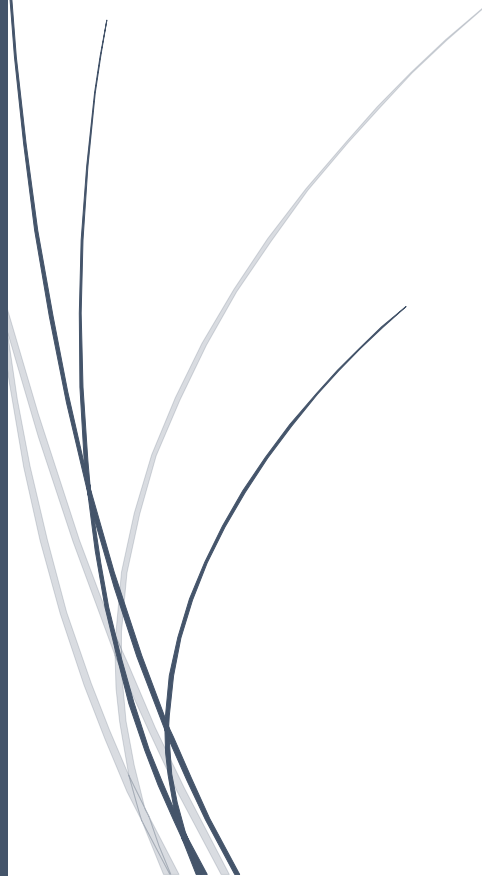


# BI Implementation for TripAdvisor



Team 1:  
Jajoo, Archit  
Desai, Ameya  
Gandhi, Ankur  
Joshi, Aryana  
Kamdi, Anupama  
Travis, Brandon

## Table of Contents

Overview	2
Description of Organization	2
Functionality and features	3
Data Model	4
Data Analysis	5
Extract, Transform and Load	7
Architecture	9
Budget Plan	12
Conclusion and Recommendation	13

## Overview:

TripAdvisor is a well-known company which helps customers to decide where to travel, what to eat, what new to discover in any area etc. In other words, it helps them to check hotel reviews, restaurants, vacation spots and decide accordingly. TripAdvisor is a very good website for travel planning, but when it comes to flight bookings, they partially fail to excite the popularity like their hotel or restaurant bookings. We propose to implement a BI project that focuses on the absolute market share KPI for flight bookings on TripAdvisor. This KPI was chosen over other possible choices because it would be good for the company to see how well they are doing in flight sales compared to other flight booking agents. This information could assist TripAdvisor in increasing flight bookings on their site and, in turn, their overall market share in this sector.

## Description of Organization

TripAdvisor is an online, travel-related site used for finding and booking vacation packages and other deals at a discounted rate and also for reviews of hotels and restaurants. We believe that TripAdvisor needs BI/reporting to better help the organization track its market influence compared to the many other discount-travel booking sites. BI/reporting could also assist the company in improving its customer service.

Key existing systems relate to sales, marketing, finance, analytics etc. TripAdvisor is a huge organization so it would be difficult to analyze existing operational data. There are very few gaps between current analytical capabilities and desired needs since TripAdvisor uses Hadoop, Hive, R, SQL, Redspark and similar other various software's to control and update the content accordingly. The organization is very advanced and known to be one of the best in several factors in travel planning. Here we try to get the report analysis of customers who opt for hotels booking but don't book the flights from TripAdvisor.

		Common Dimensions						
		Tickets Sell	Date	Other booking	Promotion	Discounts	Repeat Customers	Emp. booking
Business Processes	Daily Purchases	X	X	X	X	X	X	X
	Forecasted	X	X		X	X		
	Direct booking	X	X		X	X	X	X
	Alternative booking	X	X	X	X	X	X	
	Ticket Cancelled		X					
	Insurance Coverage	X	X	X				X
	Agent booking	X	X	X	X	X	X	
	Domestic flights	X	X		X	X	X	X
	International flights	X	X		X	X	X	X
	Package booking	X	X	X	X	X	X	X
	Individual purchases	X	X		X	X	X	X
	Tickets blocked	X	X	X				X

### Functionality and Features:

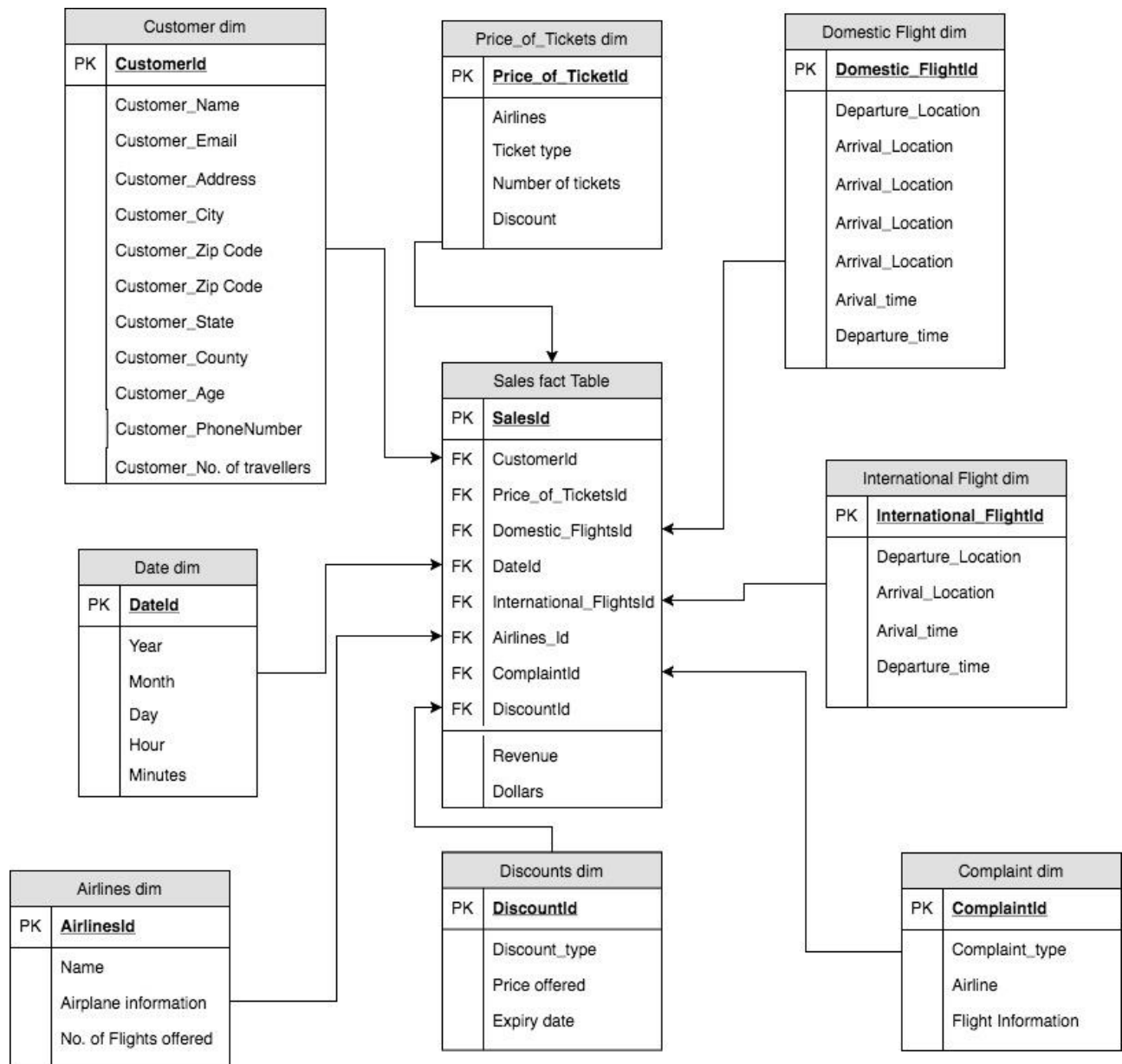
The proposed BI system to be implemented that will analyze TripAdvisor's absolute market share KPI for flights should be able to produce particular reports and analyses. We expect to generate standard reports that could include YTD flight sales, Flight Sales by Volume as a Percent of Total Sales, etc. These reports may be scheduled to run on a certain day of each week so that that week's data can be reviewed and analyzed. We

expect users of these reports to include executive management who are looking for an overview of a particular week's or day's data.

The BI analyzed would be used by the management team, along with the marketing and sales team in providing a useful solution to help increase the revenue. The report will help them to understand, which area they need to focus on more in order to increase flight bookings. Also, it will compare the data with other websites which are doing better in selling airline tickets. By comparing these data, it will help make changes to improve the sales of ticket on TripAdvisor. Major applications that will be used to extract data from the data warehouse will include SQL and Informatica and few others.

### **Data Model:**

The data model implemented in the system is star schema which has 1 fact table, which is 'Sales Fact Table' and 8-dimension tables. Since, the key performance indicator is absolute market share, the Revenue attribute has been added in the fact table so that the data analyst can extract information and calculate the absolute market share. To avoid the complex design, snowflaking has been avoided. As a result, separate dimension tables are created for domestic flights as well as international flights. The date dimension table is included in the star schema and is set at the daily grain.



## Data Analysis:

The data for building a data warehouse will be from different operational and non-operational sources. The various sources are ERP (Enterprise Resource Planning), Historical and Summarized data, Customer related information, Various promotional information and various external data sources such as Marketing, Purchase order sales and many others.

The operational data will be from ERP system having sales information, gap analysis. This will have all the monetary information done in various transaction processing like hotel reservation, flight bookings and many other. While the non-operational information will be from the sources like Customer related information which includes customer registration and contacts and all the different reviews of different hotels and locations done by customers which can be used for making decisions.

The Data set referred from TripAdvisor website ([www.tripadvisor.com](http://www.tripadvisor.com)) is Customer Reviews data set which is nonoperational as it is not used for actual transaction processing purposes but can be used for making decisions for customers.

An example of the data from the dataset is as shown below:

{"Ratings":

  {"Service": "4", "Cleanliness": "5", "Overall": "5.0", "Value": "4", "Sleep Quality": "4",  
  "Rooms": "5", "Location": "5"},

  "Author Location": "Boston",

  "Title": "\u201cExcellent Hotel & Location\u201d",

  "Author": "gowharr32",

  "ReviewID": "UR126946257",

  "Content": "We enjoyed the Best Western Pioneer Square. My husband and I had a room with a king bed and it was clean, quiet, and attractive. Our sons were in a room with twin beds. Their room was in the corner on the main street and they said it was a little noisier and the neon light shone in. But later hotels on the trip made them appreciate this one more. We loved the old wood center staircase. Breakfast was included, and everyone was happy with waffles, toast, cereal, and an egg meal. Location was great. We could walk to shops and restaurants as well as transportation. Pike Market was a reasonable walk. We enjoyed the nearby Gold Rush Museum. Very, very happy with our stay. Staff was helpful and knowledgeable.",

  "Date": "March 29, 2012"

}

The data is in JSON format i.e. (Java Script Object Notation) which is easy for data parsing and analysis for computers.

The level of granularity for the data is fine grained, as it gives detailed information of each review. The various details of the reviews which makes it fine grained are: Various individual ratings such as Service, Cleanliness, Overall, Value etc., information about author i.e. location, author id.

Data will require cleansing because sometimes data is not in proper format and may have incorrect values which could result in loss of data integrity and accuracy. This can in turn lead to creation of a faulty system. Also, data is coming from various sources which needs to be converted into a single generic format for orderliness. So, there is need for data cleansing.

For performing any specific operation, necessary information can be extracted from the dataset. Also, various drill down and drill up operations can be performed on this data set by using numerous factors such as Location of Hotel, reviews done by authors etc.

### **Extract, Transform and Load:**

Informatica PowerCenter was chosen as the major ETL tool since it is capable of handling the large data volumes that may sometimes come as an unbearable challenge for SQL developers for the TripAdvisor DW/BI system. This software was chosen over the many other available options because of its track record in the industry, its straightforward learning curve, and its ability to handle real-time data.

### ***Extracting***

The Informatica tool will extract data from multiple operational and non-operational systems and be able to generate surrogate keys and perform key lookups. The systems for data extraction will include a master Customer relational database management system (RDBMS), a master Geography RDMBS, and an Invoice COBOL flat file. In



addition to the ETL tool, Trillium Quality specialized data cleansing software is the selected software that will be purchased from a leading vendor in this particular industry, Syncsort, who recently purchased the software from Harte-Hanks. This “dirty data” cleansing tool will be implemented for its usefulness in correcting customer names, addresses, phone numbers, etc. to help maintain the integrity and quality of the data.

One of the more crucial issues that will be run into is the initial extraction of the data from these systems that have been put into place by TripAdvisor to record customer and sale data. The data will be compressed to reduce transmission time during the extraction and to encrypt it for security during the transfer.

### ***Transforming***

In order for this proposed DW/BI system to work, some transformations will need to take place to prepare the data for analysis. Some of the more basic transformations that will take place will be the deduplication and survivorship when faced with matched data when extracting from multiple systems and the cleaning and mapping of the data for format and overall data consistency, such as with gender and dates. These tasks will be accomplished using the Trillium Quality cleansing software. Because TripAdvisor services customers worldwide, currency conversion for the systems have been considered. The format revision transformation type will make it possible to convert currencies to U.S. dollars. Slowly changing dimensions will be a requirement for the system to track data history over time and to store and manage current and historical data.

One advanced transformation that needs to occur is the combining and integrating of data that will be coming in from TripAdvisor’s multiple source systems. This is a vital requirement that will be needed in order to allow the data to be standardized and consistent so that proper conformed dimensions and facts are created and maintained. Another major transformation that needs to occur is the generation of aggregates so that multiple rows of data can be summarized by city, state, region, etc. The granularity of the

data will be set at the lowest level, the atomic level, because it is the most flexible and will allow for summation of the data or drilling down from the summary level.

### ***Loading***

After the data is cleansed and transformed it is all set up to get loaded into the data warehouse. Initially, the data will be populated into appropriate tables and will be checked if it is ready to use for getting required business insights. Different dimensions will be created such as Customer dimension to load all customer related information, airlines dimension to record all information related to airlines, price which will contain the price for flights for all airlines, time dimension which will help achieve the fine granularity for data warehouse etc.

As the data transformation is done externally, direct path APIs are to be used to perform load operation. Oracle's OCI can be used to load data into data warehouse. Once the data is loaded, it is very essential to check referential integrity between all the dimensions and fact table to ensure all records are appropriately associated with each other. To access this data from data warehouse SQL, PL/SQL and JAVA will be used.

### ***Scheduling***

For a company like TripAdvisor, it is very challenging to schedule tasks of loading and extraction as the data is not centralized and it has to be collected from the various sources. The sources are not only locally distributed but as the company serves globally, the data is collected from different parts of globe, from different time zones. It is very important to consider minute factors such as seconds of delays as the website is being used globally by millions of user. It is very important to schedule data loading and transitions at proper time in order to maintain data quality and to avoid redundancy.

These data loading tasks can be automated using job scheduling. Microsoft Service Manager enables user to make use of integrated platform to automate and adopt organization's IT service management practices. To improve data richness, TripAdvisor is advised to implement real time data loading techniques. It can use scheduling tools

such as MS Service Manager to schedule data warehouse loading tasks. To analyze the absolute market share KPI, TripAdvisor needs to keep track of everyday sales. This can be used to get monthly or quarterly revenues.

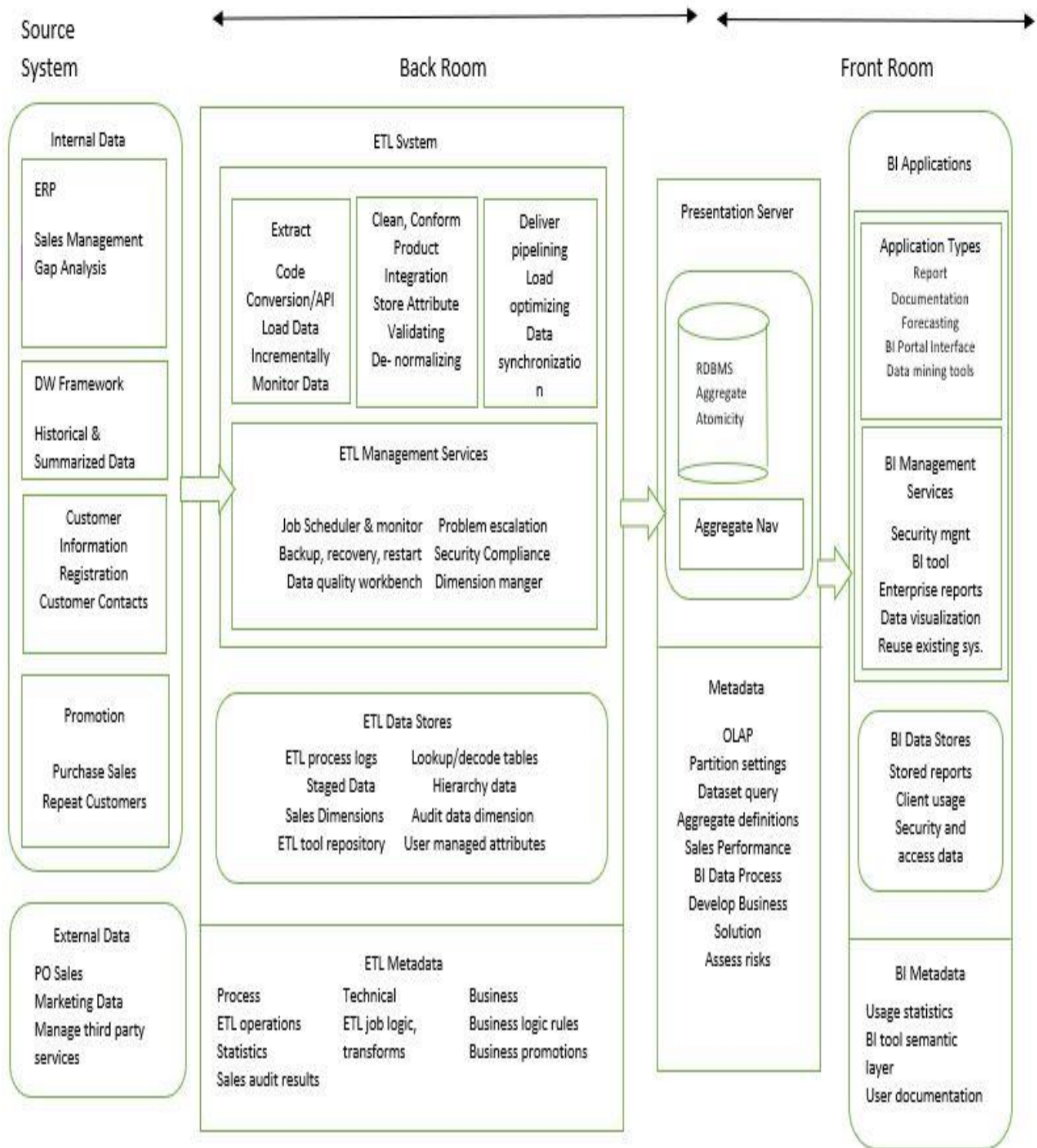
### Architecture:

The application architecture model of TripAdvisor conveys the flow and analyzes the data fetched from the data set which provides a BI analytical solution. In the source system, first collect the data from the external source. This data will be collected in the form of flat files and stored in CSV format. Then, pass it onto the ETL system stage, where it will extract the data, clean it, reform it. For this ETL tool like Informatica can be used. The main advantage of Informatica over some of the other tools like Teradata is, it can load data from anything, transform it and write to anything. Since, the project focuses on absolute marketing, Informatica allows us to transfer data from one application to another to determine how every small detail will help the company to become a high performer.

The next part of the architecture phase would be store all this information into RDBMS. For that purpose, tools like Stream can be used. The main use of stream set is that it allows to stream from MapR table to Mongo DB and vice versa. For querying purpose, on applications that use function calls and which controls all the phases of SQL statement execution, OCI is a good tool. Once a determined output is fetched from the queried data, it will then call the BI tool like Tableau, Oracle BI Publisher. Tableau as a tool gives the output in table, graphical, pie chart, etc. format which will produce the absolute marketing data more accurately. The organization will get a weekly game plan of prioritized leads that predicts each buyer's needs, including specific deals and services to highlight what action is to be taken next.

The BI Publisher will generate an overall report which will help to understand the statistics graphically along with its data. With the results obtained, it can find out the gap in those areas where it is falling behind such as promotion, discounts, points based gifts and so on to improve the sales of airplane ticketing for TripAdvisor.

## Application architecture model for Trip Advisor Ticketing Absolute Marketing



### Budget Plan:

ITEM	PRICE	No.*	TOTAL
Red Hat Enterprise Linux Server	\$349.00	500	\$174,500.00
Red Hat Jboss Enterprise Application Platform	\$8,000.00	1	\$8,000.00
MySQL Cluster Carrier Grade Edition Subscription (5+ socket server)	\$60,000.00	1	\$60,000.00
MongoDB Advanced Enterprise Edition	\$10,000.00	16	\$160,000.00
Java SE Suite	\$15,000.00	10	\$150,000.00
Cloudera Hadoop Enterprise	\$7,000.00	16	\$112,000.00
Databricks	\$1,000.00	10	\$10,000.00
StreamSets	\$8,000.00	1	\$8,000.00
AWS m4. 16xl	\$18,097.00	12	\$217,164.00
Informatica Power Centre	\$54,000.00	5	\$270,000.00
Tableau Enterprise Edition	\$840.00	100	\$84,000.00
Trillium Quality by Syncsort	\$150,000.00	1	\$150,000.00
Employees	\$100,000.00	100	\$10,000,000.00
Other Expenses	\$10,000.00	12	\$120,000.00
Service and Repair	\$10,000.00	12	\$120,000.00
<b>Total</b>			<b>\$11,643,664.00</b>

\*The No. column contains the no. of licenses, no. of users, months etc.

## **Conclusion:**

The report presents all the information management might need to initiate the DW/BI project. Although, the information provided here is theoretical, the practical implications may lead to some changes once the project has started.

This DW/BI project focuses on improving the absolute market share for TripAdvisor in the airlines segment i.e. flight bookings. This project will help the organization to increase traffic on their website and attract more customers. It will provide old as well as new customers with special offers, discounts etc. and will understand customer expectations for the airline ticket prices, simultaneously comparing it with other websites to provide better offers and trying to achieve 100% customer satisfaction.

The report describes the tools used for improving the overall process of flight bookings as well as maintaining customer information and following the buying trends from other websites as well as from TripAdvisor's website for airline ticket bookings. This project will also help the management to use data from other segments of business such as hotel bookings, vacation rentals, reviews etc. to provide better offers to customers when they are booking in any of these other segments. Also, the data can be used for performing various data mining techniques which will lead to maximizing profit for TripAdvisor.

## **Recommendations:**

Management should first make a decision whether to invest in data warehouse keeping in mind the cost and benefits of this investment by performing financial calculations and see how their profitability will increase and expenses will be recovered over the years. Looking at the current report it will be highly recommended to invest in data warehouse.

TripAdvisor should focus more on selling airline tickets to spread its name not only in hotel booking and restaurant review etc. segments but also in the airline segment. Although, this may affect the profitability initially but in the long run it would make TripAdvisor the best travel website in all segments. In addition to this, the company should provide special offers for flight bookings when a customer is looking for hotel bookings and give an option

of combined packages whose prices are well discounted when compared with other websites. Lastly, the management can focus on developing a reward based system for flight bookings which may or may not be similar to the hotel booking reward system that TripAdvisor currently has. These few recommendations can bring some major changes towards the overall profitability of TripAdvisor and there can be many changes which can be thought about since TripAdvisor is one of the biggest company in the travel segment.