

AIRLINE DIGITAL CLICK STREAM EVENT PROCESSING FOR ENRICHING THE AIRLINE BUSINESS

Md. Alauddin

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia.

Ting Choo Yee

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia.

Ian Tan Kim Teck

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia.

E-mail: alauddinm@gmail.com

Recepción: 02/08/2019 **Aceptación:** 24/09/2019 **Publicación:** 06/11/2019

Citación sugerida:

Alauddin, M., Choo Yee, T. y Kim Teck, I. T. (2019). Airline digital click stream event processing for enriching the airline business. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Edición Especial, Noviembre 2019*, 287-305. doi: <http://dx.doi.org/10.17993/3ctecno.2019.specialissue3.287-305>

Suggested citation:

Alauddin, M., Choo Yee, T. & Kim Teck, I. T. (2019). Airline digital click stream event processing for enriching the airline business. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, November 2019*, 287-305. doi: <http://dx.doi.org/10.17993/3ctecno.2019.specialissue3.287-305>

ABSTRACT

The new era of digital world with the rapid expansion of social network and mobile applications created wider scope to expand airline industry for new way of promoting their business. Due to several social media and other digital platforms, we need to emphasize on target marketing/customer profiling. Hence, to do target marketing, a new web technology is created to collect each of the raw events of their web data and mobile app data for tracking the way user is searching flights. In the proposed method BigQuery is used to process huge volume of online customers' data. The proposed method is to understand the airline ecommerce online visitors effectively by analysing the event data stream collected from various digital properties. The obtained raw digital data consists of lot information with a semi-structured and it needs to be cleansed before analysing it. So, the first stage of proposed system is to extract the data from various digital sources in real-time, then chose which data is appropriate for analysing and finally extract the key insights to improve the airline business. From the extracted variables, search patterns, the predictive models such as flight search forecast, seat sales forecast and digital channel attribution models can be developed.

KEYWORDS

Click stream processing, Big Query, Digital data processing, digital marketing, Data Cleansing and Enrichment.

1. INTRODUCTION

In recent years, most of the Asian airlines prime focus is on digital transformation (O'Connell & Williams, 2005). The prime objectives of digital transformation are to understand the online customer acquisition, digital channel attribution, online customer segmentation, and their search trend. These are the most important techniques to take right business action at right time to increase revenue. Most of the airline industries have their own online and mobile based ecommerce platform, it is possible to track and record their activities on the webpage as from which webpage they have entered, when and what they search, where they drop off, what they purchase, how frequently they book etc., (Klein & Loebbecke, 2000). These visitor data can be for customer analytics like online customer profile, sales funnel to understand at which point visitors drop off, are they price sensitive or not.

However, tracking and processing visitors' raw events from the website logs data is complicated because of the large volume of hit level data (One of the major Asian airlines has about 15 million of online visitors per month, which generates roughly 3-5 billion events of unstructured or semi-structured web tracking data) (Ananthi, 2014). In this paper, the online digital click stream dataset is obtained from one of the major Asian Airline system with 50 destinations. Each route is tracked with one way and return flights for 30 days to 120 days. This paper mainly focus more on the real-time digital data collection and pre-processing of the dataset for flight sales prediction. The overall objective of the proposed work is that, the key variables are selected from the extracted digital click stream data is to improve the airline business.

2. LITERATURE REVIEW

The growth of Internet around the world made airline business to change their way of attracting the passengers (Singh & Jain, 2014). Also this digital era made to buy tickets from anywhere in the globe at any time by comparing the different airlines. So it is becoming very difficult to predict the ticket prices and attracting the passengers becoming difficult with the influence of many factors (Gillen & Lall, 2004). However, data science showed a way to progress in this type of scenarios to study the patterns

and predict the behavior of the sales outcome. For example, it can be identifying the correlation between seat prices of particular airlines and air traffic delays. As per recent surveys of (Forbes, 2008), it is noticed that for every minute of flight delay it will affect the ticket prices about \$1.5. Low cost airlines offer ticket pricing without the baggage, food and beverages, which gives privileges to afford all common people (Groves & Gini, 2013). Hofer, Windle and Dresner (2008) explained more details of the how low cost airlines are differ from the other airlines. Lazarev (2013) described in detailed how fare variations can be influenced in various time periods. Lazarev designed very good model to predict optimum prices for low cost airlines to generate almost 90% of the profit margin. In general, all the customers always think if earlier booking flight fares might be less prices.

Based on the various studies on the airline business, the most important aspects to buy tickets online in advance according to the user's observation and their risk (Etzioni, Tuchinda, Knoblock & Yates, 2003). The user who purchases their tickets online should have a sense of control over the task they are performing over the Internet. This helps to reduce the feeling of risk or fear associated with the possibility of: making a mistake when making an airline booking online (that is, psychological risk); not receiving their ticket or the flight not even existing (performance risk) (Brons, Pels, Nijkamp & Rietveld, 2002). Several research papers described the promotions on ticket prices, gift vouchers, airline points and upgrades, which playing indirectly to attract the customers (Barrett, 2004; Gillen, & Lall, 2004). The majority of these studies conclude that the incentives employed have a positive effect on airline ticket purchase and repeat purchase and highlight that the effectiveness of the program depends to a large extent on the particular incentive offered (Aviasales, n.d.). The literature regarding the choice of Airlines has made it clear that both the benefits provided by frequent flyer programs and air fares significantly affect user's choices (Groves & Gini, 2013). Users who travel for business perceive the frequent flyer programs as more useful than other users. These authors even guarantee that business travelers are willing to pay more in exchange for reducing access time, traveling with top-ranked airlines, and traveling in a better class (O'Connell & Williams, 2005; Sabre, 2015).

3. IMPLEMENTATION OF DIGITAL EVENT DATA PROCESSING

In recent years, most of the people in the world entered towards digital era, which increases the ecommerce transactions in a vast manner compared to the offline. Also the power of digital world made people to reach the world from anywhere any time through either social media, travel blogs or meta search engine. With these available resources, the traveller's can see different travel websites, travel blogs for price comparison before they book their flight tickets. This open lot of opportunity for the airlines to track the travellers search patterns and predict passengers' behaviour using predicting models. Besides, it is also possible to find which online channel is more effective for which airline routes and geo location for predicting the cost per acquisition, which in turn save lot of advertisement costs. Further, the successful tracking of all the digital data also enable the airlines to build sales funnel of digital products, customer life time value calculation and other predictive modelling for digital marketing.

3.1. DATA COLLECTION

To collect the online digital data and analyze its patterns, five types of variables are considered for better prediction of seat sales, which are:

- Visitor.
- Flight Search.
- Device.
- Channel.
- Transactions.

The transactional, operational data are extracted using various channels such as web, mobile and tablets in the year 2016. The collection of digital data in real-time is so complicated process, but with the evolution of Java scripts tagging framework, it is possible to track each web page and its components based on visitor status on the internet. The passenger activities such as which page they search, how much time

they spent on each webpage, how many clicks and scrolls on each page etc. Also, the ecommerce related information such as add to cart, product related information and ecommerce transaction details etc. As the flight sales digital web data is very big and complex, the data collected, cleansed and processed using cloud technology. The implementation of digital analytics will help marketing to monitor the load factor (%) for future flights and how traveler is choosing origin hub to destination hub and other connecting hubs using fly through (transit). Figure 1 shows the detailed block diagram of the airline data collection from various sources and its predictive model.

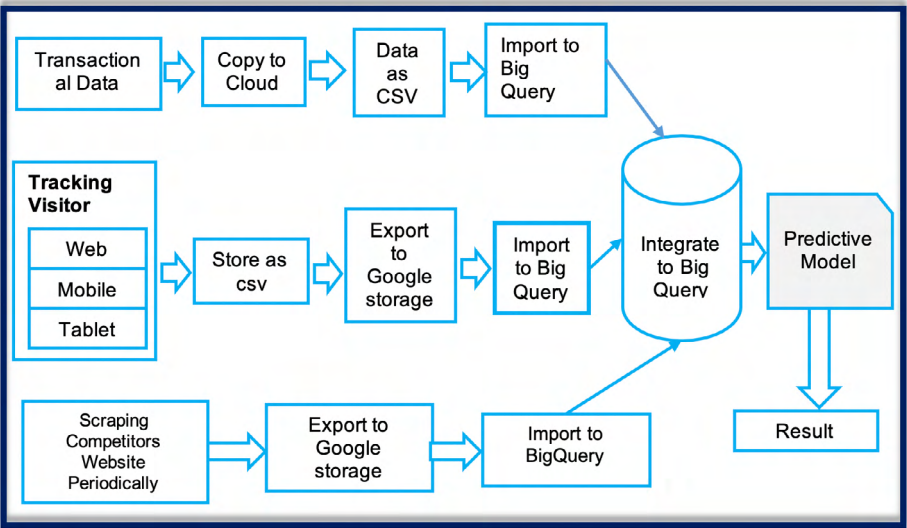


Figure 1. Airline digital data processing architecture.

Airline travel visitors search flight from different devices such as desktop, mobile devices and tablets. Therefore collecting the data from different devices is bit complicated, so it is necessary to consider each digital properties carefully. To collect the digital data (raw data) from various sources, a renowned tracking framework (*The java script which is modification of Google tracking framework*) is used. After collecting the data and tracking the gathered data, the user activity sends to the server for reporting and further analysis. The system uses different technologies to create data hits according to the types of digital properties. Hence, a new custom code is implemented for tracking web and mobile app users' activity. The proposed custom code also identifies the

new users and returning users, which provides the more information to fix the seat price dynamically. Finally, the custom code is implemented for capturing the business specific information such as Flight Search Origin, Flight Search Destination, and Departure Date etc. Also, the web server is tracked to receive HTTP request, which gives the details of the airline customers searching patterns. From the webserver log the customers details (such as, computer info, the Location, hostname, the browser type, and language they are browsing etc.,) are extracted.

In the proposed research, BigQuery is used to process high volume of customers' digital data. BigQuery is a RESTful web service that enables interactive analysis of massively large datasets working in conjunction with Google Storage. It is an Infrastructure as a Service (IaaS) that may be used complementarily with Map Reduce. BigQuery is used to process the raw data to further level. After exporting each digital properties as raw tables, which are available in BigQuery as multiple daily tables. BigQuery uses SQL syntax to process the raw data. Figure 2 shows the airline flight search data processing flow. Figure 3 shows the airline online traffic and search data processing flow from all airline digital properties in a daily aggregation. After tracking for capturing the web and mobile digital properties and the listed attributes, the captured data is exported to BigQuery on a periodic basis.

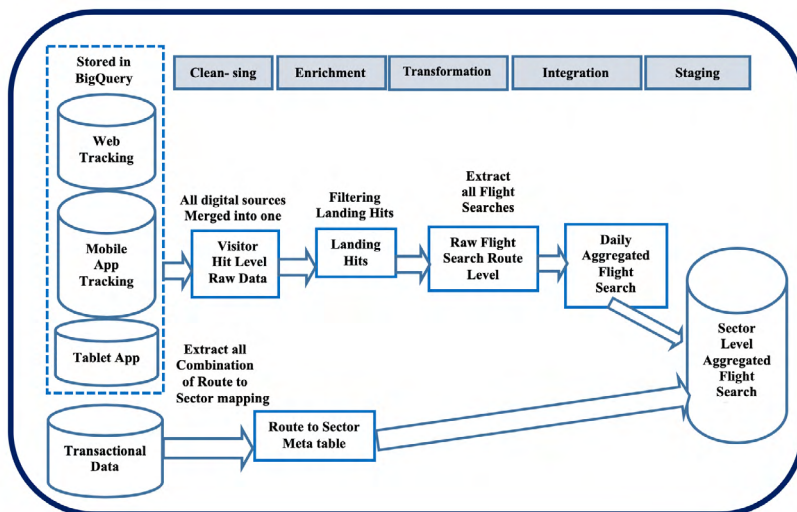


Figure 2. The block diagram of Airline Flight Search data processing flow.

In general, the open source tracking code retrieves web page data as follows:

- A browser requests a web page that contains the tracking code.
- A JavaScript Array is created and tracking commands are pushed onto the array.
- A <script> element is created and enabled for asynchronous loading (loading in the background).
- The ga.js tracking code is fetched, with the appropriate protocol automatically detected. Once the code is fetched and loaded, the commands on the array are executed and the array is transformed into a tracking object. Subsequent tracking calls are made directly to the server.
- Loads the script element to the DOM.
- After the tracking code collects data, the GIF request is sent to the analytics database for logging and post-processing.

A GIF request can be classified into few types. Table 1 shows various types of GIF request. In each of these cases, the GIF request is identified by type in the utmt parameter. In addition, the type of the request also determines which data is sent to the Analytics servers. For example, transaction and item data are only sent to the Analytics servers when a purchase is made. User, page, and system information is only sent when an event is recorded or when a page loads and the user-defined value is only sent when the _setVar method is called.

Table 1. GIF request types.

Request Type	Description	Class
Page	A web page on your server is requested	Interaction
Event	An event is triggered through Event Tracking that you set up on your site	Interaction
Transaction	A purchase transaction occurred on your site	Interaction
Item	Each item in a transaction is recorded with a GIF request	Interaction
Var	A custom user segment is set and triggered by a user	Non-interaction

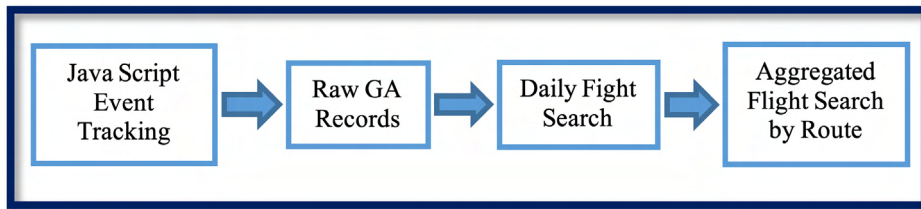


Figure 3. Airline online traffic and search data processing flow.

Raw web tracking data Processing: The volume of one year raw data is about six Terabyte. So, the first step used BQ SQL query to fetch the hits level data from BQ raw daily tables.

Data Cleansing and Enrichment: Raw tracking data have date format issues such as hit_timestmap in one format, the date extracted from page path URL has another format and custom dimension has different format. Therefore, all types of dates are converted in one standard format with same time zone. There are missing values of traffic information, flight search information, geo information, transaction information. To handle the missing values, first a metadata reference table have been created from other available attributes. Then the missing values are enriched using metadata tables. Also, the different digital properties captured with same information but different attributes name. Those need to be merged into one column. Figure 4 shows the BigQuery processing flow to predict the sector levels.

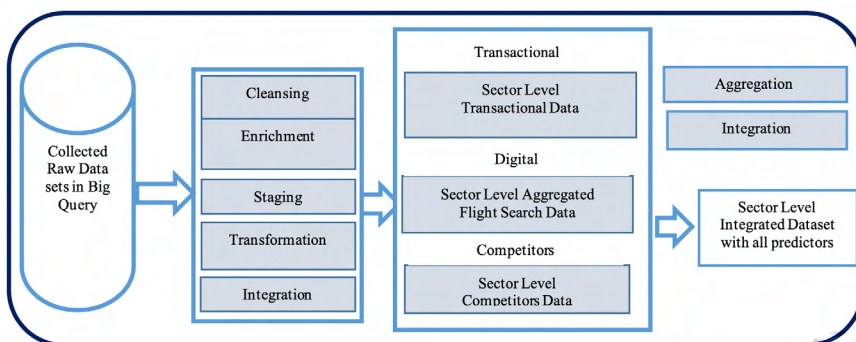


Figure 4. BigQuery processing flow to predict the different sectors data.

All digital data is stored in google storage which contains all hit level records of the visitor's clickstream. This data has been filtered to get the hits which give user interactions of flight search. The 'hit-type' filter has been set to 'EVENT' or 'PAGE'. This will filter out all interaction hits of flight-search page view or flight-search event action such as click on search button. At the same time, hit-type equals APPVIEW filters all the hits from Mobile and Tablet App (iOS and Android). These filters exclude lots of impression and other irrelevant hits records. It also helps to reduce the data volume that we must process in next stage. After that, the landing hits for the web has been filtered. This provides first hit and search hits only. It also ensures the exclusion of all other activity hits after searches such as passenger details page, add-on page, confirmation page and payment page. After that next challenge is to filter only search hit pages from the web, mobile and tablet a The search page will be identified by page-path mapping for web and screen name for mobile and tablet a However, there are different versions of web application and mobile app release with a website revamp and new version release for a mobile a Thus, the page-path and mobile screen name are not constant. To overcome this limitation all different search identifier, need to be collected for each release over time to create a reference mapping table. This table could be used to identify all search hits from all devices.

Sometimes customer searches non-operational flight route. Thus, all flight search has been examined to verify the searched route by the customer. All non-operational flight route searched by the customer have been filtered to avoid any misleading data for the final training set.

Digital data aggregation

With the clean, structured and quality data produced after data cleansing, enrichment and transformation, aggregation can now be performed to get desired data set. Algorithm 1 shows the high-level process of aggregating the digital data.

Algorithm 1: generate-aggregated digital data set

Input: $Union(D_{web}, D_{mobile}, D_{tablet})$

Output: $D_{uniqVisitorByRoute}, D_{uniqFlightSearchByRoute}, D_{NoOfFlightSearchByRoute}$

for d in (ClickStreamRecords) do

1. $search_timestamp \leftarrow$ Transform UNIX to timestamp (concat($D_{web}.visitStartTime$, date))
2. $visitID \leftarrow$ concat (sessionId, visitId)
3. $visitorID \leftarrow$ Extract ($D_{web}.fullVisitorId$)
4. $D_{FlightSearch} \leftarrow$ Extract (Max(CustomDimension.index, CustomDimension.value) by iterating each items))

end for

for d in ($D_{FlightSearch}$) do

$uniqueUsers \leftarrow$ COUNT (Distinct ($D_{FlightSearch}.visitorID$) by Routes)

$uniqueSearch \leftarrow$ COUNT (Distinct ($D_{FlightSearch}.sessionId$) by Routes)

$NoOfUsers \leftarrow$ COUNT ($D_{FlightSearch}.sessionId$)

end for

All digital platform (Web/Mobile/Tablet) data has been merged to make a one single data source. Since all the digital data are in the same structure, a UNION operation in BigQuery can merge multiple datasets of the same structure. This merged data table is named as 'clickStreamRecords'. Algorithm 3 takes this data as input. First step of the algorithm is to extract visitId and visitorId of the customer by hourly, daily, weekly and monthly basis and stored as $D_{FlightSearch}$.

After that, data has been aggregated to get the no. of flight, no. of unique user perform flight search and no. of total search as well as group by each selected route (origin and destination), search date and departure date. Furthermore, search-lead-days have been calculated by subtracting search-date from departure-date. This will compute how many days before the departure, customer searched for the flight. Output of this algorithm has been stored as $D_{uniqVisitorByRoute}$, $D_{uniqFlightSearchByRoute}$, and $D_{NoOfFlightSearchByRoute}$. Aggregated final dataset sample has been shown in Table 2.

Table 2. Sample of digital data.

Attributes name	Examples
fullVisitorId	1527445791
visitId	1527445791
SearchedOrigin	DXB
SearchedDestination	HKT
SearchedDepartureDate	2018-05-21
SearchReturnDate	2018-05-28
unique_search	6
NumberSearches	10

4. RESULTS AND DISCUSSION

The different datasets extracted from the total roll up are:

Visitor landing dataset with traffic source information: From which traffic source visitors performs the first hit at website and then what they do after landing to the website. Visitors can come from different types of online channel such as Paid Search, Organic Search, Paid Social, Meta Search, Direct etc. And after they renter into same airline webpage, it tracks the search flights as this increase the probability to purchase tickets. However, user might find irrelevant after landing to website hence drop off or visits web check-in, member sign off and other promotional pages.

Visitor Flight Search dataset: Fight Search dataset have multiple critical attributes such as Unique search visitors, Unique Search by Route, No of Total Search by Route, and other attributes

Ecommerce transaction dataset: which gives the money transactions on the seats bookings.

The reports produced from final stage of aggregated dataset is shown in Figure 5.

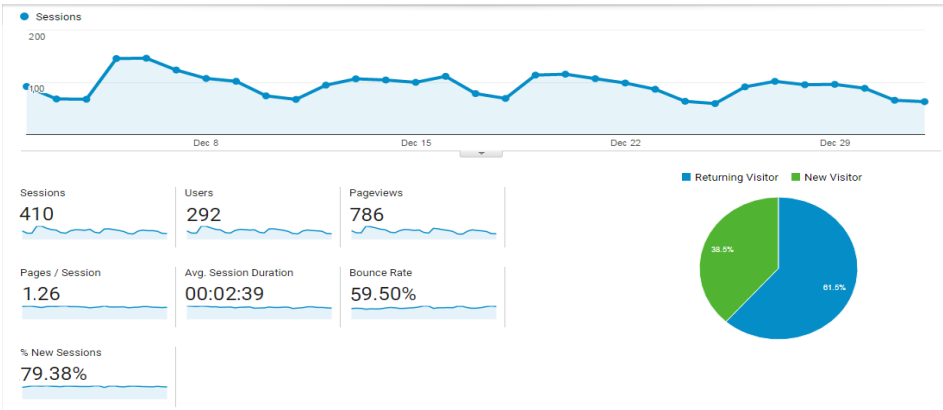


Figure 5. Digital airline Website tracking analysis report.

The digital airline website tracking analysis results shown in Figure 5 gives the summary of how many visitors visits each day and how many users log on to the website second time. Also, it shows how many numbers of sessions are in active, how long the user sessions were active. From these analysis, it is noticed that, based on the users searching patterns flight fares and seats could be decided. Few routes digital variable data have analyzed based on the seats sale using the correlation analysis and identified the best and worst routes, which is shown in Table 2.

Table 3. Correlation analysis results.

Worst Case Route		Best Case Route	
	Total seats sold		Total seats sold
Total seats sold	1.000	Total seats sold	1.000
Total unique visitors	0.179	Total unique visitors	0.576
Total unique search	0.216	Total unique search	0.609
Total number of search	0.242	Total number of search	0.572
Total unique sessions	0.198	Total unique sessions	0.593

From the Table 3 results, it is observed that, digital variables can be a strong descriptor in some routes for seat sales. This shows the potential value in including these digital variables into the model in addition to the obvious transactional variable and operational variables to get more information. It is clearly noticed that users meta and paid search rates are higher compared to the direct search rates. It is also showed that meta search rate is higher in booking also. From these analysis it is observed that, users meta search is using to book flight seats. From all the digital variable data, transactional data and operational data, the seat sales have predicted, which is shown in Figure 6.

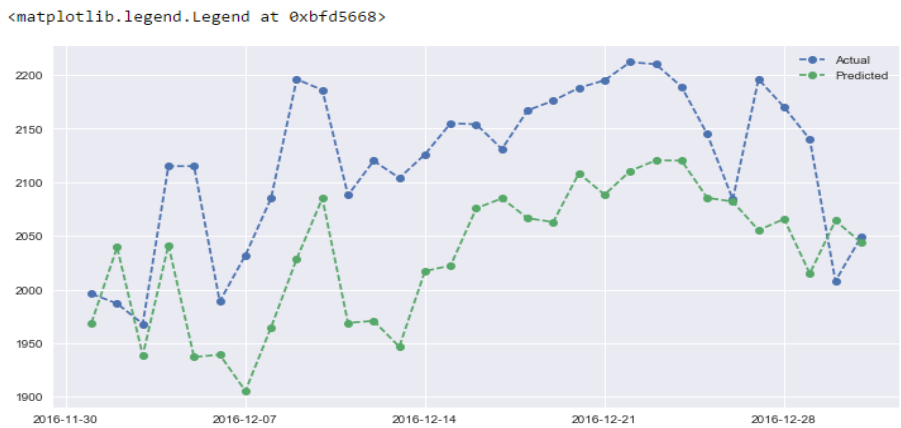


Figure 6. Seat sales forecast from various data variables data.

The forecast results showed in Figure 6 are part of the analysis. From the graph shown in Figure 6 that, the predicted values almost 6.5 to 9% deviation from the actual values. To predict accurately, hybrid models with ANN and ARIMA models are going to be implemented in further research works.

5. CONCLUSION AND FUTURE WORK

In this paper, the main approach used for selecting the important variables in flight sales forecast of each day on the route level. In this, for events tracking and web data tracking Java script is used. From the digital click stream data, the most prominent five selected variables were extracted to find visitors traffic, flight search transactions,

device data and channel data. These five selected variables data will be used to build models for predictive analytics such as Seat Sales Prediction, Revenue Optimization with Digital and Transaction data, Channel Attribution Model, Customer Life time value, which could bring tremendous business value. The proposed correlation analysis of the extracted variables, the model produced around 7% and 9% error rate when forecasting 30 days and 60 days ahead respectively. This paper discussed only the requirements and design constraints of the dynamic models. In our next paper, the dynamic predictive models will be described in detail with the suitable analysis results to predict the seat sales forecast dynamically according to the extracted real time digital data.

REFERENCES

- Singh, A. P., & Jain, R.C.** (2014). A Survey on Different Phases of Web Usage Mining for Anomaly User Behaviour Investigation. *International Journal of Emerging Trends & Technology in Computer Science*, 3(3), 70-75. Retrieved from: <https://www.ijettcs.org/Volume3Issue3/IJETTCS-2014-06-03-066.pdf>
- Ananthi, J.** (2014). A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites. *International Journal of Computer Science and Information Technologies*, 5(3), 4091-4094. Retrieved from: <https://www.semanticscholar.org/paper/A-Survey-Web-Content-Mining-Methods-and-for-from-Ananthi/36613a9e89bf3efda507568b54295b92f2f8ad23>
- Aviasales.** (n.d.). *Aviasales API*. Retrieved August 17, 2015, from: <http://www.aviasales.ru/API>
- Barrett, S. D.** (2004). How do the demands for airport services differ between full-service carriers and low-cost carriers? *Journal of Air Transport Management*, 10(1), 33-39. Retrieved from: https://www.academia.edu/2422292/How_do_the_demands_for_airport_services_differ_between_full-service_carriers_and_low-cost_carriers

- Brons, M. R., Pels, E., Nijkamp, P., & Rietveld, P.** (2002). Price elasticities of demand for passenger air travel: a meta-analysis. *Journal of Air Transport Management*, 8(3), 165-175. doi: [https://doi.org/10.1016/S0969-6997\(01\)00050-3](https://doi.org/10.1016/S0969-6997(01)00050-3)
- Etzioni, O., Tuchinda, R., Knoblock, C. A., & Yates, A.** (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 119-128. Retrieved from: <https://www.isi.edu/integration/papers/etzioni03-kdd.pdf>
- Forbes, S. J.** (2008). The effect of air traffic delays on airline prices. *International journal of industrial organization*, 26(5), 1218-1232. Retrieved from: http://recherche.enac.fr/~steve.lawford/teaching_papers/econometrics2_2019/papers/forbes08.pdf
- Francis, G., Fidato, A., & Humphreys, I.** (2003). Airport-airline interaction: the impact of low-cost carriers on two European airports. *Journal of Air Transport Management*, 9(4), 267-273. doi: [https://doi.org/10.1016/S0969-6997\(03\)00004-8](https://doi.org/10.1016/S0969-6997(03)00004-8)
- Gillen, D., & Lall, A.** (2004). Competitive advantage of low-cost carriers: some implications for airports. *Journal of Air Transport Management*, 10(1), 41-50. doi: <https://doi.org/10.1016/j.jairtraman.2003.10.009>
- Groves, W., & Gini, M.** (2013). An agent for optimizing airline ticket purchasing. *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 1341-1342. Retrieved from: https://www.researchgate.net/publication/262172314_An_agent_for_optimizing_airline_ticket_purchasing
- Hofer, C., Windle, R. J., & Dresner, M. E.** (2008). Price premiums and low cost carrier competition. *Transportation Research Part E: Logistics and Transportation Review*, 44(5), 864-882. doi: <https://doi.org/10.1016/j.tre.2007.03.004>
- Klein, S., & Loebbecke, C.** (2000). The transformation of pricing models on the web: examples from the airline industry. *13th International Bled Electronic Commerce Conference*, 19-21. Retrieved from: <https://pdfs.semanticscholar.org/3112/430cbe6730d9c3eb559b83ab373bb913ec19.pdf>

- Lazarev, J.** (2013). The welfare effects of intertemporal price discrimination: an empirical analysis of airline pricing in US monopoly markets. Retrieved from: http://www.johnlazarev.com/Lazarev_JMP.pdf
- O'Connell, J. F., & Williams, G.** (2005). Passengers' perceptions of low cost airlines and full service carriers: A case study involving Ryanair, Aer Lingus, Air Asia and Malaysia Airlines. *Journal of air transport management*, 11(4), 259-272. doi: <https://dx.doi.org/10.1016/j.jairtraman.2005.01.007>
- Sabre. (2015). *APIs. Get travel information on demand with our REST and SOAP APIs.* Retrieved August 17, 2015, from: <https://developer.sabre.com/docs>

AUTHORS BIOGRAPHY



Md. **Alauddin** is currently Masters Student in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. He is a Computer Science and Engineering Graduate from Khulna University of Engineering and Technology with major of Software Engineering. His research interest mostly on BigData, Machine Learning and Data Engineering.



Dr. **Choo-Yee Ting** is currently holding Associate Professor in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. In the year 2002, Choo-Yee Ting is awarded the Fellow of Microsoft Research by Microsoft Research Asia, Beijing, China. In 2003, he received research fellowship from Rotary Research Foundation, Rotary Club of Kuala Lumpur Diraja, Malaysia. He has been involving himself in research projects funded by MOSTI, Malaysia and Industries. He is also certified in Microsoft Technology Associate (Database) and IBM DB2 CDA.



Dr. **Ian Tan Kim Teck** is currently holding senior lecturer in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. **Ian Tan Kim Teck** is graduated with a Doctor of Philosophy (Ph.D.) from Multimedia University, Malaysia in the area of Operating Systems' schedulers. He did his Master of Science Degree in Parallel Computers and Computation, from University of Warwick, United Kingdom in 1993 and a Bachelor of Engineering Degree and Associate of City and Guilds Institute in Information Systems Engineering, from Imperial College London, United Kingdom in 1992. He is also Novell Certified Linux Administrator (NCLA), Novell Certified Linux Professional (NCLP), member of IEEE and member of ACM. His area of research interest is primarily in systems; from operating systems process scheduling on multicore systems, efficient network data transfers, to systems and network security.

