Practice Article

Big Data in the travel marketplace

Received (in revised form): 31st August 2015

Ben Vinod

Sabre Research, Southlake, USA

Ben Vinod serves as Chief Scientist and Senior Vice President at Sabre. Before rejoining Sabre in 2004, he was Vice President of Sabre Airline Solutions; responsible for pricing and yield management.

Correspondence: Ben Vinod, Sabre Research, 3150 Sabre Drive, Southlake, TX 76092, USA

ABSTRACT We are beginning to see that Big Data will have a profound impact on gaining consumer insights, improving process efficiencies and enhancing the consumer experience. In the travel industry, travel suppliers, Online Travel Agencies and Global Distribution Systems have access to vast amounts of data from across the travel value chain – marketing and lead generation, interactive selling, fulfillment and customer care. Big Data can offer unique insights into consumer preferences and behavior patterns to improve conversion rates and revenues. This article focuses on the role of Big Data, the skills required in an organization to leverage Big Data in travel followed with examples of Big Data applications related to travel as it applies to suppliers, online and traditional travel agencies.

Journal of Revenue and Pricing Management (2016) **15**, 352–359. doi: 10.1057/rpm.2016.30; published online 3 June 2016

Keywords: Big Data; pricing; revenue management; air shopping; online travel agencies; global distribution systems

Data is a key corporate asset in a range of industries. In the travel industry, travel suppliers (for example, airlines, hotels, rental car, cruise lines and so on), Online Travel Agencies (OTAs) and Global Distribution Systems (GDSs) operate at the intersection of travel and digital technology. They have access to an unprecedented volume of data captured across the travel value chain from marketing and lead generation, interactive selling, fulfillment and customer care. Yet, these entities capture, store and leverage this data for competitive advantage to only a limited extent. This vast amount of data can be used to provide unique insights into consumer preferences and behavior patterns to improve conversion rates and improve revenues. This data can also be leveraged to improve customer care and provide customers with an enhanced level of service. In today's digital world, entities that play a role in travel are awash with galloping growth in the volume of raw data that needs to be captured and harnessed, which can easily run into Terabytes and Petabytes. Big Data is the term used to describe the data that can be typically be hundreds of Terabytes or Petabytes (10^{15}) in size. The data that is being collected grows very quickly. For example, in 2015, at Sabre we generated 44 terabytes of shopping transaction data every day and this data is growing at an increasing rate, with a year over year increase of 100–120 per cent. Air shopping volumes have outpaced bookings over the past decade.

This article provides an overview of Big Data in travel, the importance for economically storing vast amounts of data, data extraction and highlights several applications that leverage Big

Data for competitive advantage. A traditional treatment of Big Data – handling outliers, missing data, multicollinearity (correlated predictor variables) and comparison of specific statistical or machine learning techniques and algorithms (Dumancas and Bello, 2015) to solve specific problems is not in the scope of this article.

Big Data consists of structured data and unstructured data. Examples of structured data are booking and ticketing transactions, post departure data and so on. Examples of unstructured data include user generated content from hotel reviews, posts on social media sites, sensor data, audio, video, click streams and log files. Insights into consumer behavior, process efficiencies and Website design can be enhanced when these different types of data are analyzed together. However Big Data is more than just handling the exponential growth in volume of the data; it encapsulates the tools that can be used to process this data efficiently, gain insights into the business and make a corporation more agile. It is data like click streams, travel reviews and social media, that are highly unstructured and infeasible to be stored and processed in a Relational Database Management System (RDBMS) such as Oracle, DB/2 or Teradata. For example, the Twitter application program interfaces (API) could be used to capture terabytes of travel-related tweets each day for sentiment analysis, trend detection, lead generation, identify hotspots to trigger aggressive inventory controls and serve as an impulse signal for demand forecasting. When an enterprise data strategy is established and steps are taken to capture data from operational systems to store in a data warehouse for analysis purposes, a key *first* question that needs to be addressed is the approach that should be taken to ensure that the framework should allow for the storage of any data type in a low cost, scalable environment that reduces the cost of processing massive volumes of data. So why collect all this data? It is a basic requirement for innovation along three dimensions – product innovation, customer innovation and data science innovation. Not too long ago, we thought we needed a cost-effective way to process large and complex data. But this remains merely a technology initiative unless it's tied to business goals and objectives. Data is all around us, in massive volumes and increasing velocity from a multiplicity of sources. These can be seen as interactions in our daily life, which is at the core of our emergence as a data-driven society. Traveler targeted technologies ranging from smartphone applications to wearable computing are capable of flooding the space with a large volume of rich data sets. Data allows the enterprise to get insights, and insights lead to finding new opportunities and doing business differently (Vinod, 2013). It provides the ability to trigger relevant targeted offers, cross-selling and upselling, variety of services including ground transportation, restaurants and events.

AVERAGE DATA FOOTPRINT

Over the course of a year everyone leaves behind a digital footprint that serves as the core for Big Data analytics. This digital footprint (Demir, 2010) is made up of all the data a person would generate in their day-to-day activity, for example, retail purchases, online banking, communications, surfing the Web, Social Media and of course any travel data.

The average person's data footprint per year is 1 Terabyte (based on an approximate global population of about 7 billion people). The digital trace generated by the average person on a daily basis was about 45 gigabytes (GB) in 2008. This includes private information such as emails, photos, VOIP calls and instant messages and today it is even greater. A recent study reveals some figures about the size of the digital universe as 281 billion GB for 2008 and 1.8 billion terabytes for 2011 (Gantz, 2008). How about passive digital traces we leave behind by credit card purchases, bank accounts, phone records, web searches, general backup data, medical and hospital records, surveillance cameras and so on? There are more passive traces collected than our active digital traces, which provide more personal information.

To visualize what we are talking about, if 1 terabyte of information was printed on A4 paper, it would be 6.5 km long and it would be comparable to a plane's cruising altitude.

TRAVELER'S DIGITAL FOOTPRINT

When it comes to the travel and tourism industry, we believe that active travelers contribute to and generate significantly more than the average person's footprint. In fact, we believe travelers generate a disproportionately higher volume of all data produced in the world.

Consider, for example, what active travelers produce in every stage of their trip. When shopping and booking their travel through OTAs, agencies and directly with suppliers, online and mobile. Or when checking-in for a flight, hotel, car or a train, travelers leave behind a digital footprint. The question is – *how can this data be leveraged to better understand and service our customers?*

BIG DATA BASICS

Volume, Velocity and *Variety* are the three dimensions of Big Data as defined by the MetaGroup's Doug Laney (now part of Gartner) in a MetaGroup Research publication (Laney, 2001).

Volume refers to the amount of data, which has been growing at an increasing rate. Travel shopping volumes have grown at an increasing rate with an annual growth rate in excess of 100 per cent! At Sabre, from consumers across the globe, we process ~129 million shopping queries per day and 1900 shopping queries per second at peak times. At Sabre, we forecast 40 billion shops in 2015. Each shopping response may have anywhere from 10 to 100's of itineraries that is about 44 terabytes a day.

Velocity refers to the speed with which data is collected and processed. Since its inception in the 1980s, revenue management systems were batch-oriented and processed data captured nightly from the reservations system. To unlock added value from the revenue management process, these systems have transitioned to process streaming data such as bookings and inventory alert messages to adapt to real time changes in the marketplace and ensure that the inventory controls are based on up-to-date information. Another example is dynamic intervention, where streaming data is leveraged by OTAs to promote offers based on the number of times a customer has visited the site.

Variety refers to the various types of data such as text, audio, video, sensor data, documents, geo-spatial data from satellites and structured data that may be required to be processed using specialized techniques. Traveler targeted technologies ranging from smartphone apps to wearable computing are capable of flooding the space with a large volume of rich data sets containing personalized information. Google has worked with the transportation departments at local, state and federal levels who have begun installing solar-powered traffic sensors on major roads to determine traffic conditions. In some cases transmitting rich data from an aircraft to the ground can be prohibitively expensive. In the context of airline safety, commercial airlines transmit some information: radio transponders identify them when scanned by radar and are fitted with ACARS (Aircraft Communications Addressing and Reporting System) that periodically relay text messages about the status of the aircraft in flight. The case of the missing Malaysia Airlines flight MH370 begs the question – why was rich flight performance and pilot voice communications data not transmitted to the ground? First, transmitting data continuously through satellites is expensive and second, pilot union contracts may not allow it. Kavi (2010) suggests some combination of encryption and privacy policies like those for medical records may be sufficient to overcome their objections. If this data was made available it could be mined in real time for abnormal behaviors and a disaster such as MH370 could have been averted. Another example is location-based services with iBeacons (Danova, 2014) that

allows iOS applications to receive location-aware notifications over Bluetooth technology. It can be used to understand how people enter and exit into specific locations and promote offers at airports, hotels, retail stores, stadiums and so on.

Of the three V's defined by Laney, the biggest challenge is *Variety* – since today most entities only capture a limited amount of data and new data sources in new formats are constantly becoming available. Steve Todd, an EMC Fellow, offered the following definition: 'Big Data is when the normal application of current technology does not enable users to obtain timely, cost-effective and quality answers to data-driven questions'.

There are additional V's that have been proposed by data scientists.

Veracity describes how accurate the data is for predictive analytics. An example of Veracity in travel is the usage of organic shopping data for predictive analytics. The shopping data used must not include robotic shops which in some cases can make up more than 20 per cent of the volume of shops. Hence pattern recognition algorithms need to be devised to detect robotic shops and eliminate them from the data before developing predictive models.

Variability describes the change in the values of the attributes typical of a large data set. Variability is very relevant for sentiment analysis as the context is important to understand how a word is being used.

Visualization is the art of displaying complex multi-dimensional data to make it comprehensible to an analyst. There are many versatile open source visualization tools like D3 Data-Driven Documents that are readily available, for example, but the challenge lies in the interpretation and the ability to explain data insights in simple words.

Value addresses the need for valuation of enterprise data based on its relative importance. Value is of significance as not all data are created equal and different types of data have different potential for monetization.

UNDERSTANDING BIG DATA

The biggest asset in an organization is the individuals, the data scientists who are passionate and with a mindset to mine the data and discover hidden signatures in the underlying data sets. It also requires close collaboration between computer scientists, operations research and econometricians. Varian (2014) provides an excellent overview of analytic tools required for data scientists.

To understand and work with Big Data, first is the ability to work with extremely large data sets that requires a working knowledge of NoSQL, which is more primitive than SQL databases but can handle larger volumes of data. Also requires is a working knowledge of the Hadoop Distributed File System (HDFS), MapReduce and Pig Latin, the language to create MapReduce jobs, for massive parallel data processing.

Google developed a proprietary, distributed file system called Google File System (Ghemawat *et al*, 2003) and a parallel programming technique and framework called MapReduce for its web search purposes (Dean and Ghemawat, 2004). Hadoop was derived from these papers. Its core components are MapReduce and the HDFS . The Google papers inspired Doug Cutting to create the Java-based Hadoop, which he named after his son's toy elephant. Hadoop and a few open source tools that complement it make huge, diverse data sets readily accessible for quick analysis using clusters of inexpensive commodity hardware. Today, Hadoop is an open source revolution, considered mission critical and is widely used by the US Government, the National Security Agency and Web giants such as Facebook, Twitter and Yahoo. Over the recent years, Hadoop has evolved into an ecosystem with many sub-projects such as Pig, Hive, Hbase and so on.

The creators of Hive wanted to make Hadoop easier to use, by making it look like a relational database. Hive translates a SQL-like query, called Hive QL, into MapReduce jobs. However, this has two consequences. First, it gives users the impression that they can use

Hadoop in the same way they use any RDBMS. Users would unknowingly submit queries that are very inefficient in the MapReduce framework and end up with very slow jobs. They would also be frustrated by the limitations Hive has. For example, Hive is not very good at optimizing queries, and some of the advanced table-joining queries cannot be done in Hive. Second, Hive tables are commonly stored in avro format in HDFS. This results in *data duplication*. We end up taking twice as much space as we actually need. In addition, the process of converting the raw XML data into Hive tables consumes CPU resources on a daily basis.

Under the hood, Hadoop is *not* a relational database and this is what makes Hadoop and Big Data revolutionary. It is important to understand that to leverage the power of Hadoop, there is not a one-size-fits-all solution. This is true in any organization. We cannot simply convert the raw data into Hive tables and hope everybody can immediately use Hadoop like they use a RDBMS or Teradata.

As an alternative to Hive, to address the reduction of data for calibration and analysis when dealing with large data sets, a library of User Defined Functions (UDFs) can be developed and used to extract the pertinent data elements from the HDFS. For example, for air shopping, each row is a very long XML string consisting of the shopping request and the shopping response – which can be 10 or 1000 itineraries. The UDFs are at an atomic level – each UDF is able to parse a single field – examples of UDFs are request origin, request destination and so on. The UDFs are written in Java to parse the raw XML shopping request/shopping response data. With a library of UDFs a user does not have to parse the raw XML files. Instead they can efficiently extract data from the HDFS. The UDFs also provide flexibility – based on the analysis at hand, an user can select the fields (UDFs) that should be included for the data extract.

A recent alternative to disk-based MapReduce is Apache Spark, an open source cluster computing framework built upon an in-memory paradigm. It allows a user to load large data sets into the memory of each node, and hence provide much better performance than disk-based MapReduce. At its core is the Resilient Distributed Datasets (RDDs), which is analogous to files in HDFS, except that RDDs stay in memory. This allows iterative algorithms to run on large data sets with much better performance than MapReduce. With more recent development efforts focused on Spark API's, Data Frames are introduced to R, Scala and Python to make Spark much easier to use. This is useful for data scientists because it opens up many opportunities for new models that were previously infeasible to calibrate in MapReduce.

TOOLS FOR DATA SCIENTISTS

In the area of Predictive Analytics, individuals with skills in one or more of the following areas are required: classical statistics, Bayesian statistics, machine learning. Machine learning is important when we are dealing with large volumes of data and supports non-linearity.

When faced with a prediction problem, an econometrician would typically think of modeling it as a regression or logistic regression model. However, when we are dealing with vast amounts of data, there may be better choices that include non-linear methods. Well-known machine learning techniques such as Reinforcement Learning, CART (Classification & Regression Trees), Neural Networks, Random Forests and so on can be used with the objective of deploying the right technique that will give *consistent out-of-sample predictions*.

BIG DATA APPLICATIONS IN TRAVEL LEVERAGING SHOPPING DATA

In the travel industry, the question remains – what types of business problems can a Big Data platform solve? Some applications and their value propositions are summarized below Vinod (2015).

Dream and plan capability for trip planning

Airline shopping is probably the most computationally complex application in the travel industry. It consists of four components that are required to display a set of itineraries that maximize conversion rates – generation of outbound and inbound schedules, seat availability by point of sale, algorithm for selecting the best set of itineraries and the ability to price these itineraries taking into consideration fares, rules, foot notes and routings. The dream and plan capability requires a shopping cache to answer complex queries to find destinations with specific attributes on a budget or determine a lead price to a destination on a 90-day calendar for a fixed length of stay. Central to the development of these new services is the deployment of simplified Representational State Transfer APIs as the information (state) for read requests is identified by the URL of the service. Hence edge caching can be used to eliminate network latency and shorten response times. SOAP (Simple Object Access Protocol) requests, which are used by traditional web services, cannot be cached. Hence, by allowing this information to flow forward and be held close to the users – for example, an air shopping cache – we can drive down the overall transaction costs and provide consistently fast response times worldwide.

Optimizing screen displays

Diversity of itineraries displayed during shopping that will resonate with customers is of critical importance to maximize conversion rates. Shopping diversity algorithms address different dimensions such as guaranteeing a minimum number of inbounds for every outbound, the quality of service expressed by number of non-stops, single-connect and double-connect itineraries, carrier diversity, fare and so on. Choice models can be calibrated to determine the utility of schedule and fare attributes to determine the best set of diverse itineraries to display as part of a shopping response. Measuring the screen quality enables OTAs to determine how effectively they convert shoppers to bookers (Rao and Smith, 2005; Vinod, 2011). On supplier sites, optimizing the screen display for each shopping request can re-direct demand from high load factor flights to low load factor flights, thereby reducing passenger displacement to generate incremental bookings and revenues.

Targeted display algorithms for air shopping based on trip attributes

Display algorithms are a critical component of any shopping response for OTAs and GDSs since the optimal display of the right itineraries maximizes conversion rates. These shopping display algorithms rely on using the output from a shopping request inclusive of itinerary diversity as input to determine the relevant itineraries to display on the target device – desktop, tablet or mobile. A display algorithm based on trip characteristics is based on a traveler's relative importance of schedule and fare attributes. Traditional travel Website filters cannot solve this problem since a filter would exclude an itinerary based on one attribute, even though it would have been outweighed by the goodness in the other attributes. An effect method is the use of TOPSIS, a Technique for Ordering Preferences by Similarity to Ideal Solution, which ranks itineraries based on a traveler's relative trade-off between schedule and fare attributes.

Fare forecasting

Predicting when air fares would go up or down is difficult since several factors need to be considered such as inventory control recommendations from revenue management, response to competitor fare and rate changes, as well as promotional fares. Air shopping data is an ideal source for developing machine learning algorithms, an artificial intelligence technique, to predict when fares would increase and when they would decrease. A 'buy' recommendation

means fares are expected to increase and a 'wait' recommendation means that fares are expected to go down. Reinforcement learning techniques are very effective in prediction. With this approach, there is an 'agent' to classify the recommendation as right or wrong and the model learns and improves the accuracy of the prediction over time (Vinod, 2013).

Similar techniques can also be applied to hotels to predict when hotel rates are expected to go up or down by star rating and neighborhood.

Competitive revenue management with dynamic availability

This is competitive revenue management by origin and destination to determine how existing inventory control recommendations need to be modified based on prevailing competitive market conditions (Ratliff and Vinod, 2005). This requires monitoring of competitive selling fares from shopping responses to determine the optimal inventory controls. To determine the attractiveness of each itinerary in a shopping response requires a choice model to be calibrated from a shopping request and response data extract with pertinent variables such as displacement time, elapsed time, fare, screen presence and so on.

Competitive revenue management with dynamic pricing

Dynamic pricing is closely related to dynamic availability. Both techniques leverage competitive selling fares to arrive at an inventory control or dynamic price recommendation. The session-based fare optimizer already determines the optimal price point for the host airline based on the competitive set and current selling fares of competing airlines in the marketplace. Instead of converting the optimal price point to an inventory control recommendation, the dynamic price is used to approximate the ticketed price. While dynamic pricing of opaque travel products (Zouaoui and Rao, 2009) has minimal business process changes, the potential impacts of deploying dynamic pricing on third-party systems can be quite significant (Choubert *et al*, 2015).

Optimal markup of net fares

Net fares are fares negotiated between an airline and a travel agency. What the agency owes to the airline is the negotiated net fare and the agency has the opportunity to markup the fares for sale. While net fares are less prevalent in North America, they are more prevalent in Asia Pacific and the Middle East. Shopping data can be used to calibrate a choice model to determine the optimal markups based on prevailing market conditions.

Pricing opportunity model

A pricing opportunity model helps airline pricing analysts evaluate their past performance. In hindsight, what would the correct market pricing have been for a specific carrier, considering its competitors' prices and its own available capacity? Most pricing analysts currently focus their competitive monitoring efforts on filed ATPCO fares without considering actual low fare search results – seat availability, carrier and schedule quality of service. This approach can systematically detect persistent recent patterns of under or overpricing across markets and provides pricing analysts with important new insights; by comparing future situations with the recent past, they can see if the recent trends still hold true for current or future dates.

Marketplace transparency – Analytics and actionable insights

The marketplaces created by GDSs bring buyers and suppliers together on a global scale. For example, Sabre transactions US$300 million a day through the GDS marketplace. Marketplace analytics applications allow airlines to better

understand the dynamics of the marketplace and identify opportunities to grow business by integrating shopping activity and travel demand patterns and automatically alerts suppliers to opportunities to improve or grow their revenues in the marketplace (Sabre, 2015). With more than 2 million markets in the marketplace, it helps convert demand to bookings by answering key questions such as: *What are travelers shopping for? When are travelers shopping? Where is the travel demand coming from? How are my products performing?*

CONCLUSIONS

The age of Big Data in travel has arrived. It is anticipated that we will see various new value propositions in travel that leverage these new data sources. Today we are merely scratching the surface to provide insights into customer behavior patterns and make internal operations of enterprises more efficient. The role of the data scientist in organizations will grow in importance, and the underlying scrutiny and challenge will lie in estimating and documenting the return on investment in small incremental steps for investing in a data infrastructure.

ACKNOWLEDGEMENTS

This article is based on a presentation the author made at the AGIFORS Revenue Management Conference in Shanghai, 14–15 May 2015. The author thanks Sunny Ja and Tassio Carvalho for having provided the opportunity to present a key new enabler for pricing and revenue management, and travel in general, that may not be foremost on the radar of revenue management practitioners.

REFERENCES

Choubert, L., Fiig, T. and Viale, V. (2015) Amadeus Dynamic Pricing, AGIFORS Revenue Management Conference presentation, 15 May, Shanghai, China.

Danova, T. (2014) Beacons: What they are, how they work, and why Apple's iBeacon Technology is ahead of the pack, Business Insider, 23 October.

Dean, J. and Ghemawat, S. (2004) MapReduce: Simplified Data Processing on Large Clusters, Sixth Symposium on Operating System Design and Implementation, OSDI, Vol. 6, San Francisco, CA, December.

Demir, H.I. (2010) Our unique digital footprint, *The Fountain Magazine*, Issue 77, September–October.

Dumancas, G.G. and Bello, G.A. (2015) Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high-performance data mining, http://sc15.supercomputing.org/sites/all/themes/SC15images/tech_poster/poster_files/post111s2-file3.pdf, accessed 15 January 2016.

Gantz, J.F. (2008) 'The Diverse and Exploding Digital Universe' an IDC White Paper sponsored by EMC, Framingham, MA 01701.

Ghemawat, S., Gobioff, H. and Leung, S.-T. (2003) The Google File System, 19th ACM Symposium on Operating Systems Principles, Lake George, New York, October.

Kavi, K.M. (2010) Beyond the black box, *IEEE Spectrum*, August, pp. 46–51.

Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity, Variety, Application Delivery Strategies, META Group, Stamford, Connecticut, 6 February.

Rao, B.V. and Smith, B.C. (2005) Decision support in online travel retailing. *Journal of Revenue and Pricing Management* 5(1): 72–80.

Ratliff, R.M. and Vinod, B. (2005) Airline pricing and revenue management: A future outlook. *Journal of Revenue and Pricing Management* 4(3): 302–307.

Sabre Corporation (2015) Sabre introduces Sabre Marketplace Analytics – A data analytics solution to help airlines capitalize on travel demand, PR Newswire, 24 March.

Varian, H. (2014) Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28(2): 2–28.

Vinod, B. (2011) The future of online travel. *Journal of Revenue and Pricing Management* 10(1): 56–61.

Vinod, B. (2013) Leveraging big data for competitive advantage in travel. *Journal of Revenue and Pricing Management* 12(1): 96–100.

Vinod, B. (2015) Leveraging Big Data in the Travel Marketplace, AGIFORS Revenue Management and Second Chinese Airline Revenue Management Summit Joint Conference, 13–15 May, Shanghai, China.

Zouaoui, F. and Rao, B.V. (2009) Dynamic pricing of opaque airline tickets. *Journal of Revenue and Pricing Management* 8(2/3): 148–154.