Big Data in SATA Airline finding new solutions for old problems

Armando B. Mendes, Hélia Guerra, Luís Gomes
Algoritmi, Universidade do Minho
Universidade dos Açores, Departamento de Informática,
Rua da Mãe de Deus, 9501-801 Ponta Delgada
e-mail: {armando.b.mendes, helia.mg.guerra,
luis.mp.gomes } @uac.pt

Ângelo Oliveira
SATA Airlines
Av. Infante D. Henrique, nº 55 - 6º
9504-528 Ponta Delgada
e-mail: Angelo.Oliveira@sata.pt

Luis Cavique
BioISI-MAS, FC, Univ. Lisboa, DCeT
Rua da Escola Politécnica, 141-147, 1269-001 Lisboa
e-mail: Luis.Cavique@uab.pt

***Abstract*- With the rapid growth of operational data needed in airlines and the value that can be attributed to knowledge extracted from these data, airlines have already realized the importance of technologies and methodologies associated with the concept of big data. In this article we present the case study of SATA Airlines. The operational and the decision support systems are described as well as the perspectives of using these new technologies to support knowledge creation and aid the solution of problems in this specific company. The proposed system provides a new operational environment.**

I. INTRODUCTION AND MOTIVATION

With the rapid development of technology, airline companies generate huge volumes of data at high speed. These data are created not only by the PNR's (Passenger Name Record, best known as reservation code) to keep data as the name, itinerary, number of passenger identification, taxpayer number, passport number, country of origin, number of tickets, etc. But, they also generated by internal company processes associated with the characteristics of the aircraft fleet, the schedule of each aircraft, the company's sales history, flight history, etc.

External data, such as market competition, population, national and international economic situation, the weather, among others, play an important role for the airline companies when they manage their current routes or planning of new routes.

In the top of the volume and variety of data in airlines is also important to notice two different types: the real-time data and stored data. The former concerns information captured in the moment (e.g., the current status of a flight: if the flight has seats available, which classes are available, etc.) while the latter are past data, with at least one day.

The vast majority of companies are developing its own software for data analysis to support decisions in multi-application environment, such as revenue management, customer care, etc. [3]. SATA airlines is no different. Since for long time the development of software in their IT department is common practice. However, for departments such as Revenue Management, where data analysis is of great complexity and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

urgency, and where the slightest mistake can cause huge losses of revenue, most companies buy specialized software from well-known software houses.

Using data to support airline decisions came a long time now, wherein pioneer experiences were made by Continental Airlines. During eight months testing, Continental Airlines, based on passenger internal information created three separate groups of high-level clients (passengers who frequently fly on Continental). When passengers suffered delays of over 90 minutes, one group received a letter with an apology request, another group received this letter and also a trial for the company's President's Club, and the 3rd group did not receive anything [3]. Passengers who received only letters of apologies spent 8% more in the following 12 months, and about 30% of those who received the trial for the President's Club joined the program. The profit was $6 million.

Because of its size, Air France-KLM has developed a new system of Revenue Management which contains considerably more data than their previous systems, including data about their passengers dated the previous two years. The system calculates and optimizes revenue for itineraries of origin / destination and bases its price according to the profiles of passengers. The system also estimates the probability of cancellations and no-shows on flights, permitting to evaluate the amount of overbooks (sell more seats than physically exist on the plane) allowed, with a certain level of security [13]. The system runs on the Hadoop architecture and was the first major company to use this technology [8]. The tests produced in this new system leave to a substantial improvement in system performance revenue management systems.

As authors like Davenport [9] and [10] recognizes Big data is a major opportunity for travel businesses empowering companies to enhance both the business and experience of travel. Big data can be the foundation for innovation, but demands big ideas and the courage to implement them. Managing and analyzing data is no longer an issue for IT departments alone, since all collaborators should be involved.

If data is made available, using big data technologies, it could be used to support more precise situations like the ones described by Christine Currie [8] in the sequence of discussions on the meeting of Revenue Management and Pricing International. The intention is to include other supply chain companies in the models concluding that it will be possible to taking account of the huge number of origin destination pairs in railway and flights, fairly pricing the many different holidays on offer from tour operators, where demand for each individual holiday is relatively low; taking account of multiple capacity constraints and treating vehicles on ferries and cargo as variable-sized entities rather than identical products.

B. Vinod [15] also introduce expected impacts of big data in Online Travel Agencies which are partners to SATA in selling seats, so this change can eventually impact on the company. In this article impacts like demand forecasting based on consumer preferences and hotel shopping, and dynamic ranking are discussed. The article concludes assuring that "Big Data revolution has arrived and is the new definitive vehicle to create competitive advantage by providing insights into consumer behavior patterns and improving process efficiencies for profitability that was previously not possible". Profound strategic alterations are also expecting in industry, as Alnoor Bhimani [4] advocates. These are possible directions of change also for SATA Airline.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The purpose of this paper is to analyze the software systems in SATA Airlines, and propose a system which includes Hadoop technology in order to create a new operational environment. This paper highlights two main subjects:

- the current information system in SATA Airlines
- the proposal of an information system more customer oriented and personalization oriented

The structure of this paper is the following. In the next section (section 2) we explain the basic technology of big data. The information system of SATA Airline is described (section 3) and the expected impact of the introduction of the new technologies in SATA (section 4). The article finishes with a conclusion (section 5) and further work.

II. BIG DATA TECHNOLOGY

With the advent of web 2.0 associated with mobile devices and the Internet of things, the classic operational applications were greatly surpassed in volume data. In a study of 2012, the estimated value of information on the planet was 2.8 ZB (1 zettabytes = 10^{21} bytes). The change of scale in data volume and its refresh rate gave rise to what generally is called Big Data. The name Big Data is associated with the acronym of 3Vs: volume, update velocity and variety of formats. Some authors include a fourth V for veracity. Others prefer focus on usefulness (or value of knowledge discovered). Mark Lycett [14] introduces the term “Datafication” like dematerialization, liquefaction and density proposed as the foundations of understanding big data analytics a key means of deriving value.

The existing 2.8 ZB, 85% are unstructured data, i.e. media such as video, photography and sound. Of the remaining 15%, formatted and text data only 3% have been analyzed. It is concluded that only a small percentage of 0.45% of the planet's data are studied. As in astrophysics where dark matter eventually contains several explanations for the origin of the universe, the 99.55% of the raw data are coined as "dark data".

With the emergence of new structured data formats arose within the Big Data the concept of NoSQL [2]. NoSQL, or Not only SQL, allows the storage, treatment and very efficiently query data. The NoSQL solutions use several data models:

- Key/value Storage, as Voldemort used in LinkedIn
- Super-columns Storage, HBase or Cassandra used in Facebook
- Document Storage, as XMLdatabase or MongoDB
- Storage of Graphs, as HyperGraphDB or ArangoDB
- Storage of Object Oriented, as Db4object

With the relational model and the declarative language SQL (Structured Query Language) in mind, used in most of the companies’ databases, NoSQL is presented as the alternative to handle large volumes of data.

For data aggregation the concept of MapReduce [2] is used, which is implemented in two phases and easy to apply in distributed file system like Hadoop from Apache™. The operator function is to divide the Map problem into subsets that can be distributed by other cluster nodes. The Reduce operation takes all the subsets

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

of data, gather all pieces and processes information to respond to the original query. Unlike SQL, which stores the results in memory and returns them to the client without storing it permanently, while in MapReduce, it runs in two phases, firstly organize data by the subsets from different sources, and secondly count the contacts for each subset, see Fig. 1.

In SQL the queries are processed in one phase, while in MapReduce, it runs in two phases, firstly organize data by the subsets, and secondly count the contacts for each subset, see Fig. 2.

Hadoop [2] is an ecosystem that provides management, storage and processing of large-volume, in a distributed and redundant way. Hadoop has the following functional components - Job Traker, a central process that manages and coordinates the parallel processing of different nodes; Task Tracker, a process that makes the information processing in each node; Name the Node, a central process that manages and coordinates the storage of information; Data Node, which is the storage of information on each node.

Consequently, the Big Data creates new opportunities in decision-making based on data, "data driven decisions". As said Peter Norvig, director of Google Research, "we do not have better algorithms but we have more data" [8]. Data Science is the current term for the science that analyses data, combining statistics, operational research, data mining and database technologies, to meet the challenge that Big Data presents. The term coined in the 2010s match what in the 1970s was named Decision Support Systems, DSS, in the 1980s the Executive Information Systems, EIS, in the 1990s the Online Analytical Processing, OLAP, and in 2000 the Business Intelligence, BI [8].

If the complexity of the algorithms cannot be controlled, the answer may lie in reducing the dimensionality of the data. Recent works using two-phases can be mentioned. In the first phase the raw data is collected in a condensed data structure: network in Topological Analysis of Data [5], Petri net [1], Markov chain [4] or graph [6]. In the second phase patterns are searched in the condensed data structure.

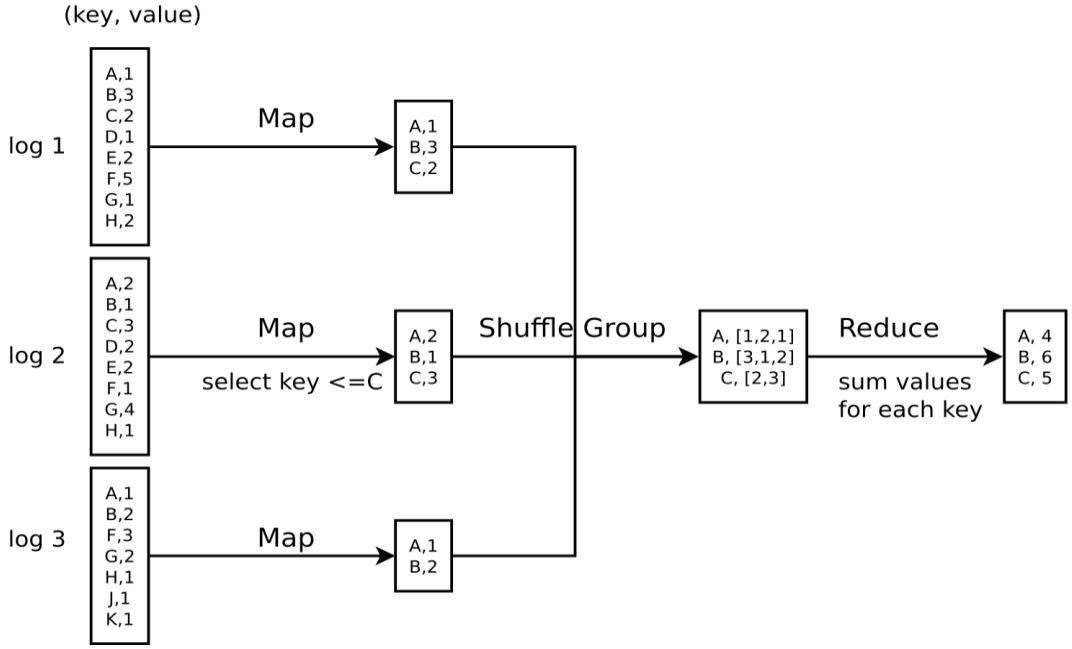

Figure 1. MapReduce workflow deals with data from different sources.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. DATA AND INFORMATION SYSTEMS IN SATA

SATA Airlines is a business group whose mission is to connect the nine islands of the Azores archipelago and also link it to the rest of the world. Owned by the Azores regional government, it includes two airlines, two tour operators and an airfield management division, comprising over 1200 employees and associates.

Its origins go back to the year 1941 when the five founding members created the "Azorean Society for Air Travel Studies", whose aim was to end the isolation of the islands. Like many other successful projects, SATA was born on a simple idea. The founders believed that the strategic location, uniqueness and natural beauty of the Azorean islands provided sufficient motivation to create a strong, successful airline company. Over six decades have passed, and despite the many difficulties encountered over the years, time has proven that they were right.

As an ambitious project, SATA took off on the 15th of June of 1947, when a small Beechcraft airplane, symbolically named "Açor", made the first commercial flight. Today the company plays a major role in providing accessibility for those who live on the islands. It has also become an invaluable ally in the growth and development of tourist activity in the Azores.

Since 1995 it flies also to several destinations in the European and North American continents. The company mission is to develop sustainable mode of air transport activity related to the Azores, with a vocation for the Atlantic based on a reliable service, hospitable and innovative. This mission implies several air flight related activities defining the SATA group: full service air transportation (middle and short range), handling service, cargo management, touristic operators, and aerodromes management, near shoring back-office and front-office services, and even a training academy.

Six decades of flying have provided the SATA Group with real aeronautics 'know-how'. Under the blue skies and over the blue ocean, the company strives to bring the soul of the Azores to every aspect of its activities. Everything supported by a myriad of interlinked systems. In Fig. 1 we can see a schema concerning its passenger business model, the main systems and the communication connections between them.

The airlines companies, working in international environments, deal with many external data sources, regarding the reservation and ticket selling. The internal applications are mostly related to revenue, operations and customer data, as present in table 1.

TABLE I
DATA SOURCE

External	**Internal**
Global Distribution System (GDS)	Revenue Management (RA)
Marketing Information Data Transfer (MIDT)	Revenue Accounting (RM)
Departure Control System (DCS)	Operations, Planning and Control (OPC)
Pricing	Frequent Flyer Program (FFP)
Lost & Found	Customer Care (CC)

In the operational side, the main system is the Global Distribution System (GDS), which is manly a B2B service offered by an external provider. As examples of GDS's, Sabre, Amadeus, Apollo, Galileo and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Worldspan could be mentioned. It is the communication channel between the airline companies and flight seat retailers. Through it, journeys between two points are available, independently of carriers and itineraries, regarding the required travel dates, services and range of fares. The retailers, like travel agencies, can search flight seats for the costumer's necessities. It also provides train trips, car rental, hotel and travel insurance services.

The GDS have specific sub-products for data analytics. MIDT (Marketing Information Data Transfer) is an example of these products which provides the buying behavior.

This system is central because almost everything in the operational part of SATA depends on the information about seats sold. The GDS communicates using EDIFACT (Electronic Data Interchange for Administration, Commerce and Transport) messages, which is a special short coded text message. In SATA, it embedded in its website, where it is necessary to show the user alternative seats in different flights in response to user specified queries.

Another system is the Operation Planning and Control (OPC), which is supported by AIMS, an Air Olympic spin-off company that installs of-the-shelf fully integrated crew management and aircraft fleet control solutions. Installed in SATA datacenter, it generates the sets for GDS to sell and, of course, it needs to know about sales to efficiently manage seat availability and real time flight status to control fleet and crew promptness.

The DCS (Departure Control System) is the system whose frontend is present at check-in counters. It is another external system that matches GDS reservations and inventory tickets with planned flights, managing passenger and freight allocation, organizing aircraft load, and controlling aircraft departure procedures. It also communicates via EDIFACT messages with other systems like FFP and RA.

Lost&Found is an external system, provided by Worldtracer, for dealing with lost baggage. It has no interaction with other systems and it is only used to post information about lost and not reclaimed bags in airports and process the matching.

The Frequent Flyer Program (FFP) is a system that keeps track of clients and bonus offers. It is the main directory of information about SATA passenger customers. It is an optional adherence program so only a small fraction of final customers is tracked using this system. GDS and DCS communicates which seats the client bought and, via a frequent flyer card number, this system records offered miles and can also offer other treats like premium service or discounts in associated companies as rent-a-car and hotel service. It also communicates with Customs Care (CC) for checking complaints about miles and services offered.

This last system collects and process all the passenger input inside and outside the flight. It is also a way to collect information about the end users. It also checks with GDS flights and seats.

Revenue Accounting (RA) system is mainly for accounting, recording ticket sales and coupon usages, audit fare bases and forms of payment, prorating city pair revenue, etc., keeping track of flown coupons and other usages. It obviously needs sales information from GDS and lift usages from DCS. In SATA, it is also an off-the-shelf system.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Revenue Management (RA) and Pricing are two systems and services interlinked and very interesting from the point of view of decision algorithms. They will be best explained in the next section.

The two operational systems that use optimization, forecasting and AI or some form of learning heuristic are used in the revenue management (RM) and pricing (see Fig. 1). For instance, revenue management, after the best model is chosen for predicting the flow of passengers in peak seasons (e.g. Christmas) and off peak (low season), the software applies the optimized model for future flights, i.e. calculate the passenger flow. It has also the possibility of showing programs on "special" times such as school holidays, regional festivals, important games, etc., thereby allowing a forecast based analysis of the route at different times, and enabling profit maximization per seat sold or even enhance / reduce aircraft rotations. When a certain flight has low bookings, the revenue management has this information in real time, and can create a promotional / campaign fare along with the pricing to monetize the flight, thus creating lower tariffs and providing attractive passenger. The customer will be satisfied with the tariffs and the company can fill an unattractive flight, thus optimizing company revenue. These are complex decisions based on forecast data, in fact, a flight can have a maximum of 24 price classes.

The actual DSS is supported by a data warehouse fed by every piece of information the company can set the hands on. Traditional Business Intelligence is also implemented including OLAP system and reporting used mainly to strategic decisions in the company. The data warehouse also feeds the site, which is used as the main channel with the end costumer to sell seats, frequent flyer program and complaints. The company use also other channels like a call center and mobile apps.

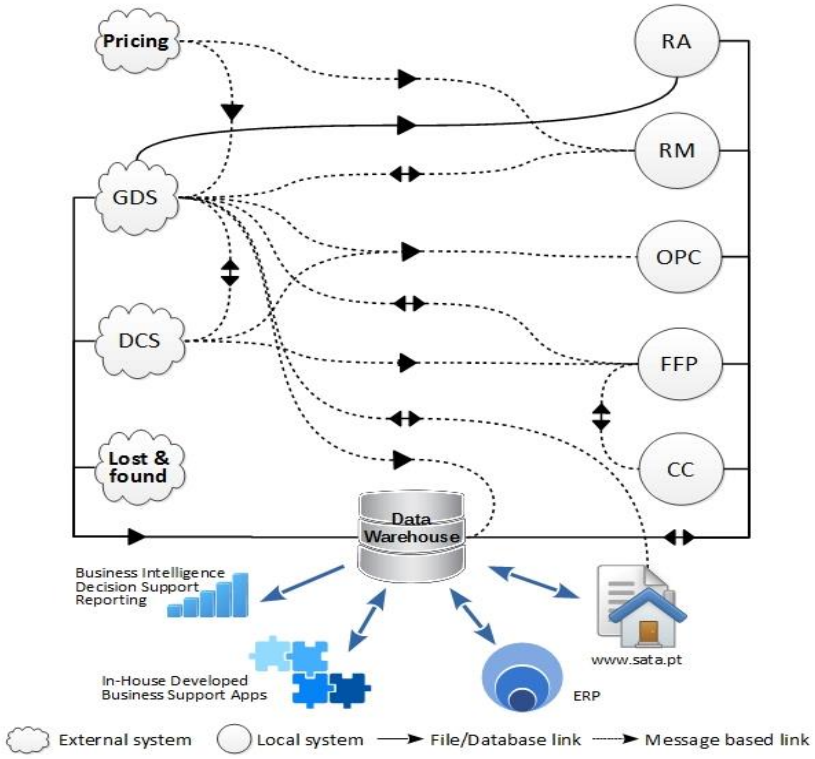

Figure 2. Schema of the front office systems implemented in SATA for airline business.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. BIG DATA IN SATA AIRLINE

The main driver for big data adoption in SATA is to provide in real time personalized services to its customers. Personalizing provides better services by predicting customer needs. What customers will be buying in the future is fundamental to gain advantage in the competitive airline market.

In this paper we propose a Data Lake, based on Hadoop technology to support the customer management, which will work in parallel with the current Data Warehouse.

A Data Lake is a massive repository for storing Big Data with low cost computer hardware. Data Lakes differ from Data Warehouses and Data Marts, which are oriented for fast queries over pre-aggregated attributes. Data Lakes retain all data, even when the scope and the purpose of the storage is not known. Data Lakes support all data types and any kind of user, offering an environment that easily adapts to change. The data is transformed to fulfill the needs of each analysis and providing a way to faster insights [7], [11] and [12]

SATA has already a lot of customer data, although not totally structured, that can be used for segment customers according to their needs. A Data Warehouse was developed for integrate information provided by different unconnected systems, namely bookings, check-in counters and luggage complaints. The real challenge with the Data Lake is to cross reference transactional (structured) data with unstructured data obtained by customers from different means, such as voice interactions in call centers, website activities (e.g., sequence of events such as mouse clicks), reviews on blogs and social media, and ultimately single out the individual customer.

SATA only knows well its frequent flyer customers. Most of passengers are unknown to SATA for several reasons. Some passengers are not members of the frequent flyer programs, or do not identify themselves as members when travelling. Since about 2/3 of SATA seat reservations are made by its partners, the majority of passenger details, including contacts, may not be available in a structured way or can be protected and not accessible.

Thus, building algorithms to parse existing data and cross reference with historical records in order to get a complete characterization of every passenger is a vital necessity for SATA.

Hadoop technology will be used to support services offering amenities to passengers, such as arrangements of onboard meetings, taxi or bus sharing if they go to hotels in the same vicinity, organization of special care services for passengers who have singular needs or who have had less positive experiences in the past. Moreover, big data technologies can even be used to pinpoint travel agency misbehavior by identifying conduct patterns.

Best customers who are not enrolled as FFP members could be isolated and presented with alternative rewards or personalized services. Also more information concerning to passengers and operational issues could be available on personal contact points either on ground or inflight.

Although the analysis of internal data is central, the analysis of external data, such as MIDT's (Marketing Information Data Transfer) is also an essential part, in spite of the fact that these data represent only about 60-70% of the airlines total information. In fact, MIDT's do not include data of reservations made in companies such as contact centers, on site companies, etc., neither include data on low-cost airlines. MIDT’s only include GDS's (Global Distribution System) reservations.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Combining these two sources of data, the airline can check what their passengers are doing, especially if the analysis of aircraft and crew are available. With the aid of MIDT it can also be check out the competition and flux of passengers to each destination, contributing to the airline make decisions about whether to invest in a certain new route and / or discard existing routes.

The Data Lake (Figure 3) will store external and internal customer data, in order to replace the current Data Warehouse.

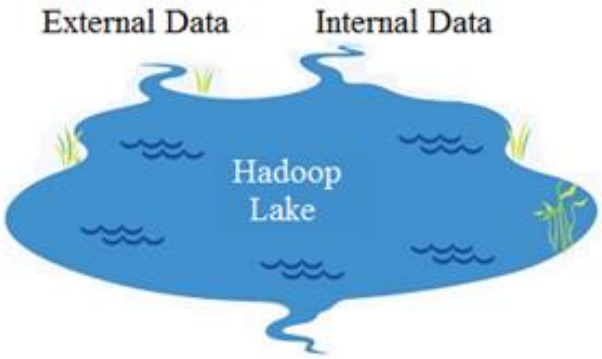

Figure 2. Data Lake.

The conditional phases of the Data Lake project are the following:

- phase 1: Customer Data Lake and Data Warehouse work in parallel;

- phase 2: If the Customer Data Lake seems competitive, discontinue the Customer Data Mart in the Data Warehouse.

The proposed architecture allows the use of scalable data mining models for customer management, such as, probabilistic predictive models, segmentation based of historical data, etc., detailed in Table 2.

TABLE II
EXAMPLE OF DATA MINING ALGORITHMS USED IN CUSTOMER MANAGEMENT

Model	**Problem Type**	**Algorithm type**
Customer Segmentation	Clustering	k-means clustering
Customer Loyalty Analysis	Classification	Decision Tree, SVM
Costumer Life Time Value Analysis	Classification & Regression	Decision Tree, Generalized Linear Model, Regression
Frequent Flyer Passenger Prediction	Classification	Decision Tree, SM

V. CONCLUSION AND FUTURE WORK

In this article we described the systems and information in SATA and suggest new systems based on big data technology to support particular decisions especially relevant in SATA Airline.

In order to create a new customer approach in this paper we propose a Data Lake supported in Big Data technologies which has a capacity of holding a great amount of raw data in its native format. The Data Lake will use external and internal data, from MIDT, FFP and CC applications and any additional sources with customer information.

SATA is looking forward for implementing big data technologies and we intend to write follow up articles describing the impact of these technologies in the operational decisions, but mainly in decision support of more unstructured decisions like the ones discussed in this article.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ACKNOWLEDGMENT

We would like to provide a special thanks to Carolina Amaral Dias for the idea of this article and the initial work. Paulo Ornelas, the head of IT department in SATA, was also fundamental in this project. It was also relevant, in the early days, the collaboration of Catarina Costa and Diogo Silva, for their endless patience and availability to explain the terms, concepts, and strategies of Customer Care and Revenue Management.

REFERENCES

[1] Aalst W. van der, "Process mining: discovery, conformance and enhancement of business processes, Springer, 2011. ISBN 978-3-642-19344-6.

[2] Alexandre J., L. Cavique, "NoSQL no Suporte à Análise de Grande Volume de Dados," Revista de Ciências da Computação, nº8, pp. 37-48, 2013.

[3] Anderson-Lehman R, Watson H. J., Wixom B. H., Hoffer J. A., "Continental Airlines Flies High with Real-time Business Intelligence". IS Quarterly Executive Vol. 3, 4, pp. 163-176, 2003.

[4] Bhimani, A., "Exploring big data's strategic consequences". Journal of Information Technology, vol. 00, pp. 1–4. 2015. http://dx.doi.org/10.1057/jit.2014.29.

[5] Carlsson G., "Topology and Data", Bulletin of the American Mathematical Society, vol. 46, 2, pp. 255–308, 2009.

[6] Cavique L., "Network Algorithm to Discover Sequential Patterns", in Progress in Artificial Intelligence, J. Neves, M. Santos and J. Machado (Eds.), EPIA 2007, LNAI 4874, Springer-Verlag Berlin Heidelberg, pp. 406-414, 2007.

[7] Coral W. and Hassan A., "Personal Data Lake with Data Gravity Pull," 2015, pp. 160–167.

[8] Currie, C.S.M., "How far will the airline model stretch?" Journal of Revenue and Pricing Management, pp. 1-3, 2015. http://dx.doi.org/10.1057/rpm.2014.39

[9] Davenport T.H., "Big Data at Work: Dispelling the Myths, Uncovering the Opportunities", Harvard Business Review Press, 2014. ISBN 978-1422168165.

[10] Davenport, T.H., "At the Big Data Crossroads: turning towards a smarter travel experience", Report. Amadeus IT Group, 2013.

[11] Huang, F. "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem," 2015, pp. 820–824.

[12] Ignacio T., Peter S., Mary R., and John E.C., "Data Wrangling: The Challenging Yourney From the Wild to the Lake," presented at the biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, California, 2015, pp. 1–9.

[13] Lawrence, R.D, Hong, S. J., Cherrier, J., "Passenger-Based Predictive Modeling of Airline no-show Rates". In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 397-406, ACM New York, NY, USA, 2013.

[14] Lycett. M., "'Datafication': making sense of (big) data in a complex world", European Journal of Information Systems, vol. 22, pp. 381–386, 2013. DOI:10.1057/ejis.2013.10

[15] Vinod, B., "Leveraging BIG DATA for competitive advantage in travel" Journal of Revenue and Pricing Management, vol. 12, pp. 96-100, 2013. DOI: 10.1057/rpm.2012.46.