α denoted the scale parameter.

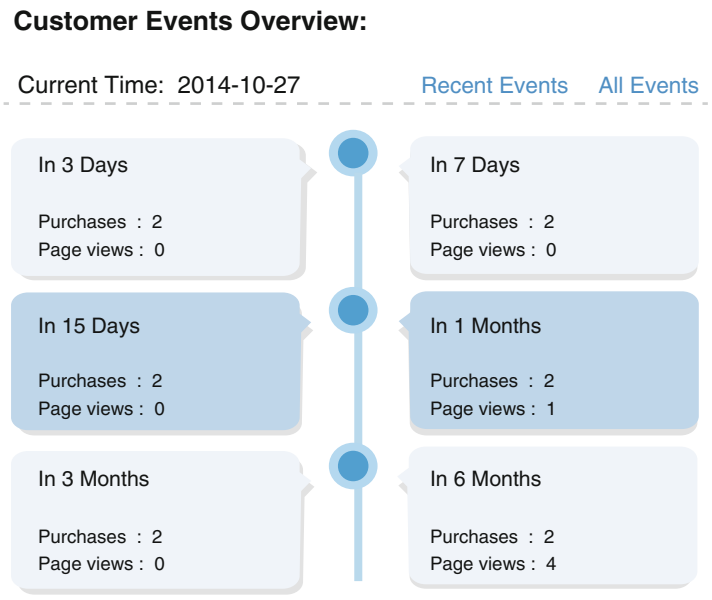

Fig. 12.3 Passenger event overview

Occurring Time	Event Details	
2013-03-10 10:12:31	2013-03-10 10:12:31	We chat Check-in
2013-01-14 09:33:23	2013-01-14 09:33:23	Log in member card 692212812028 though B2C website
2013-01-11 09:33:17	2013-01-11 09:33:17	Transfer service
2013-01-11 09:33:44	2013-01-11 09:33:44	Buffet service
2013-01-08 09:33:41	2013-01-08 09:33:41	Self-help luggage service
2013-01-07 09:33:24	2013-01-07 09:33:24	Counter check-in service
2013-01-04 09:33:18	2013-01-04 09:33:18	Ipad service
2012-12-25 09:33:50	2012-12-25 09:33:50	Book flight 20130117 TV5510 ZSSS-ZGGG though B2C website
2012-12-25 09:33:36	2012-12-25 09:33:36	Check flight 20130117 TV5510 ZSSS-ZGGG though B2C website
2012-11-27 09:33:41	2012-11-27 09:33:41	Click flight 20130117 TV5510 ZSSS-ZGGG though B2C website

Fig. 12.4 Passenger event details

3. The dropout rate μ obeys exponential distribution.
4. Heterogeneity in dropout rates across customers follows a gamma distribution $\Gamma(s, \beta)$. s is the shape parameter and β denoted the scale parameter.
5. The transaction rate λ and the dropout rate μ are independent across customers.

The Pareto/NBD model requires only each customer's past purchasing history: "regency" (last transaction time) and "frequency" (how many transactions in a specified time period). The information can be described as $(X = x, t, T)$, where

x is the number of transactions observed in the time period (0, T] and t is the time of the last transaction. With these two key summary statistics, the Pareto/NBD model can derive $P(active)$, the probability of observing x transactions in a time period of length t, and $E\left(Y(t)\middle|X = X, t, T\right)$, the expected number of transactions in the period (T, T + t] for an individual with observed $(X = x, t, T)$[8].

With passenger activity and other conditions, airlines could analyze the influence factors of activity degree which could be used to improve passenger activity. Three pieces of information were selected from the database where large passengers' records stored.

ORD_FAR.Far_Idnum:Customer id
ORD.Ord_Bok_Time:Booking time
ORD_CAS.CASH_TOTAL_TICKETPRICE:ticket price

In the database provided by China Southern Airlines, we put the passenger data from 2013-01-01 to 2013-06-30 into the Pareto/NBD model, and forecast the purchase number of each passenger in July and August, 2013. The Pareto/NBD model was implemented with R language.

Figure 12.5 shows a density distribution of the passengers' activity. We can find that the activity of most passengers is between {0.1, 0.2}. Table 12.3 lists the range of the passengers' activity. The total number of passengers is 202,370. Based on the passengers' activity the number of flying times predicted, airlines could make more effective marketing strategy for their customers.

This customer activity analysis has several important implications.

1. Customer segmentation could be done based on the passengers' activity degree. For example, customers could be divided into highly active, active and inactive. Then, airlines can carry out targeted management.
2. With the average spent by passengers and predicted flying numbers, airlines could calculate the revenue this passenger would bring to them and predict future returns.

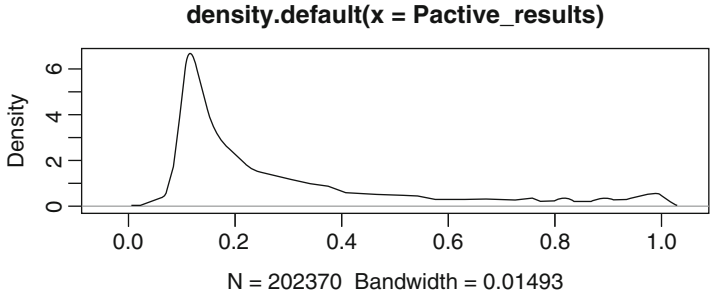

Fig. 12.5 Density distribution of the passengers' activity

Table 12.3 The scope of activity and corresponding number of passengers

The scope of activity P(Active)	Number of passengers	The scope of activity P(Active)	Number of passengers
[0,0.1]	8004	(0.5,0.6]	8337
(0.1,0.2]	96,269	(0.6,0.7]	5990
(0.2,0.3]	31,634	(0.7,0.8]	5722
(0.3,0.4]	19,538	(0.8,0.9]	6562
(0.4,0.5]	10,788	(0.9,1]	9526

3. Combining passenger active degree with life cycle length, airlines can calculate and estimate the customer lifetime value to allocate marketing resources and provide the basis for effective customer management.

12.3.6 Customer Segmentation by Clustering Analysis

Clustering analysis can be applied to segment airline passengers and explore their purchase behavior [9]. In this case of China Southern Airlines, the company website has around 6.2 million effective booking data, with passengers involved reaching the number of nearly 4 million in the past 2 years. A sample data set is retrieved from the company website, which includes 2.5 million booking data.

First, 12 variables of airline passenger behavior are selected for principal component analysis. Table 12.4 shows the 12 variables involved in principal component analysis (PCA).

Using PCA method, 12 analysis variables were integrated and transformed to nine principal component factors. The accumulative contribution of the nine extracted factors is over 0.98, indicating these factors carry more than 98 % of information which can be provided by the original 12 variables. Among the original variables, total mileage is affected by two indicators—frequency and average flight distance, average ticket price is affected by average discount and average flight distance, sum of consumption is affected by frequency, average discount and average flight distance, so the PCA analysis reveals that they can't act as separate factors. The result of PCA analysis is shown in Table 12.5. Therefore, only nine factors were selected for further cluster analysis.

Then, K-mean value analysis was conducted to explore passenger groups. Iteration experiment was used to select the combination of a group of clustering numbers and random seed which yields the best grouping result. Iteration experiment generates the best grouping number, and eight passenger groups with typical purchase characteristics are identified. The groups are described and labeled in Table 12.6.

Next, through comparing the mean values of group characteristics, we can identify the advantages and disadvantages of targeting different passenger groups. The following five findings would be useful for the airline company's business decision.

Table 12.4 Airline passenger lifetime value and purchase behavior

Characteristics	Indicators	Descriptions
Passenger lifetime value characteristics	Number of booking legs	Number of take-off and landing city pairs for client bookings
	Sum of consumption	Gross purchase sum
	Average ticket price paid	Quotient of purchase sum and number of flights
	Average discount	Price published for each city pair
	Number of days as of the last booking up to today	Difference between the last booking date and analysis date
	Total flight mileage	Sum of flight mileage of each city pair
	Average flight mileage	Quotient of Sum of flight mileage and number of flights
Behavior characteristics	Average number of days for booking upfront	Average of difference between purchase date and flight date
	Average booking time	Average of each purchase time point
	Rate of weekend flight	Quotient of number of weekend flight and total flights number
	Rate of holiday flight	Inclusive of certain days before and after the holiday
	Rate of flight for an international expo	Destination being Guangzhou in period of outward voyage, while departure from Guangzhou in the period of back trip

Table 12.5 The result of PCA analysis

Factors	Eigenvalue	Contribution	Accumulative contribution
F1: Frequency	2.89	0.241	0.241
F2: Average flight distance	1.89	0.157	0.398
F3: Average discount	1.46	0.122	0.520
F4: Advance booking	1.20	0.099	0.619
F5: Last flight trip	1.00	0.083	0.702
F6: Holiday flight trips	0.99	0.082	0.784
F7: Booking period	0.98	0.082	0.866
F8: Weekend flight trips	0.80	0.067	0.873
F9: Flight trips to an international expo	0.57	0.048	0.981
F10	0.10	0.008	0.989
F11	0.07	0.006	0.995
F12	0.03	0.005	1.000

1. Group 1, Group 5 and Group 7 have fewer numbers of bookings, with middle level of ticket price, supposed to be the ordinary mass groups, while the difference among the three groups is about the rate of flights on weekends, holiday and workdays.

Table 12.6 Passenger clusters

Clustering	Label	Advantage characteristics	Disadvantage characteristics
Group 1	Ordinary business client	None	Few number of booking, lower rate of flight trips on weekends
Group 2	Happy flight—not lost	Big number of days for advance booking	Few number of booking, the lowest discount
Group 3	Expo event	Higher rate of flight for an expo event	The smallest group
Group 4	Occasional high-end flight	High average ticket price, High average price	Few number of booking
Group 5	The masses—flight trips on weekends	Higher rate of flight on weekends	None
Group 6	Happy flight—already lost	None	Few number of booking, the lowest discount, the longest time interval since the last booking up to now
Group 7	The masses—flight trips on holiday	Higher rate of flight on holiday	None
Group 8	High-end, flight often	Big number of booking, high average price, big sum of consumption	None

2. Group 2 and 4 are groups purchasing discounted tickets, the difference is that Group 2 is still active, while Group 4 is basically lost already. Group 4 bears similar characteristics with A-type group supposedly.
3. Group 3 are travelers who flow in the same direction with those attending an expo, so we infer quite many of them are participants of the event.
4. Group 4 have a fewer number of booking, but with a higher price, while a small number of days for advance booking, suggesting this is a group with occasional travel needs, paying attention to prices seldom, so it could be a high-end group.
5. Group 8 have a big number of booking, with a high average price, and are a high-end group who fly often indeed. This high-end group are in pursuit of trends, and enjoy new technology while traveling. We can see this high-end group tend to handle special check-in, showing an obvious higher rate than other groups, at 40 %, especially online check-in and SMS check-in.

12.3.7 Recency-Frequency-Monetary (RFM) Analysis

Recency-Frequency-Monetary method is considered as one of the most powerful and useful models to implement consumer relationship management. Bult and Wansbeek [10] defined the variables as: (1) R (Recency): the period since the

last purchase; a lower value corresponds to a higher probability of the customer's making a repeat purchase; (2) F (Frequency): number of purchases made within a certain period; higher frequency refers to greater loyalty; (3) M (Monetary): the money spent during a certain period; a higher value means that the company should focus more on that customer [9].

This case study adopted an extended RFM model to analyze the airline passenger behavior. The extended RFM model incorporated average discount factor as an additional variable, because average discount factor is an important indicator to measure the price level of passenger's airline purchase. The average discount factor defined here is ratio of purchase price to published price of the airplane seat. Therefore, the extended RFM model involves four variables: number of days from the last order date to modeling (R), number of flight trips (F), sum of consumption (M), and average discount factor (D). In this way, a traveler's ID generates the consolidated data.

Principal component analysis was used to score individual travelers based on the RFMD variables, and 16 consumer groups were identified. The findings could help marketers to recognize those most valuable consumers and establish profitable consumer relationships. The procedure of the RFM analysis is described below.

12.3.7.1 Exploratory Data Analysis

This step involves taking a closer look at the data available for investigation. Exploratory data analysis consists of data description and verifying the quality of data from the airline company's databases. Tables 12.7 and 12.8 provide a general understanding of the passenger data set.

Table 12.7 reveals that the difference between the maximum and the minimum of the two variables: number of flight trips and sum of consumption is huge. The data distribution plot also indicates that the original data is heavily right-skewed. Therefore, using the original data directly in our modeling will be a big problem. In order to fix this data problem, logarithmic transformation is used regarding number of flight trips, sum of consumption and average discount factor. We also take the opposite number regarding the difference of dates from the last order date to modeling date, and then standardize the data to remove dimension's influence.

Table 12.7 Descriptive data analysis of RFMD variables

Modeling variables	N	Mean	SD	Minimum	Maximum
R: Number of days from the last order date to modeling	1,624,293	188.34	175.07	1	730
F: Number of flight trips	1,624,293	2.25	2.11	1	128
M: Sum of consumption	1,624,293	1729	2062	18	173,190
D: Average discount factor	1,624,293	0.62	0.22	0.02	3.4

Table 12.8 Correlation matrix of RFMD variables

	R	F	M	D
R: Number of days from the last order date to modeling	1	−0.006	−0.06	−0.25
F: Number of flight trips		<0.0001	<0.0001	<0.0001
M: Sum of consumption	−0.006	1	0.839	−0.075
	<0.0001		<0.0001	<0.0001
D: Average discount	−0.063	0.839	1	0.197
	<0.0001	<0.0001		<0.0001
	−0.250	−0.075	0.197	1
	<0.0001	<0.0001	<0.0001	

Pearson correlation, N = 1,624,293; When H0: Rho = 0, Prob > |r|

Table 12.8 indicates that the number of flight trips positively correlates with income (sum of consumption). The more flight trips, the bigger sum of consumption, which corresponds with the flight reality.

12.3.7.2 Principal Component Analysis

Principal component analysis is used to determine the weight of each RFMD variables. Figure 12.6 shows the steps of principal component analysis of RFMD modeling.

Through the principal component analysis, the result shows that the three RFM modeling variables account for 95 % of the overall variance. The four RFMD variables and weights were determined to further depict the passenger's value model in Table 12.9. In particular, weight of number of days from the last order date to modeling is 1.23, weight of number of flight trips is 1.21, weight of sum of consumption is 1.43, and average discount factor is 0.54.

12.3.7.3 Clustering Analysis

K-mean value clustering method was applied to generate 16 passenger groups. The four RFMD indicators can be used to analyze specific target groups in more details [11]. The four RFMD indicators can help to rank the level of passenger lifetime values, and determine individual marketing strategy and realize precision marketing in respect of individual high-end travelers.

The concept of Customer lifetime value (CLV) is adopted to evaluate the profitability of each cluster. CLV is the present value of all future profit generated from a customer [12]. In this case study, the average CLV value of each cluster can be calculated with the equation:

$$CLV_{ci} = NR_{ci} \times WR_{ci} + NF_{ci} \times WF_{ci} + NM_{ci} \times WM_{ci} + ND_{ci} \times WD_{ci} \quad (12.5)$$

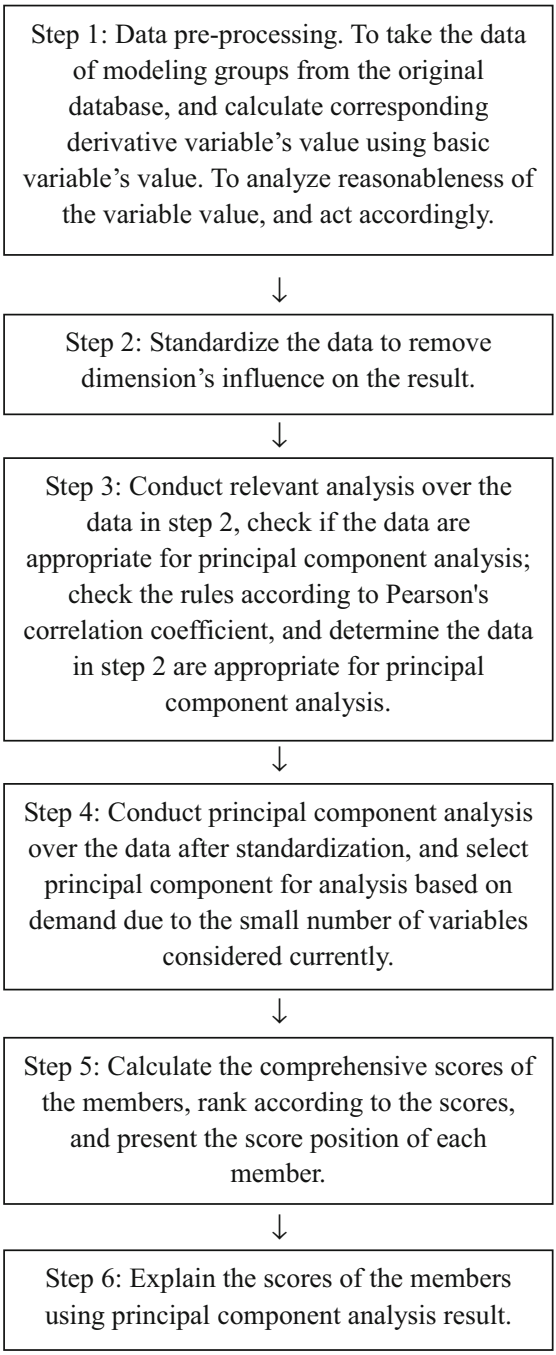

Fig. 12.6 Principal component analysis steps

Table 12.9 Basic statistics of RFMD data

Modeling variables	Weight (all principal components)	Weight (the former three principal components)
R: Number of days from the last order date to modeling	1.23	1.23
F: Number of flight trips	1.32	1.21
M: Sum of consumption	1.32	1.43
D: Average discount	0.60	0.54

NR_{ci} refers to normal recency of cluster ci, WR_{ci} is weighted recency, NF_{ci} is normal frequency, WF_{ci} is weighted frequency, NM_{ci} is normal monetary, WM_{ci} is weighted monetary, ND_{ci} is normal duration of cluster ci, and WDci is weighted duration. The result of clustering analysis is shown in Table 12.10.

Based on the result of clustering analysis, some insights of customer segmentation and corresponding business strategies can be developed. For example, for those who don't take a flight for more than half a year, mark as low value directly, for example, group 16's score is 74.89, mainly because this group took flights a lot, but they fail to purchase on the website for more than 300 days, so they're treated as a lost group (lost herein refers to being lost to other channels or other companies).

Group 12 and 13 have higher average scores, value of variables is higher than average value, so they are company's important clients who are supposed to be developed and maintained; continuous client care is necessary, service measures and client experience need to be improved; these two groups aren't sensitive to prices, so market measures like promotion and fare reduction aren't suitable. Each client of each group has a score, and priority of resources can be given to these groups in light of scores and ranking.

Group 1 have lower value scores, because their average discount is low, but they made purchases in the last two months, indicating they're active relatively, so this group needs attention. Further study can be made, paying attention to age, purchase website source and other information of the group. It could be an individual traveler base that forms the long tail of website sales.

Figure 12.7 describes the characteristics of each cluster and corresponding business strategies.

12.4 The Construction of Passenger Intelligence Applications

In order to better utilize big data techniques, airline companies need to start to develop some big data applications. This chapter introduces a multi-channel intelligence customer marketing platform for airlines.

The platform is developed to build a scalable, high-efficient 360-degree view of customers. The platform collects passenger's multi-channel data through every trigger point of each client, analyzes the consumer behavior and shopping habits

Table 12.10 Result of clustering analysis

Group	Number of days from the last order date to modeling	Sum of consumption	Sum of consumption	Average discount	Customer lifetime value	Percentage	Label
1	64.86	2.121	876.8	0.394	47.77	7.78	Low value
2	94.95	1.004	1580	0.873 (↑)	62.25 (↑)	6.25	Customer to develop
3	601.5 (↑)	2.086	1100	0.386	7.303	6.36	Low value
4	76.42	3.637 (↑)	2254 (↑)	0.485	85.96 (↑)	5.58	Promising
5	301.1 (↑)	2.081	1476	0.564	45.94	7.55	Low value
6	322.9 (↑)	2.105	870.4	0.363	18.68	8.78	Low value
7	96.87	1.024	748.5	0.873 (↑)	36.98	8.08	Low value
8	88.22	1	1087	0.628 (↑)	40.79	8.29	Low value
9	85.95	1.004	532	0.599	17.08	8.9	Low value
10	494.2 (↑)	2.19	2122 (↑)	0.773 (↑)	44.76	4.96	Low value
11	121.1	2.291 (↑)	2450 (↑)	0.815 (↑)	82.94 (↑)	8.38	Customer to retain
12	100.2	5.476 (↑)	4808 (↑)	0.674 (↑)	96.22 (↑)	4.89	Customer to develop
13	75.09	14.31 (↑)	126,694 (↑)	0.705 (↑)	99.65 (↑)	1.38	Customer to retain
14	176.1	1.997	4016 (↑)	1.665 (↑)	88.33 (↑)	0.9	Customer to retain
15	79.64	2.07	1526	0.588	70.83 (↑)	8.27	Low value
16	333.9 (↑)	3.827 (↑)	2664 (↑)	0.508	74.89 (↑)	3.64	Low value
Mean	188.3	2.255	1729	0.619	50.5		

(↑) indicates the value for the cluster is higher than the mean

of travelers to recommend personalized travel products. This application gathers data from internal and external sources including website, ticketing robots, GDS, loyalty, check-in, flight, marketing, CRM, social media, and industry databases. It identifies records relating to the same individual to produce a single customer view, then further enhances the customer profile with external data, segment codes, and recommended treatments. The results are available to execution systems to use for

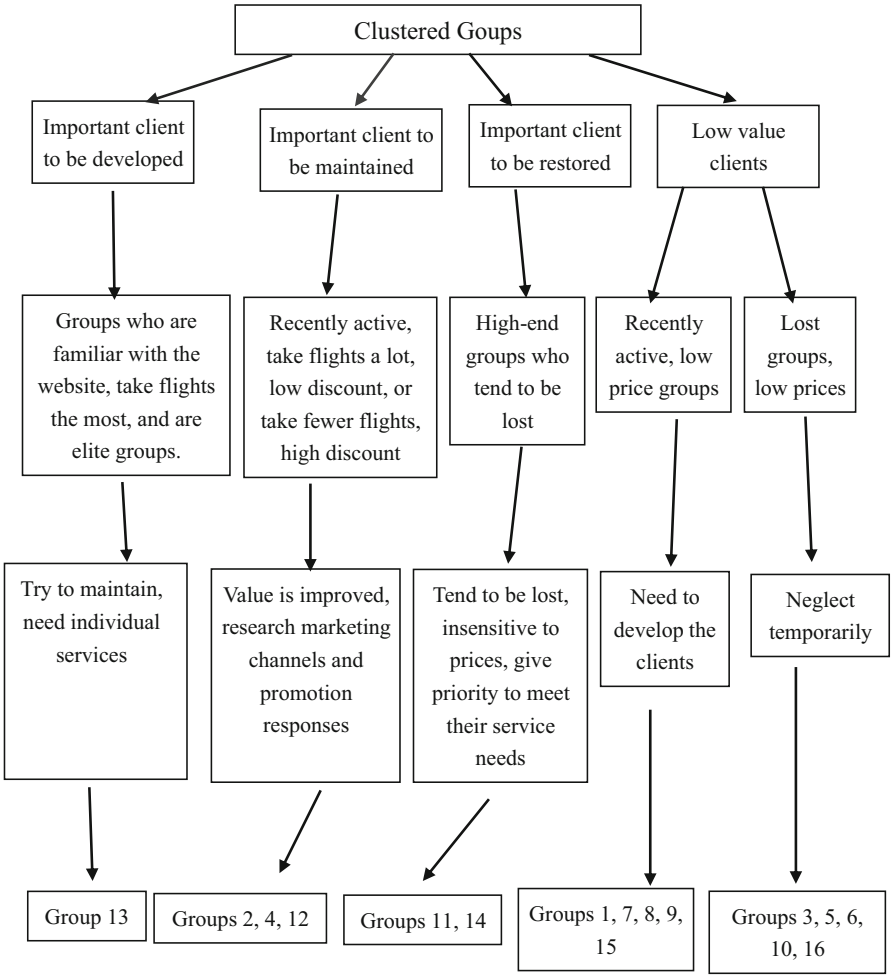

Fig. 12.7 The characteristics of each cluster and corresponding business strategies

personalization, offer and program selection, campaign execution, in-flight services and other purposes. The same data is also available for results reporting and other types of analysis.

This platform has two main benefits. First, it enables all systems to work from a complete, consistent customer view, enabling coordinated treatments across channels. Second, it saves each channel system from having to assemble its own detailed customer database. Traditional data warehouses require separate, expensive software to load and transform the data, to associate related records, and for modeling, analytics, and reporting. This all runs on expensive hardware and is operated by expensive IT staff and data scientists. These barriers have prevented all but the largest, most sophisticated B2B marketing organizations from building

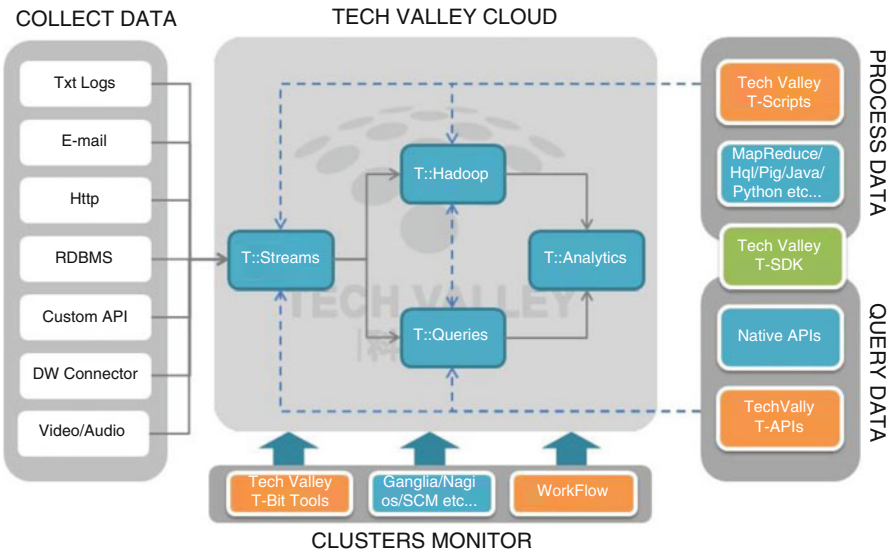

Fig. 12.8 Multi-channel intelligence customer marketing platform

a truly comprehensive central customer behavior and prediction database. In big data time, this is changing. Airlines and data solution providers cooperate with each other. Figure 12.8 presents an overview that shows how actionable insights can be derived out of big data. This platform has the following capabilities:

1. **Accept data from any source system.** This includes traditional, structured data such as purchase transactions right through to unstructured data such as web log files and contact center notes. The platform has an opt-in option to gather external data sources to enrich customer profiles.
2. **Allow access by external systems.** Upon request, the platform provides an Application Program Interface (API) that lets systems read its data during customer interactions.
3. **Associate different identifiers for the same individual.** The platform can chain together multiple identifiers: so if a Web system captures an email address and cookie ID, and the contact center captures email address and phone number, the system recognizes that the phone number, cookie ID, and email address all belong to the same person.
4. **Customer segmentation and customer behavior.** Behavioral segments can be created based on the aggregated customer data to combine transactional sights with shopping behaviors and social signals.
5. **Personalization and prediction.** The platform provides recommendations for customer treatment. The recommendations control which offers are considered and how they are prioritized and captures whether the recommended offer was actually presented and if it was accepted. This is controlled through an interface

Table 12.11 Different stages in the Big Data multi-channel intelligence customer marketing platform

Platform component		Source/products/technologies
Collect data	Structured data	CRM data, reservation system, call center logs, ERP, website logs,
	Unstructured data	Social media sites, blogs, location data, browsing behavior, mobile data, website content, enterprise data not recorded in CRM or ERP(e.g. marketing e-mail responses, survey results etc.), customer-employee interaction data, weather data, news, reviews etc.
Process data & query data (Technologies)		Hadoop (HDFS, Mapreduce), Cassandra, Hbase, Hive, Cognos, Hyperion
Big Data analytics		R, Sas, SiSense, Mahout, Datameer
Monitor		Mondrian, JGraphX, mxGraph, JavaScript Inforvis, Excel

that is accessible to non-technical users and captures whether the recommended offer was actually presented and if it was accepted. The information is used in reporting and to help the system make more accurate recommendations in the future.

Table 12.11 gives a few examples of the sources of structured and unstructured data, technologies and products that can be used for the different stages in the platform.

12.5 Conclusion and Implication

Using the data mining techniques discussed in this chapter, airline companies can learn important marketing implications of big data analytics. This chapter introduced Weibo customer value modeling, social network analysis, website click-stream analysis, customer activity analysis, clustering analysis, Recency-Frequency-Monetary (RFM) analysis, and principle component analysis, and the data mining results reveal passenger travel patterns, preference, travel social networks and other aspects of purchase behavior.

To help with formulating better business strategies, the airline companies may consider adoption of the following implications.

1. To set up an easier-to-use system: flight inquiry with high-level customization; ticket price calculation support with high-level customization, customized function design and product design, relevant product support, especially support for hotel products, and one-stop service should be provided to the greatest extent.
2. To track travel value chain closed loop further: from the whole process, such as inquiring products, reservation, payment, ticket issue, check-in, security check, waiting, cabin service, luggage claim, mileage accumulation, notice and start of next trip, record traveler's behavior details through contact of travelers with

the airlines, and provide prioritized individual services to important clients, for example, onboard seat preference or meal habit, etc.

3. Timely and effective client care, including SMS wishes, posted gifts, and so on. Weibo emotion analysis can be applied to judge if the attitude of a Weibo is positive, neutral or negative. It would be useful to extract information from the Weibo information and the social networking relationships (e.g., authentication information, number of fans and number of comments, etc.) of a passenger.
4. Individual online experiences. Priority can be given to important clients to be developed and maintained under the circumstance of limited resources, analyzing traveler's trip habit (behavior and preferences) to conduct cross-selling, identifying attractions to them based on user's preferences, filtering unnecessary information to present individual recommendations, and offering the most valuable product portfolio for clients during a specified time period. For instance, system analysis finds out that some clients purchase air tickets within a certain price range only. In this case, precision marketing can be applied to these clients, with SMS, mail, SNS and other means to keep them posted of product information.

References

1. Harteveldt HH (2012) The future of airline distribution: a look ahead to 2017. s.l.: special report commissioned by IATA
2. Davenport TH (2013) At the Big Data crossroads: turning towards a smarter travel experience. Available via AMADEUS. http://www.bigdata.amadeus.com/assets/pdf/Amadeus_Big_Data.pdf. Accessed 1 May 2014
3. Ghee R (2014) Top 5 in-flight trends to look out for in 2014. Available via http://www.futuretravelexperience.com/2014/01/top-5-flight-trends-look-2014/. Accessed 7 May 2014
4. Nicas J (2013) How airlines mine personal data in-flight. Available via http://www.wsj.com/articles/SB10001424052702304384104579139923818792360. Accessed 8 November 2013
5. Peveto A (2011) KLM surprise: How a little research earned 1,000,000 impressions on Twitter. Available via http://www.digett.com/2011/01/11/klm-surprise-how-little-research-earned-1000000-impressions-twitter. Accessed 11 January 2013
6. Chen J, Xiao YB, Liu XL, Chen YH (2006) Airline seat inventory control based on passenger choice behavior. Syst Eng Theory Pract 1:65–75
7. Schmittlein DC, Morrison DG, Colombo R (1987) Counting your customers: who are they and what will they do next? Manag Sci 33:1–24
8. Fader PS, Hardie BGS, Lee KL (2005) "Counting your customers" the easy way: an alternative to the Pareto/NBD model. Market Sci 24:275–284
9. Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. Addison Wesley, Upper Saddle River
10. Bult JR, Wansbeek T (1995) Optimal selection for direct mail. Market Sci 14:378–395
11. Khajvand M, Zolfaghar K, Ashoori S, Alizadeh S (2011) Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study. Procedia Comput Sci 3:57–63
12. Gupta S, Lehman DR (2003) Customers as assets. J Interact Mark 17(1):9–24