DATA PIPELINE PROJECT

TRAVEL AGENCY RECORDS – XML MODELING

DATA PIPELINE 1

Joyee Banerjee - 20210300

Applied MSc in Data Analytics 2021-23

Data ScienceTech Institute

Data ScienceTech Institute

Table of Contents

I. Introduction

OVERVIEW

This report presents a summary of the course project which required us to model and implement an XML database of a tour operator. The instructions given in the assignment description have been followed while modeling the database. This project involved selecting a framework to base the modeling of the XML database, designing and fabricating the required XML Database, creating and validating the corresponding Schema, utilizing XPATH and XQUERY to create and write the XSL stylesheets and their corresponding XSLT transformations are presented in output formats of HTML, XML and JSON respectively. Additionally, the XML Database has also been presented in JSON and its corresponding JSON Schema has also been validated.

II. XML Database Modeling

XML TREE

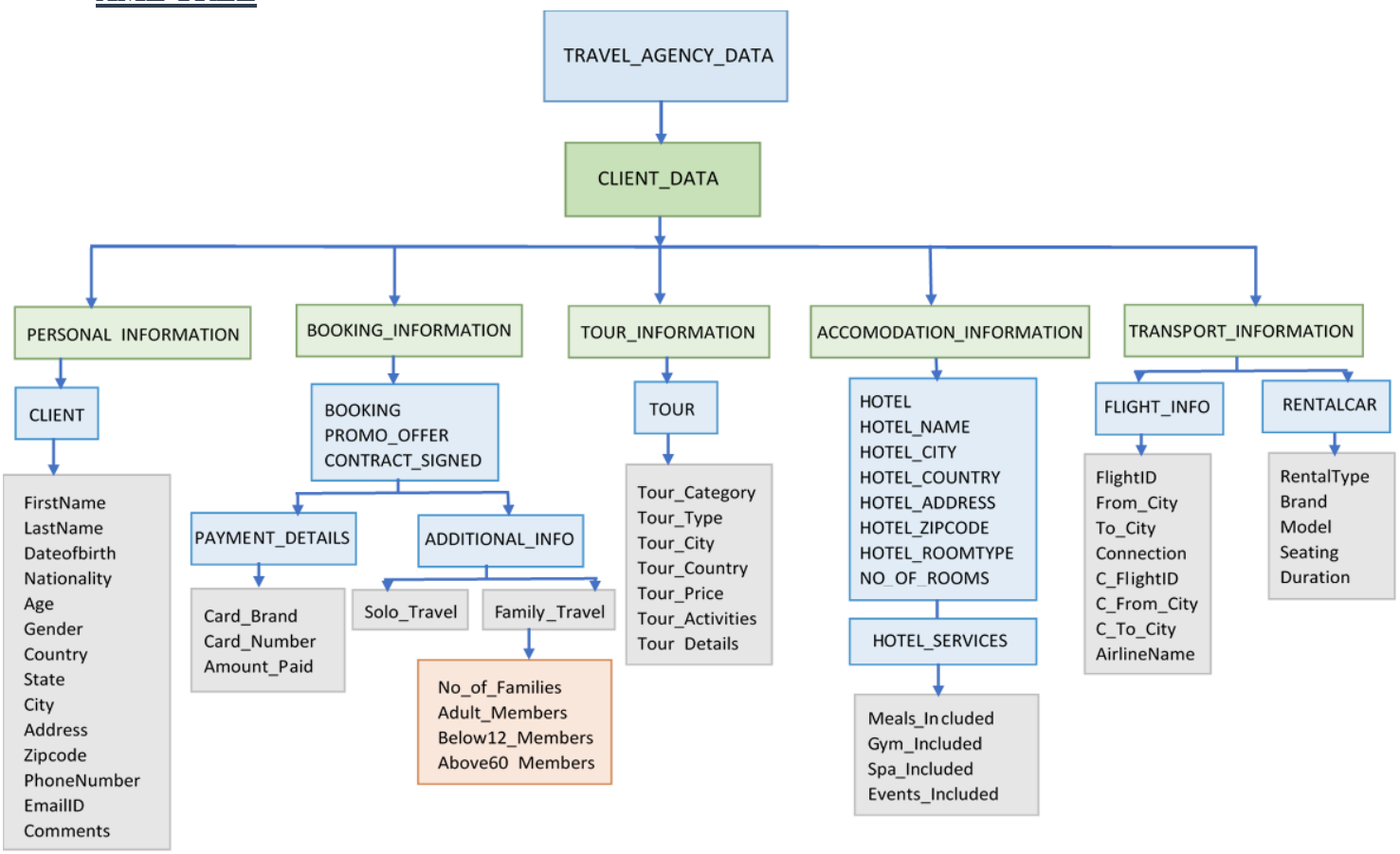

Figure 1:Graphical Representation of the XML Tree used in the modeling of the XML database.

XML FILE

The selected model, i.e., structure for the XML database of a travel agency is shown in Figure 1 above. As illustrated in the figure, the model has a tree structure with Travel Agency Data (TAD) as the root node with Client Data as its child node. The child node Client Data is a list of all the clients containing a sequence of all elements of Client Data. Even though the number of clients contained in the Client Data node can be unbounded, in my modeled database I worked with twelve clients made at random selection. Hence, the root node Travel Agency Data has the number of child nodes equivalent to the total number of clients in the database.

For each Client Data node, there are five other child node elements namely i) Personal Information ii) Booking Information iii) Tour Information iv) Accommodation Information and v) Transport Information, each group containing information about client personal, booking, tour, accommodation and transport details respectively.

i)Personal Information - The element 'Personal_Information' comprises of the general information about the client. In addition, to all the necessary details required for identification purposes like name, date of birth, gender etc., it also includes information about the client's ID number and nationality which could help decide their travel preferences and requirements.

ii)Booking Information - The element 'Booking_Information' is for the travel agency's internal use purpose as it contains information about any promotional offer given to the client, their payment details and the amount paid. Further, it also contains information about the number of people traveling pertaining to each client and if they are traveling with family or solo. The payment details and additional info node, each contain their respective sub-child nodes.

iii)Tour Information - The 'Tour_Information' element holds information about the category, type, city, country, price, activities, and duration of each client's travel itinerary. This helps to categorize if a client is traveling to an international or a domestic destination and aids in knowing about the type of tour package that they have chosen like EcoTourism, BeachTourism etc.

iv)Accommodation Information - The 'Accomodation_Information' consists of the details of the hotel like hotel name, address, city, country, room type, no. of rooms and the services included.

The services included child elements further contains four sub-child nodes related to meals, gym, spa and the events that are include for each client.

v)Transport Information – The 'Transport_Information' consists of important information corresponding to the flight information and rental car of each client, each further containing sub-child nodes. The flight information node contains details about the flight ID, airline name, and if or not there is a connecting flight etc. and the rental car node contains details about the type, brand, model, duration etc.

ADVANTAGES AND DISADVANTAGES OF THE CHOSEN MODELING APPROACH

The advantages and disadvantages of the specific modeling approach explained above are stated as follows:

1) Advantages:

i) Well defined model and Ease of Implementation: All the information in the travel agency database is stored in an organized manner inside the 'Client_Data' element. Hence all the information in the database can be accessed inside the 'Client_Data' element and this element provides an easy overview on the clients. This specific style of grouping makes it easier to implement the model, make changes to the model and to fetch information from the database.

ii) Clear and simple XQUERY for XSL: The XML Tree illustrated in Figure 1 makes it easier to comprehend and write queries for the XSL stylesheet and it makes retrieving information from the database simple.

iii) Easy Maintenance: Maintenance of the XML database is made easier with the tree structure as it makes parsing through the tree straightforward and flexible.

2) Disadvantages:

i) Complex XPATH: Sometimes to retrieve specific data linked with each other, a long and complex chain of query needs to be written to get information from the tree structure model.

ii) Alternate approach: Instead of storing all the data under the 'Client_Data' element, perhaps it could be restructured to include all the data under Booking or Tour Information for easier retrieval.

XML Schema

The schema for the travel agency data model is written as a reference to the XML file. The Schema helps to define the different data types present on the XML file and their corresponding occurrences namely minimum or maximum within an element. The XSD schema once written needs to be validated against its related XML file.

XML vs. JSON MODELING DIFFERENTIATION

As an additional attempt the XML database modeled was also written in JSON with its corresponding JSON Schema. This approached attempt indicated that modeling the database in JSON appeared comparatively simpler than its XML counterpart modeling.

III. XSL Transformations and their utility

There were six possible scenarios imagined for making six transformations from the Travel Agency Data model discussed above.

TRANSFORMATION 1

This scenario was written to fetch information which would help to get a general overview of the clients in the database.

Base Client Information

Client Name	Client ID	Nationality	City	Gender	Age
Brian Whittaker	21500	British	Lyon	Male	48
Annika Terrel	215001	French	Caluire-et-Cuire	Female	36
Jason Smith	215002	British	Aix-les-Bains	Male	55
Julie Copeland	215003	Belgian	Lille	Female	29

Figure 2: Transformation 1 output sample.

TRANSFORMATION 2

This scenario was written to retrieve information about the clients with specific booking specifications like if they received a promotional offer, credit card details and amount paid.

Additionally, a query was written to fetch information only for those clients who are traveling with their families.

There are 9 Clients traveling with Family in the Database

Client Name	Offer Recieved	Credit Card Brand	Amount Paid	No. of Families	No. of Children
Brian Whittaker	true	*Unionpay*	131.651	2	3
Annika Terrel	false	*Mastercard*	243.996	5	0
Jason Smith	true	*Voyager*	150.325	1	0
Sally Wright	false	*Visa*	192.141	3	2

Figure 3: Transformation 2 output sample.

TRANSFORMATION 3

The tour category and tour type pertaining to each client is seen in the output of the query written for this scenario which would help differentiate the clients as per their tour packages. Information like the co-related tour ID, city, country, cost and duration can also be seen illustrated here for easy referencing.

Information about Clients and their corresponding Tour Package

Client ID	Client Name	Tour ID	Tour Category	Tourism Type	Tour City	Tour Country	Tour Cost	Tour Duration
21500	**Brian Whittaker**	2149	Domestic	BeachTourism	Marseille	France	335.750	*5 days & 6 nights*
215001	**Annika Terrel**	6304	*International*	CulturalTourism	Bogota	Colombia	587.900	*7 days & 6 nights*
215002	**Jason Smith**	3278	*International*	EcoTourism	Tromso	Norway	405.880	*4 days & 5 nights*
215003	**Julie Copeland**	2641	Domestic	AdventureTourism	Grenoble	France	267.840	*3 days & 2 nights*

Figure 4: Transformation 3 output sample.

TRANSFORMATION 4

This transformation on the database was made to create an XML sub-database to provide an overview of the client name, age and their corresponding tour type, tour city, hotel name, room type and airline name. This helps to get a broader picture of the tour preferences of the respective clients.

SUB-TRANSFORMATION 4

The table below was created from the XML sub-database created in transformation 4 as part of the fourth scenario.

Client Tour Overview

Client Age	Tour Type	Tour City	Hotel Name	Room Type	Airlines Name
48	**BeachTourism**	Marseille	*Seascape Hotel*	Deluxe	RyanAir
36	**CulturalTourism**	Bogota	*Atlantis Pyramid Resort*	Double Room	AmericanAirlines
55	**EcoTourism**	Tromso	*Majestic Mantle Hotel*	Suite	Lufthansa Airlines
29	**AdventureTourism**	Grenoble	*Pristine Hotel*	Single	EasyJet

Figure 4: Sub-transformation 4 output sample.

TRANSFORMATION 5

This scenario illustrates part of the data such as the client ID, name, nationality, city of residence and tour data extracted from the main data XML database in JSON as the output of the transformation.

IV. Conclusion

This report presents comprehensive overview of the Data Pipeline part 1 project by following the assignment description and instructions in modeling an XML database for a tour operator.

TOOLS AND WEB SERVICES USED & FILES INCLUDED:

Following are the tools and web resources used for the project:

i) Notepad + + for creating the XML database, XSD, XSL and JSON files.
ii) https://jsonformatter.org/ for validating the JSON Schema.

The files created for this project are as follows:

i) XML Source File & Schema: TravelAgencyData.xml & TravelAgencyData.xsd
ii) XSL Files: TAD_Transfo1.xsl, TAD_Transfo2.xsl, TAD_Transfo3.xsl, TAD_Transfo4.xsl, TAD_Output4_Transfo.xsl, TAD_Transfo5.xsl
iii) HTML Output Files: TAD_Output1.html, TAD_Output2.html, TAD_Output3.html, TAD_Output4_Output.html
iv) XML Output File: TAD_Output4.xml ; v) JSON Output File: TAD_Output4.json
vi) JSON Files: TravelAgencyData.json & TravelAgencyData_JSON_Schema.xsd