2013 International Conference on Advance: in ICT for Emerging Regions (ICTer): 207 - 216

Big Data solution for Sri Lankan development: A case study from Travel and Tourism

Rinusha Irudeen [#1] and Sanjeeva Samaraweera [*2]
Database Competency Excellence Group
Virtusa (Pvt) Ltd
Colombo 09, Sri Lanka.
Email: nr.irudeen@gmail.com [#1], sanjeeva.samaraweera@gmail.com [*2]

Abstract—**In this paper, the key aim is to provide conceptual, technical solution design to implement a supportive dashboard which integrates with Big Data indicators and online transaction processing (OLTP) systems. The proposed solution is mainly focusing on a case study of Travel and Tourism industry in Sri Lanka. We carried out in-depth analysis to identify the necessity of a dashboard and examine significant challenges need to consider when designing the solution. Moreover, we evaluate suitable big data technologies for implementation and Hadoop, Hbase, MapReduce has been proposed. Centralized repository for user Meta data, Contextual Search, Early Warning Alerts, Index Indicators, Analytic tool and Reports, Marketing campaign optimization, Link with social media features, Sales and marketing forecasting are main dashboard features that has been designed based on the requirements. Our results attest importance of Index indicators, one of the major functionality which is built-in to dashboard. In this work, we present a detailed analysis of a total efficiency index using four indexing strategies of varying complexity including Visit Index, Wealth Index, Health Index, and Lifestyle Index. We conclude by designing an open architecture, that can track and leverage data on the behavior of tourist via a dashboard which consider trends, to make better decisions, reduce risks and drive personal tourist experiences.**

Keywords— **Big Data, MapReduce, Hadoop, Hbase, Dashboard, Apache Sqoop, Apache Mahout**

I. INTRODUCTION

Big Data has become very comprehensive and a tech buzzword these days. It is breaking down the barriers that existed with exploding amount of data, historical data, expensive and complicated databases. With the rapid evolution in the Information Technology (IT) industry, organizations are contemplating in moving towards Big Data solutions instead of conventional Relational Database Management Systems (RDBMS) and Datawarehouse (DW) [1]. Despite the global financial crisis meltdown directly impacting the Sri Lankan economy, leaders in the public, commercial, and social sectors are focusing on new business opportunities within the country to boost up revenue through foreign exchange [2]. There has been an incrementing trend in travel and tourism industry after the end of Sri Lanka's 30 year civil war. Sri Lanka as a nation is economically growing with increasing GDP growth rates and has become open for new business opportunities and constantly attracts foreign direct investments.

In a nutshell, most of Big Data related researches are based on two main areas named: Technology of how big data to be processed[3], and marketing of tools by different vendors [4]. It is a significant constraint to apply big data concepts without extensive re-work by expert data scientists. This paper mainly discuss on how to utilize existing big data technologies and tools in a more practical manner for the up-liftmen of society. However there is less research work on centralized repository or personal dashboard which includes Big Data indicators to analyze trends, alerts, personal behavior and data. With the growing size of dataset it requires repository to be scalable and highly efficient.

The research method followed full life cycle process while providing descriptive and practical insight into big data solution. We will discuss on: Use of Hadoop and Map Reduce big data distribution concepts, Crawling big data from internet and other sources and mapping into key value pairs, Store in Hbase database, reduce the data set into dashboard requirements, Design of dashboard with thresholds especially with big data indicators. The Dashboard embedded with features to analyse web user clicks and categories click data to analyse a specific tourist, examine specific tourist behavior to guide in various mode, ability to turn customer transaction history into intelligent recommendation for logistics, sell products and travel and tourism.

Personal indexes are very useful in many industries [5]. This paper examines the applicability, usability and design of such indexes. The dashboard is screening personal data as an index and exploration probes categorized into four indexes including Visit Index, Wealth Index, Health Index and Lifestyle Index. While carrying out insight into theory behind Big Data architecture, Map Reduce and problem solving process derives these indexes efficiently. The proposed Big Data solution assists to solve a real world problem of summarizing massive quantity of data from different sources. Mainly, it enables to provide the potential value across stakeholder cost effectively.

This paper mainly consists following sections. Literature review presents the summary of the research that was carried out, findings and decisions made. Case description and solution design section present the analysis of the research, system requirements and methodology selection. This also outlines the design phases of the solution; diagrams are presented that illustrates the design. Furthermore this section discusses the process carried out and the problems faced and also how the problems were given the best possible solution. Testing phase was carried out in order to identify weaknesses of the system and correct them. Discussion provides a critical evaluation of the research, and limitations of the current solution and possible future enhancements will be outlined. Conclusion presents the personal reflection for the research and presented solution.

II. LITERATURE ANALYSIS

In the literature it is recognized that necessity of integrated big data solution while investigating a dichotomy exists between big data technologies and solutions. Big data solution approach is still not prominent in Sri Lankan travel and tourism industry to generate massive revenue and deliver

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

analytic for decisions making and customer services. Analysis of recent studies carried by vendors, market researchers, solution developers and government intervention statistics, policy form the focus of big data technologies, studies of the objectives and decisions confronting to design big data solution which is fitting for Sri Lankan travel and tourism industry. This research combines two disciplines: In contrast, Analyze existing big data technology and solutions used in travel and tourism industry. Secondly, how to utilize, leverage existing technologies and design big data solutions in a more practical manner which will suites the travel and tourism industry in Sri Lanka.

The statistical analysis published by the Sri Lanka tourism development authority (2011) shows the enlargement of trends and structural characteristics of tourist traffic, as International tourist arrivals grew up to a total 980 million in 2011. Revenue from tourism, scheduled airline operations and passenger movements, Income and Employment, Tourist Prices indicates that travel and tourism industry has a direct involvement in core foreign exchange earners in the overall economy of Sri Lanka [6].

Various methods has been used to extract, store, process and present data. With the advancement of hardware, the price of storing devices and processing devices has gone down. Advancement of web and sophisticated devices generate the need to work with massive loads of different kinds of data. Bigdata can be defined as the high volume, high velocity and high variety of data assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [7]. The primary challenge is to build frontier dashboard to analyze data by converting seemingly unstructured data into useful information.

Bigdata processing got it's initiation through commercial giants like google and yahoo. Currently, it has been taken up by open source organizations including apache and commercial organizations like Oracle and IBM [8].The Google/OTX study finds that (2011), most of the travellers gather information before booking [9] [45]. Davenport (2013) asserts that travel industry should begin considering real big data solutions to provide better services and tailored travel experience to their customers, as creating an integrated data source, data storage issues, and working in a hybrid technological environment becomes challenge to big data solutions in the travel industry. In fact, recent research reported that case studies of big data adoption, KAYAK travel meta search engine find the best possible flight or hotel, Amadeus IT solutions for Air Lines "look to book" ratio, Facebook's ad, British Airways (BA) Opera Solutions, Marriott, Hipmunk, Munich Airport solution heavily depend on variety of big data technologies and analytics for both internal decisions and customer services [10]. Mitra (2007) also identifies the gap existing between travel search engines including issues in poor content, search function and vertical search issues [11]. Ventana (2012) research highlights those large scale organisations, working with large scale data beginning at one terabyte (TB) to petabyte (PB) range still using relational databases for their large-scale data processing [12]. As states by recent research there are different types of big data solutions provide by different vendors and there are limitations in solutions. Defining such solutions which is suite for Sri Lankan Travel and tourism by avoiding limitation is a huge task. Apart from that, how this challenge is met is critical because organisations are highly relying on effective analysis of the information to stay on competitive. Therefore given solution needs to align with the stakeholder requirements, limitations and policies.

Furthermore, research proves that typically high percentage of RDBMS, Flat files, data warehouse appliances and in-memory databases are used technologies to manage and analyze big data [12]. There are combination of new technologies to manage and analyze big data including Hadoop and specialized analytical technologies such as columnar databases [13]. Chen et al (2012) organizations run reports and queries against of historical data, however with the growth of data it is not practical and hard to navigate through them to find the most useful items. Predictive analytics, planning and forecasting and visualization techniques are the process of examining large amounts of different data. Moving forward organisations are emerging these techniques into big data solutions [14]. As data volumes grow use of Hadoop will assist for predictive analytics and for visualization [15]. Comparing to the other technologies Ventana (2012) research asserts that 67% of organizations are using Hadoop to build their new products and services [12]. Apache promotes Sqoop import data from numerous relational databases into HDFS, Hive or HBase [16]. We have used Apache Mahout to classify data and use as a platform for our research since the collection of data to be processed is very large and based on collaborative filtering, clustering, and classification [17]. Several practical case studies imply that use of standard Hadoop's MapReduce model to investigate large data issues [18]. MapReduce framework based on Hadoop and it is easy to design efficient MapReduce algorithms for an instance there may be numbers of documents where each document needs a set of terms and need analyse a total number of occurrences of each term in all documents therefore by using MapReduce algorithms and Basic MapReduce Patterns including Counting and Summing, Collating, Filtering, Parsing, and Validation, Distributed Task Execution, Sorting helps to design efficient solutions [19], [20].

GigaSpaces solutions carried out big data survey by aiming IT and business professionals. They assert that business decisions are heavily reliant on analysis of the data or to handle rapidly growing data inputs. Therefore organisations consider Big Data processing as mission critical and it is essential to data processed in real time. Furthermore the survey clearly indicates that organisations seeking combination of Big Data and cloud computing solutions to achieve maximum output [21]. Therefore in the future development, authors needs to focus on the combination of big data and cloud computing solution as well.

III. Case Description

With the end of 30 year civil war Sri Lanka is going through a rapid development phase and one key goal is to be transform Sri Lanka into a tourist hub [22]. Big Data technologies have the potential to accelerate a country's development in various aspects. This paper attempts to provide insights into a case study of Travel and Tourism industry in Sri Lanka while designing a solution to build new conceptual breakthroughs.

The proposed dashboard can be used by organizations which already have assets of Big Data and need to kick start without any consultation fees and overheads. It is integrated with comprehensive Big Data technologies to build an

enterprise level reporting and business insights platform. The main challenge in the Big Data industry is extracting proper value from Big Data. Nowadays Organisations are investing millions on flashy dashboards and reporting tools. However due to the lack of capabilities of these systems they are poor in presenting useful insights and have not achieved the expected return on investment (ROI) from expensive dashboards and reports [23].

Need for understanding Big Data solutions, adoption, and demand of Big Data technologies and how it could revolutionize the business in novel ways has got increased mainly due to amount and the proportion of unstructured data in the whole data volume [24].

Most solutions stick to conventional structured solutions where as others who are brave enough to get into the unstructured data will ultimately show information not worthwhile. Due to the popularity of buzzword 'Big Data' the market is looking for proper solutions but it is imperative that industry has to come up with usable solutions. Although it is not an easy task, the main idea of this paper is to show a practical way of using the Big Data. The proposed solution includes examples for unstructured and unpredictable social network data. Organisations value their data as corporate asset because data can be effectively transformed into valuable information that is used to make business decisions. With the growth of unstructured, unpredictable data, this initiative is really about installing the concept of managing data and providing Big Data solutions in different aspects. In Travel and Tourist industry there are lots of possibilities in leveraging on Big Data Solutions to predict their desired travel destinations.

- Social networking data on purchase patterns and the idea about their buying power can be used for promotions of commodities, hotels and travel agents.
- Information and feedback about visits to SL in their blogs will immensely help hotels, boutiques, airlines, airports and government institutes to improve their services.
- Government authorities can use summarized data to improve infrastructure and plan for capacity and accommodation trends.
- Stakeholders can estimate the number of tourists to cater in next season by predictions using Big Data analytics.
- Know the current trends in tourism using social network data of similar stakeholders in tourist destinations in other countries.
- Stakeholders can plan customized tour packages, promotional discounts, end of tour souvenirs etc according to consumer needs.

This is done by designing personal dashboards and grouped total reports using Big Data analytics.

IV. Challenges

As revealed in the introduction, Sri Lankan government is looking for to increase profitability, return on Investment, modernize operations, improve tourist retention [25], extend product lines and reduce risk through a solution. Existing services, products and solutions are constrained by traditional data integration approaches that hinder productivity and benefits anticipated.

There are significant challenges that need to be considered when designing the solution for Dashboard. Data Storage and complexity are key factors. Therefore dashboard should leverage data from multiple internal and external sources, including structured, semi-structured, un- structured and Big Data types such as Hadoop. Large volume of data needs to analyse right through the solution considering the performance issues should be considered. In a rapidly changing business environment, information has to be up-to-date, accurate, and accessible and well governed [26]. New data sources need to be brought rapidly and existing sources need to modify according to the current requirement.

V. Solution Design

A) Requirement Gathering

Requirement gathering is extremely important to the success of the study. This solution was developed using a thorough review of the literature on stakeholder analysis [25] [28], policy process [27], consider the requirements [22], benefits [25], obstacles and future work as well. Initially, it is essential to discuss with all stakeholders and survey literature of tourism industry. A brainstorming session was carried out on how best a manual reading of internet resources help to uplift the industry. Furthermore, lists of reliable web resources were identified in order to gather information.

Identifying and understanding the stakeholders and their interests is important to provide appropriate engagement solution. As explained above, the proposed solution is based on a case study from travel and tourism industry to apply the Big Data concepts. It is important to understand the stakeholders and their objectives in order to ensure that all aspects have been addressed. Therefore we define stakeholders based on interest, ownership, specialist knowledge, impact or influence and contribution. In working on this paper we gathered information from all stakeholders including businessmen working in the Tourism Industry, tourists as well as government officers. With that information we designed a common personal dashboard customized for tourism industry.

To get information onto the dashboard we researched on best available Big Data technologies to populate accurate, fast and important indicators for end users to take decisions [28]. Unlike a RDBMS which stores only transactional data or a Datawarehouse which facilitates management to take decisions using aggregated data, the proposed personal dashboard will be very useful in end user to get summarized information about a person from different sources like internet and make decisions [29].

Especially here we are looking at the possibility of getting information on individual tourists. Going through all these resources manually and understanding about the tourists personal information is vital. The following are some basic criteria for evaluating the appropriateness: Wealth, Health, purposes of visiting the country and Life style. Regardless of the purpose of visit, Tourist's life style is very important for a tourist hotel or any other organization which is trying to invest on tourist [25] [30].

Even though manual research proves to be best method, it is not effective and profitable to read about all tourists and keep information beforehand and also reading on a particular tourist from all sources is not practical. Also if the task is distributed among several persons the criteria they used to see how much efficient a particular tourist for investment is vastly different.

Due to these reasons it is necessary that the process should be automated and proper efficient algorithms are in place in

getting efficiency index of each tourist. Also a computerized dashboard will be very helpful for stakeholders to compare efficiency with the average samples. Therefore authors came up with the concept of tourist personal dashboard showing a single efficiency index derived from various indexes gathered using social network sites. This approach is likely to optimize the dashboard effectiveness and adoption by using sophisticated algorithms in future. Proposed solution builds on top of rich platform (Meta layer) and it helps to reduce the database complexities. Therefore, provides end users with smooth interaction with reporting, indicators, alerts on their own.

B) Big Data Architecture

As explained earlier we are mainly concerned about facilitating for stakeholders who are end customers like tourist hotel cashiers and Visa officers. Using the solution they should be able to take decisions upfront. Unlike a conventional big data solution integrated with a data warehouse, our aim is to integrate with the OLTP system. Open architecture for the proposed dashboard illustrated in figure 01 [31].

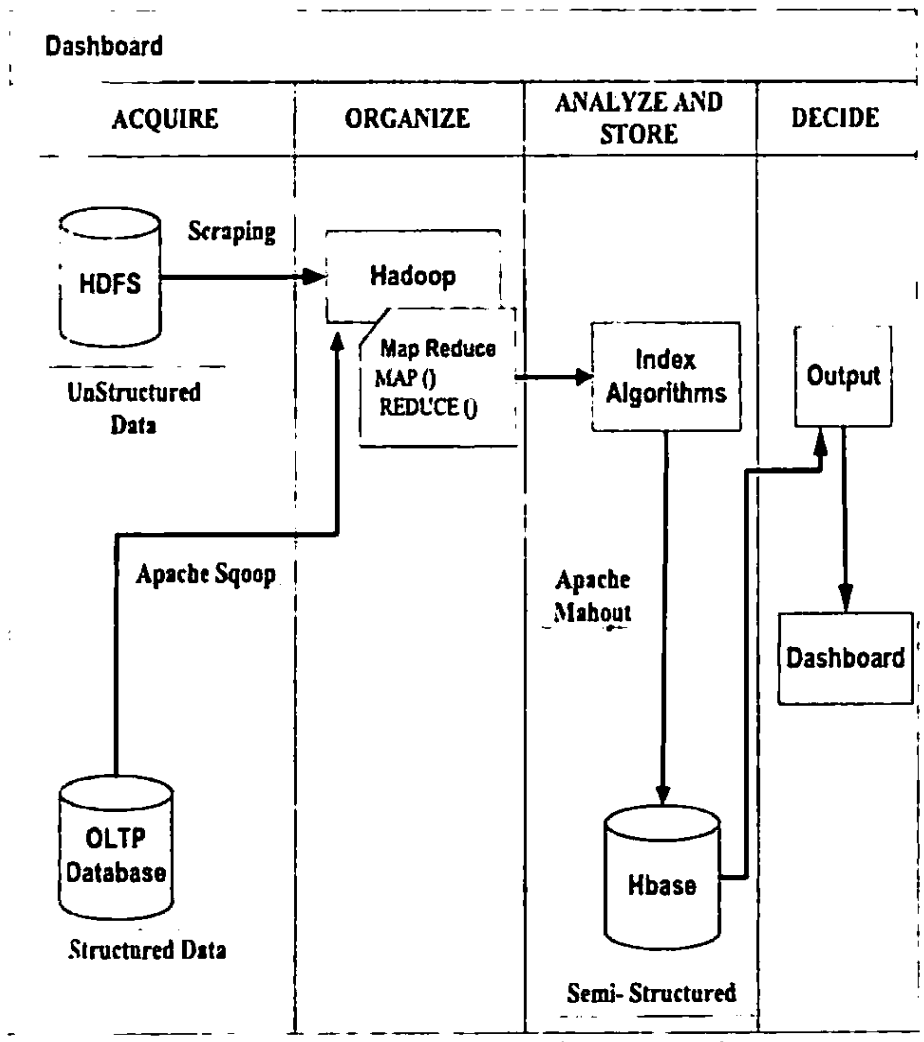

Figure 1: Architecture Diagram for Big Data Solution

1) Crawling Data from Internet: There are four main sites considered for personal data extracting. These four sites named Facebook, Twitter, LinkedIn, and Google+ were analyzed for their behavior below.

The purpose of web crawling scraping in this project is to collect information about a tourist from social websites in order to provide information about that particular person to the dashboard. This information feed will provide the dashboard with the relevant details of the persons' behavior and habits and will aid the stakeholders in the rating process.

	Facebook	Twitter	LinkedIn	Google +
Tourist Usage/ Tourist information	High	High	Low	Moderate
Posts/Tweets Update Frequency	High	High	Low	Low
Ability to find if account is verified or not verified	Moderate	High	Moderate	Moderate
Accuracy of information in posts /tweets	Moderate	High	Low	Moderate

Table 1: Crawling Data from Internet

Following are few open source scraping tools based on java which was evaluated for this project.

Jsoup is the main tool that is used here for extracting specific data from HTML and Twitter4J will be used to scrape tweets from Twitter [32] [33].

Scraping tools	Purpose
JTidy	To use a XML based tool to traverse it.
Jsoup	To extract specific data from the HTML
HtmlUnit	To unit test the HTML.
TagSoup	To parse a non-formatted HTML document.
NekoHTML	To parse a HTML document having many mistakes.
Twitter4J	To scrape tweets from twitter.

Table 2: Scraping tools

2) Integration with OLTP Database: This is done using Apache Sqoop software [34]. Tourist Data from Oracle database of Department of Immigration and Emigration is planned to be imported into HDFS and will be used for scraping.

3) Algorithm to distribute scraped data in HDFS: This will map clients according to their country of origin [35]. This is done inside the crawling algorithm and it will check a proper country of origin data column from a reputed social network engine and shard data according to country of origin. This will be very important as it's easier to identify similar trends and behaviour from tourists in the same country. Natural Language processing algorithms also can be enriched with inherent qualities of a certain nation [36].

4) MapReduce algorithm to get the indexes: Total Efficiency Index Indicator for tourist is the single indictor that is shown in this dashboard as a single value calculated from this dashboard which shows the possibility of doing business with a tourist.

- Greater than 75% - high possibility of doing business
- Greater than 50% - possible and need to incorporate strategies

- Greater than 25% - less possibility of business, worth trying
- Less than 25% - no possibility of business, not worth investing

This is calculated with several factors and these factors are given a parameterized weightage. Following is a brief description of how each factor is calculated.

Factors	Suggested Rate
Visit Index	25%
Wealth Index	25%
Health Index	25%
LifeStyle Index	25%
Total	100 %

Table 3: Parameterized weightage for Index

a. Visit Index

No of times visited out of country, No of times visited SL etc has to be counted in the reduce algorithm and added into counters. A NLP algorithm searches following criteria [19] [20]. Apache mahout is used as the NLP –machine learning tool and some of the key words used in creating knowledge base are as following [21].

Max number of visit count is defined as a threshold and word count will be counted as a percentage. These values are calculated for current status and will be compared with previous values in a graph inside the dashboard.

Words
Toured
Visited
Foreign Trip
Sri Lanka
Asian Country [check list]

Table 4: Parameters for Visit Index

b. Wealth Index

This is the measure of buying power. NLP algorithms designed to find how much spent on items.

Words
Bought for <> dollars
Its <> dollars
Sold me for <>
Ticket was <>
Words
Bought <good>
Sold me <goods>
Watched a drama/movie

Table 5: Parameters for Wealth Index

c. Health Index

This is the index of health of a person. If the health is not good the possibility of a person visiting and spending time is very less. Following words will be counted for a period in a NLP algorithm and index is calculated [37].

Words (-counts)
Got a flu
Have arthritis
Diabetic
Words (+counts)
Yoga
Did abs
Did a workout
Diabetic

Table 6: Parameters for Health Index

Each ailment will count and healthy habits accumulated and health will be measured accordingly.

d. LifeStyle Index

This is very important to understand the matching characteristics to a particular hotel or destination. The words here are parameterized and can be changed by client.

Words (-counts)
Rowing
Ayurveda
kiribath

Table 7: Parameters for Life Style Index

5) Store indexes in Hbase: Keyword Counts for above indexes as well as other personal and keyword information are stored for each tourist as column family in Hbase [38].

Key	Column family: keywords					
Name	Toured	Bought	Watched	Flu	Diabetic	Jetwin

Table 8: Store indexes in Hbase

C. Design

It enables with visualization tools to track personal, business and travel and tourism metrics. It derives data from multiple enterprise data sources and integrated with the Department of Immigration and Emigration database which is an OLTP Database.

Data Ingestion layer collects data of current Visa Holders from different modes including collecting personal information, Bio metric data, and visa details, flight details from Department of Immigration and Emigration database and collecting social networking data from data sources. Basically it proceeds as a centralized repository for user Meta data which is embedded with pre-store social data repository. Social network sites can be utilized to see the potential of certain tourists.

The Dashboard's Contextual Search Criteria is based on information gathering process of proposed solution. It allows the users to define their search criteria. For instance, end user can use the given name of a tourist as well as name and personal details. In order to gather accurate information best solution would be to search standard social media web sites like Facebook and LinkedIn. Tourism sites also can be used but it would be less helpful if comments are made anonymously. The crawling algorithm should take all precautionary steps to ensure the searched data belongs to

correct individual. Hence for the proposed solution, discover content provided via a centralized search engine and federated data sources. For an instance, personal details taken from an integrated OLTP database is crawled.

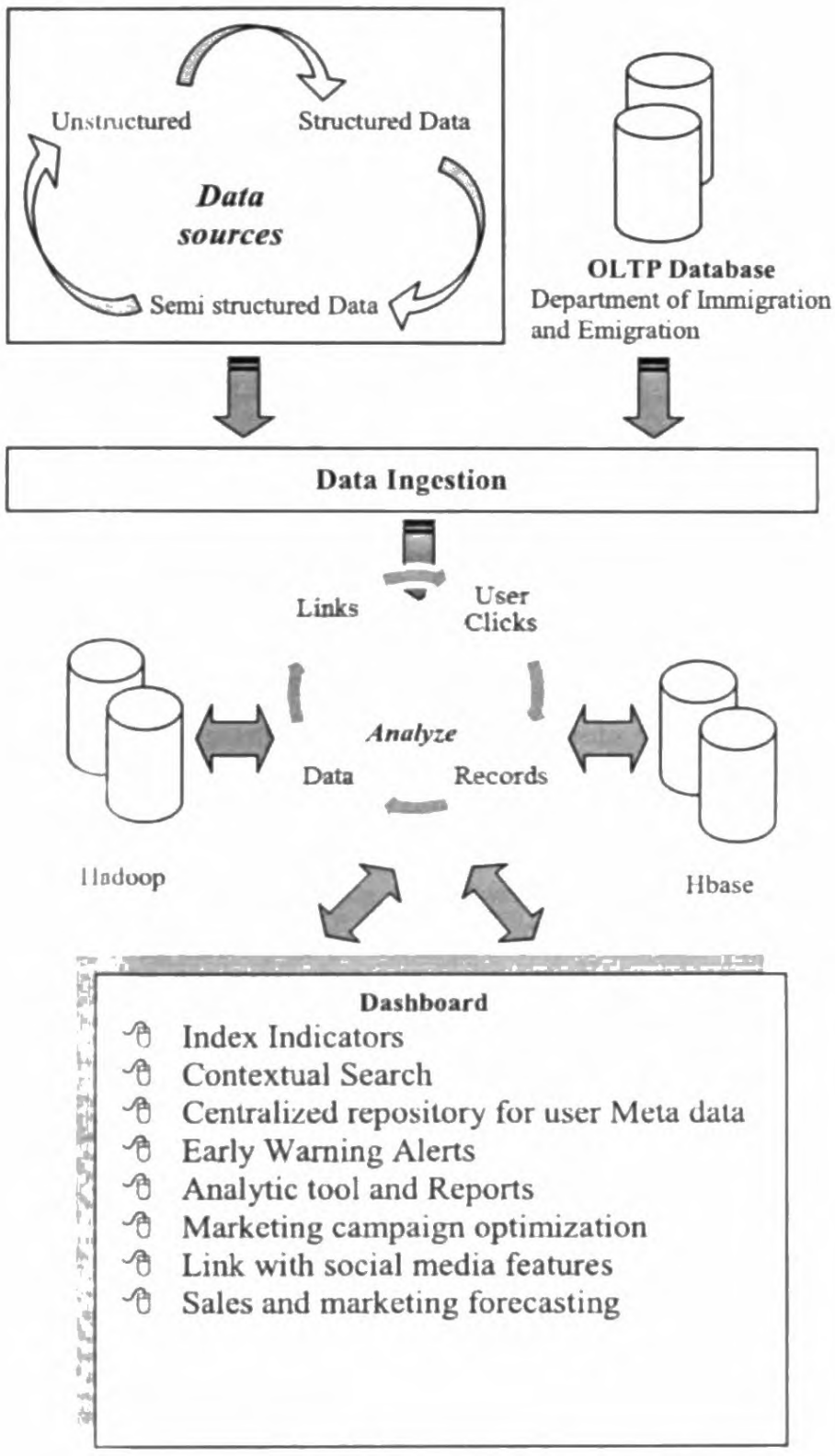

Figure 2 Solution Design Process

Dashboard enrich with early warning alerts and which is based on the pre-store social data on tourists , information which is given by tourist himself and any other information communicated through regulatory institutes. In particularly end users can see alerts via dashboard by searching a specific tourist. For instance, based on the alerts visa officer can decide whether a tourist can enter the country or not.

Index indicators are one of the major functionality in the proposed dashboard. Initial information gathering process is defined based on information which is available on web. This method allows extended solution to gather data including basic information, needs, desires and preference of a tourist.

Searching information for a particular tourist on web is not an easy task. The selection of a unique key is one of the most critical decisions when deciding key criteria. Therefore constraints should ensure that the selected keys are unique.

In particular, it focuses on the issue of searching tourist information on web by using only "Name" as key word and also it is a well-known fact that the similar name refers to a different person. It is not coherent to use only name as constraint when designing Index for the solution. Therefore key criteria defined based on collection of constraints including fields like full name, age, country, address, and passport number. That may be useful in cross checking although it is highly unlikely that a person has published all these details in sites accurately.

The basic information shown for any person like age, race, gender, and country are the main details taken. Also the published date of any information is very important as indexes are compared over the time.

The main task in index calculation is to get information for the four indexes.

- Visit Index
- Wealth Index
- Health Index
- Lifestyle Index

Visit Index is the probability of a tourist to visit a particular client on his tour. This is mainly done using keywords used in the tourist's blogs. If a tourist has indicated that he is visiting the client then the visit index becomes 100. If the person has commented that his lifestyle matches with clients business and he is visiting Sri Lanka then the visit index is 50. Likewise sophisticated algorithms have to be designed in order to forecast the probability of a particular tourist visiting the client site.

Wealth Index is the index showing his wealth status. If a person has keywords saying his salary is Millions or his expenditure is Millions the index will be increased accordingly. Also standard economical sites can be used to find whether this person is listed as a wealthy man we can increase the index. Also his educational qualifications and professional data are gathered and if he is highly qualified and is occupied in a good profession, the index can be increased.

Health Index is another index we calculate. Any person's health is very important in deciding to select as as an investment for a particular client.

While for a health care provider an unhealthy customer may be a better prospect while the normal tourist hotels will look for healthy persons. It is possible to parameterize these indexes and its effect on total efficiency index, so a health care provider can customize the calculation accordingly. Depending on the requirement, algorithms involved for the normal tourist care providers can be defined.

Lifestyle index is another factor that is analysed to understand the purchase decision patterns of a tourist. Mainly this is the index that is customized according to the tourist care provider. An environment friendly hotel will look for environment concerned tourists where as an Eastern Food specialist will look for tourists who like Eastern cuisine. Also a particular client may add more weightage to this index and influence the Total efficiency index more towards lifestyle.

Considering each of these four indexes, total efficiency indicator is calculated for each tourist. Also the average total efficiency indicator is calculated for age, race, gender, country of a tourist and this is another guideline to compare a particular tourist's efficiency compared to his peer groups. Also these four indexes as well as the total index are shown over the time for a tourist and the trend as well as future predictions can be identified from this.

Analytic tools and reports which embedded into dashboard capture most innovative statistics with open architecture. Because it is enrich with Index indicators and big data technologies such as Hadoop and Hbase.

In addition to gathering information about an individual tourist, end user can track and leverage data on the behaviour of tourist based on clickstream data from the web and data on historical purchases. Basically proposed dashboard can utilize as a predictive analytics models to make better decisions, reduce risks, to analyse trends, deliver more personal tourist experiences.

D. Analyse and Justification

There are number of technologies which were introduced with the evolution of Big Data. Database vendors demonstrate the advantages of their products along with hardware and software configuration. Therefore it is not easy to choose appropriate technology or databases for a particular case study. In order to identify suitable technologies for the proposed solution design a brief vendor independent comparison research based on productivity, performance, cost and effectiveness is carried out. The proposed solution designed based on Hadoop Distributed File System (HDFS), MapReduce, HBase, Apache Sqoop and Apache Mahout.

Apache Hadoop is an open-source implication [15], [16]. As a result, the proposed solution can promotes for free redistribution and allows access an end solution's design and implementation. One of the benchmark of the proposed output is to design a solution within the budgetary constraints. There should be strategies for collecting massive amounts of data from multiple sources. Hadoop enrich with facilities to store, analyze and access massive amounts of data from variety of sources across clusters of commodity hardware[34][35].In addition to that, Hadoop integrated with components including Hadoop Distributed File System (HDFS) and Hadoop MapReduce. Therefore Apache Hadoop is the suitable platform for the proposed solution. Apache Mahout use as a supporting platform for research and have faced both of these issues in employing it as part of our work in collaborative filtering [17].

As explained earlier, contextual search, early warning alerts, index indicator features are designed to execute queries and other batch read operations beside massive datasets. Thus, MapReduce is a high-performance parallel data processing engine. Ideally it is a suitable framework for processing large amounts of data in parallel on large clusters of commodity hardware in a reliable and fault-tolerant manner. With the growth of the data it has ability to potentially scale up to thousands of nodes [35][39].

The proposed solution enriches with the feature to capture trillions of bytes of information about tourists, and embedded into Social networking data, Department of Immigration and Emigration database. Therefore solution has an ability to host very large tables, store various types of data including structured, semi-structured and un-structured data. Also it has to captured, communicated, aggregated, stored, and analyzed. HBase provides better solution in this context than other NOSQL databases like Cassandra, Voldemort, Redis and VoltDB [15] [35].

When comparing to the scalability of database Hbase provides fault tolerant way of storing large quantities of sparse data. Since it has been built on top of HDFS distributed file system and allows fast individual record lookups in files, with random real-time read or write access to data. HBase provides two run modes including Standalone and Standalone HBase and distributed [40] [38]. Depends on the requirement and with the minimum configuration has an ability to switch between two modes. Each table is stored as a multidimensional sparse map including rows and columns. One of the other major testing is on the performance with major volumes of data [41].

E. Extract Data For Dashboard

End user allows making decisions more quickly, has visibility into key metrics across visa details, personal information, social media details and visualizing accurate data via dashboard. They can have a better understanding of the effects of marketing efforts. End users can rely on proposed dashboard to as reliable source for finding person information and obtaining records about them. Find tourist by current name, maiden name, address, phone number, or email. Also include a country or city to narrow tourist search results. It retrieves billions of records to easily locate tourist individually and retrieving data almost instantly. End users can sign in to preview actual records of each individual.

As mentioned earlier similar name may be refer to a different person and nearly all email addresses are linked to more than one name. End users allow customize searching criteria depending on their requirement. For an instance if the end user selects email address as a searching criteria dashboard will show information about every person connected to that particular address. Therefore end user has option to select exact person and will retrieve accurate information.

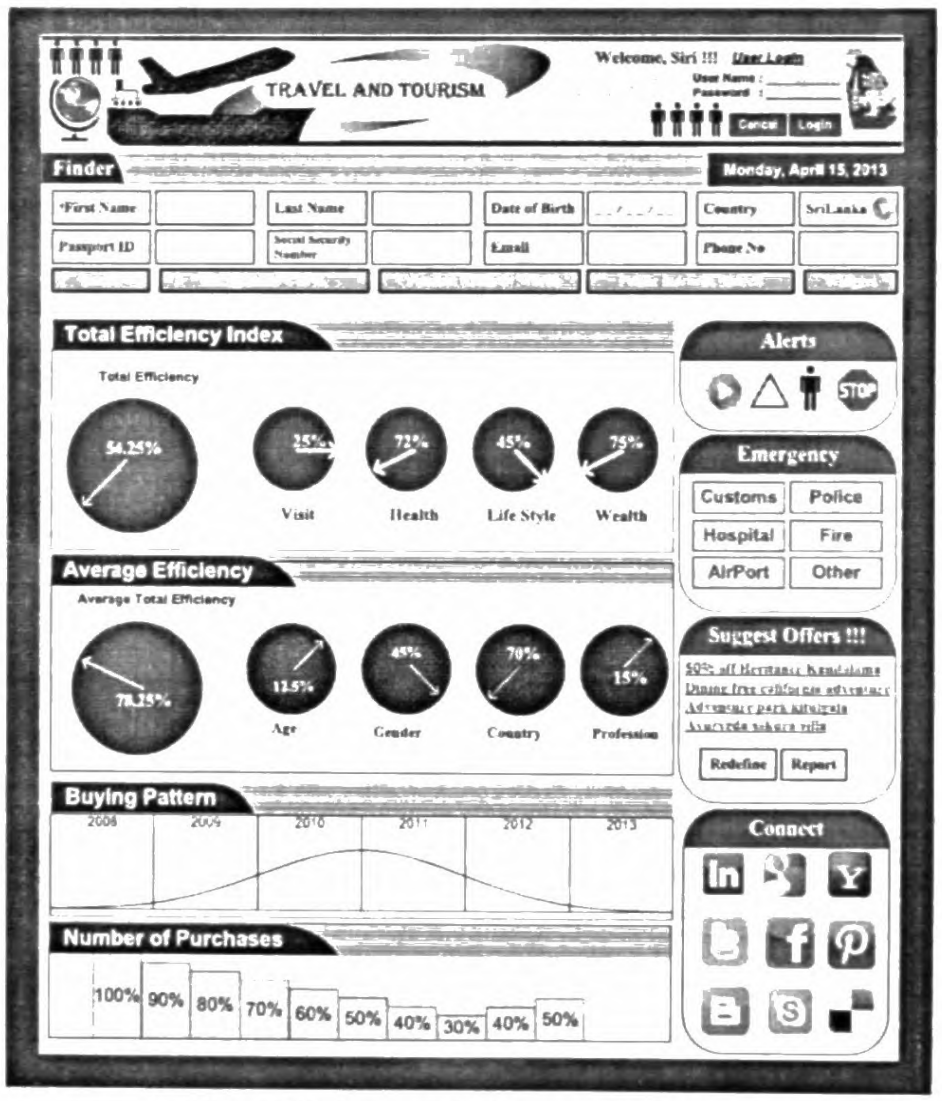

Figure 03: Proposed Dashboard Design

Furthermore dashboard can customize search tourist by phone number. That is one of the accurate and easy methods for finding the current holder of phone number and can

identify the accuracy of given details by visa holder. Also this can search public records associated with that particular phone number. Social Security number (SSN) can be used as one of the searching criteria to follow individuals' accounts within the Social Security program. End users can verify comprehensive Background Check via dashboard. For instance visa officer can check if anyone has a criminal record or finding out if someone has gone through bankruptcy in their history.

Basically dashboard shows everything that should be known about anyone before entering to the country obtaining useful facts before making important decisions. If there is any security concern it will show in alert section and display as security warning alert.

Total Tourist Efficiency index is reflecting the investment capability of a tourist. This index makes clear how efficient a tourist as a whole for a purticular client.

Total Efficiency Index = (Visit Index + Health Index + Wealth Index + LifeStyle Index)/4

Average Efficiency Indexes are defined for group of tourists of same criteria. Criteria can be age, gender, country of residence, ethnicity etc. This is a good indicator to compare individual tourist efficiency with that of the average of each criteria he belongs to.

Average Efficiency Index = (Sum of Total Efficiency Indexes) / no of tourists
Buying pattern is the buying frequency and amount spent in which a tourist purchase goods or services in a period of time

VI. TESTING

One of the major problems in much invested big data projects is the lack of confidence due to non-practical results. Therefore a sound quality control, using standards and a quality check with comparison testing of results with real web sites is very important. During the testing three dimensions (volume, variety and velocity) [42] and each of the proposed phases of big data processing need to be tested. Test plan has been proposed as follows;

A. Validating based on the characteristic of big data

Testing high volume of data is very complex task and needs faster approach. Testing approach needs to carry out by using number of datasets sampling strategy based on the data requirements. These Structured data needs to test by comparing data using compare tools. Therefore data discrepancies can be identified. Converted semi structured data into structured format since data crawling from different website to the dashboard it does not have defined structure to validate. Subsequently validation should carry out by comparing expected data with actual data. There is no define format for unstructured data like social media data, web pages, web log files, search indexes, e-mail, documents. It converted into structured data by using Pig scripting [43] and validated aggregated data against the data output. Performance testing needs to be out in order to test data speed.

B. Functional Testing

1) Validating data crawling and scraping phase: During the testing incorrect data can be captured when loading source data files into HDFS. Therefore it needs to initially validate the data requirement and compare the input data file and source data including social data, streaming data, and structured data from OLTP database. Subsequently needs to validate that whether the data files are loaded into HDFS appropriately.

2) Validating MapReduce Phase: Coding issues in map reduce jobs can be occurred therefore developers needs to highly concentrate to identify and fix issues. Business logic, Data process, map reduce process, aggregation, output files and file format need to be validated during this phase.

This solution heavily involves java code. It is important that code is written according to java standards and best practices to ensure proper functioning of the application. Java unit testing of this application is out of scope.

3) Validating Analysing and storing phase: Once the data from data sources loaded in to HDFS and map reduce process is completed, processed data will move to this phase. Data transformations are very complex and it needs more processing time. Validation needs to consider in terms of data integrity and data quality. Hence transformation rules and aggregation of data needs to be validated.

This solution involves inserting and selecting of data from NoSQL databases like Hbase. Still there are no hard and fast standards and best practices for NoSQL based data operations. Even though it is in the very scope of this paper the time factor has not enabled authors to explore much on this area. Authors plan to explore standards and best practices of NoSQL data operations in a future research paper.

4) Validating decision phase: During this phase considered that whether the fetched data appeared as expected. Dashboard data and reports needs to be validate by ensuring whether web data are up to date and available.

C. Non Functional Testing

Performance Testing is another main area authors need to pay a lot of attention. Bigdata involves high volumes, high variety and high velocity in data which is vulnerable for performance issues if any of these attributes are changed. Therefore millions of data needs to be prepared and loaded into the application and complete cycles needs to be test. Also as mentioned earlier there are different varieties of data. Therefore data needs to feed to the application in high speed simulating a real world scenario where web pages will rapidly update. Any issues in performance will rectify from application, hardware, network and other factors continuously to get real value from big data

HDFS architecture is designed to detect and recover to proceed with in the processing for three common types of failures including NameNode failures, DataNode failures and network failures [44]. However validation should take place when switch over in to other data node to ensure failover testing.

D. User Acceptance Testing (UAT)

UAT involves significant participation from end user and authors did lot of work in preparing test specifications as this

is very important for the end user confidence. We checked whether the solution is functionally fit for use and behaves as expected; validate end-to-end business process, user access, data availability, integrity and quality. The method of this testing mainly is on comparing the end result with that of the source web sites. For example let's get a person whose visit index to Sri Lanka is very high. Testers have to manually access the web sites involved and see whether they reflect these visits. Also they have to compare this with different users and see whether the visit index for this user is reasonable compared to web published data and other users. In this way all the indexes has to be tested with sample users and also see whether average indexes are realistic and also other trends are correctly shown in web.

VII. Discussion

As discussed in the paper this research was based mainly on the end users like tourist hotels and tourist guides etc. Therefore it was essential that we get the feedback from them on the solution design as well. We got their feedback mainly from the dashboard prototype which was a very good tool for them to get good idea on our design.

Users were very enthusiastic on this kind of a solution. Their main worry was whether they will have to pay for this application. They knew that we were based on readily available data in internet and some of them were already using some web pages to get an idea of their tourists. So it was not an easy task to convince a new application. But the rich interface and the comparative indexes authors designed were making sense to them. After looking at the return on investment most of them were happy on this venture.

One of the main valid points was whether we can rely on web data. As they pointed out some of them had relied on these web information earlier and had mixed results. Some tourists sites were publishing genuine data where as some others were publishing fake data which we also had to agree. Therefore what they were pointing was rather than relying on data that is published by tourist themselves we have to go for some independently confirmed data source, since they have to subscribe also. Authors pointed out the facility of integration with the Department of Immigration and Emigration which users welcomed and were suggesting further to get integrated with official authorities of countries of relevant tourist as well. This was a good suggestion and authors think this is possible if they can get the sponsorship from government of Sri Lanka on this venture.

Bigdata solutions mainly go hand in hand with datawarehousing solutions. Software companies market them as cheap options for datawarehousing due to the open source stack it uses. But penetrating into the market with this strategy is difficult when there is an existing datawarehousing solution. Client thinks he already has the given solution using the datawarehousing data and if he has already spent on software he may not be interested in going for a new solution.

Due to this problems in penetrating into the market with datawarehosing solutions authors came up with this novel idea of integrating bigdata solutions to On Line Transaction Processing Solutions. Rather than marketing this as a Management Information System authors are proposing to get additional data using this solution for front office transaction systems.

One of the main attractive points in this solution is rather than working on historical data of the available databases authors is going for social media to explore additional and independent data. This was the main key factor in introducing bigdata into tourist sector as this impressed the tourist operators. Even though this data is known to the end user they readily accepted the way authors were trying to analyze this.

Even though bigdata is a buzzword the use of it is very limited and this a practical solution design which will attract users to use this technology.

Authors' idea was to implement this solution design before the ICTer conference for the presentation. But due to other full time commitments they did not have time to show a fully executable prototype of this solution. Especially both authors being database specialists it was no easy task for them to start on a venture of a java based project. Therefore the solution will be limited to the total solution design with fully tested hadoop and hbase components.

We would have validated the data files in different data nodes in order to complete the validating process during the data crawling and scraping phrase. In the testing, we mainly tested standalone node only. Different issues may occurred when validating the map reduce process run on multiple nodes.

VIII. Conclusions

In this paper we proposed a solution design for a dashboard which is integrated with big data technologies. Different types of big data technologies were analyzed and identified. After evaluating stakeholder requirement and possibility of implementation proposed solution were designed based on Hadoop.

Four fusion logics were applied to design main index indicators. Index Indicators, Contextual Search, Centralized repository for user Meta data, Early Warning Alerts, Analytic tool and Reports, Marketing campaign optimization, Link with social media features, Sales and marketing forecasting are the main functionalities will be embedded in the proposed dashboard.

The proposed personal dashboard enriches with customized features and enables end user to get summarized information about a tourist from different sources. On other hand it is very useful to take decisions in various aspects. The concept has been applied in one of Sri Lanka's most emerging industry named, travel and tourism to show the practicality of this technology to anybody.

Acknowledgment

Our gratitude is expressed to Database Competency Excellence Group (DB-CEG) in Virtusa and the support provided by Big Data Proof of Concept team (POC) who was sharing their expertise coding knowledge in implementing the solution.

References

[1] T.Lock (2012) The Register: The Big Data revolution.[Online].Available:http://www.theregister.co.uk/2012/10/08/big_data_revolution/

[2] R.Henschke. (2009) Asia Calling: Sri Lanka Calls for Tourism DevelopmentBoom.[Online].Available: http://www.asiacalling.org/en/special-reports/after-the-war-the-hard-work-begins-in-sri-lanka/1441-sri-lanka-calls-for-tourism-development-boom

[3] J.Manyika, M.Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H Byers, (2011) Big data: The next frontier for innovation, competition, and productivity. California: McKinsey Global Institute [Online].Available:http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[4] J. Kelly, D. Vellante, and D. Floyer, (2013) Wikibon: Big Data Market Size and Vendor Revenues. [Online]. Available: http://wikibon.org/wiki/v/big_data_market_size_and_vendor_revenues

[5] M. Fcuz, M. Fuller and F. Stalder (2011) Personal Web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalisation [Online]. Available: http://firstmonday.org/article/view/3344/2766

[6] Research and Internationalrelations Division. Sri lanka tourism development authority: annual statistical report (2011) [online]. Available:http://www.sltda.lk/sites/default/files/Annual_Statistical_Report-2011.pdf

[7] P. Russom, Big Data Analytics, TDWI Best Practices Report, Fourth Quarter, 2011.

[8] P.Zikopoulos and C.Eaton, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data" McGraw-Hill, 2011.

[9] Google/OTX(2011)Traveler's Road to Decision 2011:Google/IPSOS OTXMediaCT,USA.[Online].Available: http://www.thinkwithgoogle.com/insights/emea/library/studies/travelers-road-to-decision-2011/

[10] T.H. Davenport, (2013) At the Big Data Crossroads:turning towards a smarter travel experience: Amadeus IT Group .[Online]. Available: http://2013.amadeusblog.com/wp-content/uploads/2013/06/Amadeus-Big-Data-Report.pdf

[11] S.Mitra, (2007) Web 3.0 and Travel Search Engines. [Online]. Available: http://www.sramanamitra.com/2007/06/01/web-30-travel-search-engines/

[12] Ventana research (2012) The Challenge of Big Data: Benchmarking Large-Scale Data Management, California: USA. [Online]. Available: http://www.ventanaresearch.com/uploadedFiles/Content/Landing_Pages/Ventana%20Research%20Benchmark%20Research%20Big%20Data%20White%20Paper%202012.pdf

[13] D J. Abadi, D.S.Myers, D.J.DeWitt,and S.Madden , Materialization Strategies in a Column-oriented DBMS. In: Proc. of ICDE (2007) pp.466–475.

[14] H.Chen, R.H. L. Chiang and V.C. Storey, Business intelligence and analytics:from big data to big impact: Business Intelligence Research ,MIS Quarterly Vol. 36 No. 4, 2012, pp. 1174-1175

[15] T.White, Hadoop: The Definitive Guide: Storage and Analysis at internet scale, CA, USA: O'Reilly Media, 2012.

[16] K.Ting, and J.J Cecho, Apache Sqoop Cookbook: Unlocking Hadoop for your relational Databases, CA, USA: O'Reilly Media, 2013.

[17] I.Drost, Scaling Data Analysis with Apache Mahout, CA, USA: O'Reilly Media, 2011.

[18] J. Lin, Exploring Large-Data Issues in the Curriculum: A Case Study with MapReduce (Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics), Ohio: USA, Association for Computational Linguistic, 2008, pp 54–61.

[19] R. Baraglia, G. D. F. Morales, and C. Lucchese. Document similarity self joins with MapReduce. In ICDM, 2010.

[20] Y. Kim and K. Shim. Parallel top-k similarity joins algorithms using MapReduce. In ICDE, 2012.

[21] Gigaspaces (2012) "Big Data Survey: Real-Time Stream Processing and Cloud-Based, Big Data Increasing in Today's Enterprises": USA. [Online].Available:http://www.gigaspaces.com/sites/default/files/product/BigDataSurvey_Report.pdf

[22] Explore Srilanka, Laya: Comfort, Peace And Serenity (2012) [Online]. Available: http://exploresrilanka.lk/2012/12/laya-comfort-peace-and-serenity/

[23] S. Few and P. Edge (2007) Why most Dashboards Fails [Online]. Available:http://www.perceptualedge.com/articles/misc/WhyMostDashboardsFail.pdf

[24] S.Rosenbush, and M. Totty (2013) U.S. edition of The Wall Street Journal, with the headline: How Big Data Is Changing the Whole Equation for Business [Online]. Available: http://online.wsj.com/article/SB10001424127887324178904578340071261396666.html

[25] Research and International relations Division. Sri lanka tourism development authority: annual statistical report (2011) [online]. available:http://www.sltda.lk/sites/default/files/Annual_Statistical_Report-2011.pdf

[26] Data Protection Acts 1988 and 2003: A Guide for Data Controllers [Online].Available:http://www.dataprotection.ie/documents/forms/NewAGuideForDataControllers.pdf

[27] L.Wijesiri (2012) DAILY NEWS: Developing tourism in Sri Lanka and challenges [Online]. Available: http://www.dailynews.lk/2010/02/27/fea03.asp

[28] NewVantage Partners: Big Data Executive Survey (2013) [Online]. Available: http://newvantage.com/wp-content/uploads/2013/02/NVP-Big-Data-Survey-2013-Summary-Report.pdf

[29] R. E. Bryant, R.H. Katz and E. D. Lazowsk, "Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society" Version B, 2008, pp 2-3.

[30] The Authority on World Travel & Tourism: Travel & Tourism Economic Impact 2012 WORLD (2012) [Online]. Available: http://www.wttc.org/site_media/uploads/downloads/world2012.pdf

[31] An Oracle White Paper in Enterprise Architecture—Information Architecture: An Architect's Guide to Big Data (2012) [Online]. Available:http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

[32] jsoup: Java HTML Parser [Online]. Available: http://jsoup.org/

[33] P.Houston, "Instant Jsoup How-To: Effectively extract and manipulate HTML content with the JSoup Library", Packt Publishing. Birmingham: UK, 2013.

[34] Integrating Hadoop with Enterprise RDBMS Using Apache SQOOP and Other Tools (2011) [Online]. Available: http://www.hadoopworld.com/session/integrating-hadoop-with-enterprise-rdbms-using-apache-sqoop-and-other-tools/

[35] H.Liao, J.Han and J. Fang (2010) Fifth IEEE International Conference on Networking, Architecture, and Storage: Multi-dimensional Index on Hadoop Distributed File System (2010) [Online]. Available: http://www.cs.odu.edu/~mukka/cs775s11/Presentations/papers/liao.pdf

[36] G.Satell (2013) Why Facebook's Graph Search Really Does Matter: Big Data + NLP [Online]. Available: http://www.forbes.com/sites/gregsatell/2013/02/04/why-facebooks-graph-search-really-does-matter-big-data-nlp/

[37] The Blog of the International Computer Science Institute:Big Data or Expert Annotation - What's Best for Natural Language Processing? (2013). [Online]. Available: http://www.icsi.berkeley.edu/icsi/blog/data-versus-experts

[38] The Apache Software Foundation. Apache HBase. [Online]. Available: http://hbase.apache.org/.

[39] K.Shvachko, H. Kuang, S.Radia, and R. Chansler (2010) The Hadoop Distributed File System. O'Reilly Media, Yahoo! Press.

[40] J. Dean and S. Ghemawat (2004) MapReduce: Simplified Data Processing on Large Clusters.Vol 06.

[41] Avinash, Lakshman. Cassandra-A Decentralized Structured Storage system. In LADIS, 2009.

[42] M.StoneBreaker.SQL databases V. NOSQL databases, Communications of the ACM, Vol. 53 No. 4, pp.10-11, 2010.

[43] J.Hurwitz, A.Nugent, F.Halper,and M.Kaufman,"Big Data For Dummies: Big Data management ", NJ, USA: John Wiley & Sons,2013.

[44] C.Lam," Hadoop in action: Programming with Pig" Manning Publications, 2010.

[45] D.Borthakur. (2008) Hadoop 1.2.1 Documentation: HDFS Architecture Guide [Online]. Available: http://hadoop.apache.org/docs/stable/hdfs_design.html

[46] Big data and analytics in travel and transportation: Beyond the hype: Solutions that deliver big value, IBM Corporation, 2013.[Online].Available: http://public.dhe.ibm.com/common/ssi/ecm/en/gbw03215usen/GBW03215USEN.PDF