DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Western Engineering

ECE 9014A - DATA MANAGEMENT AND APPLICATIONS

FINAL REPORT

Submitted By:

Dharanikota Rajendra Kamal
Harseerat Sohal
Shivani Solanki
Madhava Babu Ella
Srujana Bushireddy

MOTIVE:

To study and implement the various concept of data management and applications that are taught in class to create and store the database based on the railway system containing passenger's trains and their bookings. We need to assume and list out various components that are taken to create and store a database of a railway system and also explain the methods chosen to implement these techniques.

OBJECTIVE:

The main objective of this project is divided into three different parts, with each part emphasizing on different aspects of Data management and applications dealing with different aspects like how to deal with, handle, create , relate, normalize, store , extract, transform, and utilize the data which is available In different forms and these various tasks are segregated into three most important parts of data management and they are as follows.

1) Relational Data Base Management (RDBM tasks)

2) Data Warehousing and

3) Data Mining.

RELATIONAL DATABASE MANAGEMENT:

The first part is Relational Data Base Management which involves the ability to visualize the data, interpret it and be able to put together an entity relation diagram.

ERD (Entity relation diagram) and Relational mapping:

In an entity relation diagram the various relations between two entities are defined in a sequential and easy to understand manner. An entity is nothing but a real-life part or measure which needs to be digitalized to be stored in a database. There are two different kinds of entities and they are called strong and week entities. These entities are connected by relations which describes how two or more entities are connected or are dependent on each other. Hereafter we put up a diagram, we need to map various entities and attributes depending upon the cardinality ratios between two entities and the attributes of each entity. Then we need to check if this mapping of models is in normal form or not and if they are not in the normal form they need to be normalized based on the guidelines or constraints set by first, second and third normal forms. The relational table of each entity has a unique attribute which is used to identify that particular relation, and this is attribute is called as the primary key and it is occasionally used in order to identify or connect other relations as a foreign key (The primary key in relation is used as a foreign key in another relation thus connecting the two relations)

The first part is for us to implement and create a database for railway system containing its all its key components and in order to do so we need to make some assumptions other than the attributes and entities are given beforehand. The following are the attributes that are given and also that are assumed by us.

ASSUMPTIONS and GIVEN (ATTRIBUTES AND ENTITIES):

ENTITES	ATTRIBUTES
TRAIN	TRAIN_ID, TRAIN_NAME
STATION	STATION_ID, STATION_NAME
TRAIN_SCHEDULE	TRAIN_ID, STATION_ID, IN_TIME, OUT_TIME, SEQUENCE
PASSENGER_BOOKING	BOOKING_ID, TRAIN_ID, PASSANGER_NAME, BOOKING_DATE, FROM_STATION, TO_STATION, SEAT, COACH

TRAIN: This entity is used to store different aspects related to the train

STATION: This entity is used to store different aspects that are related to a station

TRAIN_SCHEDULE: This entity is related to a train or n number of them but is used to store the timings of arrival and depart of a train (in this training schedule we have TRAIN_ID and STATION_ID as the foreign keys which are primary keys in TRAIN and STATION relations respectively).

PASSENGER_BOOKING: This entity is used to show the booking of a particular passenger and the other aspects that relate him to the train

TRAIN_ID: The train id attribute which is a part of train entity and it is a unique identifier for the trains and since it is used to uniquely identify the trains it is made the primary key. (It is also used as a foreign key in the BOOKING_ID in order to get the details of the train that a passenger booked.)

TRAIN_NAME: The train name attribute stores the names of the trains by linking them with their respective id's and is also a part of the training entity

STATION_ID: The station id is also a primary which is used to uniquely identify the different stations along a train route and is a part of the station entity.

STATION_NAME: The station name is also a part of the station entity and is used to link and store the names of each station along the train route to the respective station id's.

IN_TIME and OUT_TIME: The in time and out time are used to determine the time at which the train arrives and departs from a particular station depending on the station id.

SEQUENCE: The sequence is an attribute in the train schedule attribute which is used to store the information on the sequence of trains arriving and leaving a particular station.

BOOKING_ID: When passenger books a train then the details of the train booked are linked to his personal details with the help of a unique identifier called the booking id and since it is used to uniquely identify the booking details of a passenger it is called the primary key for the PASSENGER_BOOKING entity.

PASSENGER_NAME: The passenger name is used to store the name of the passenger who is booking the seat in a train.

BOOKING_DATE: This is used to store the date of the booking.

FROM_STATION and TO_STATION: These attributes are used to store the starting and ending points of the booking being made.

SEAT and COACH: The seat and the coach are the attributes that are used to store the information of the seat and the coach that has been selected by the passenger (passengers choice) during the time of the booking.

We then need to draw an entity relation diagram which is then used to do the mapping based on the functional dependencies and then reduced into their normal forms. The following is the ERD for the above-given attributes and entities.

ENTITY RELATION DIAGRAM:

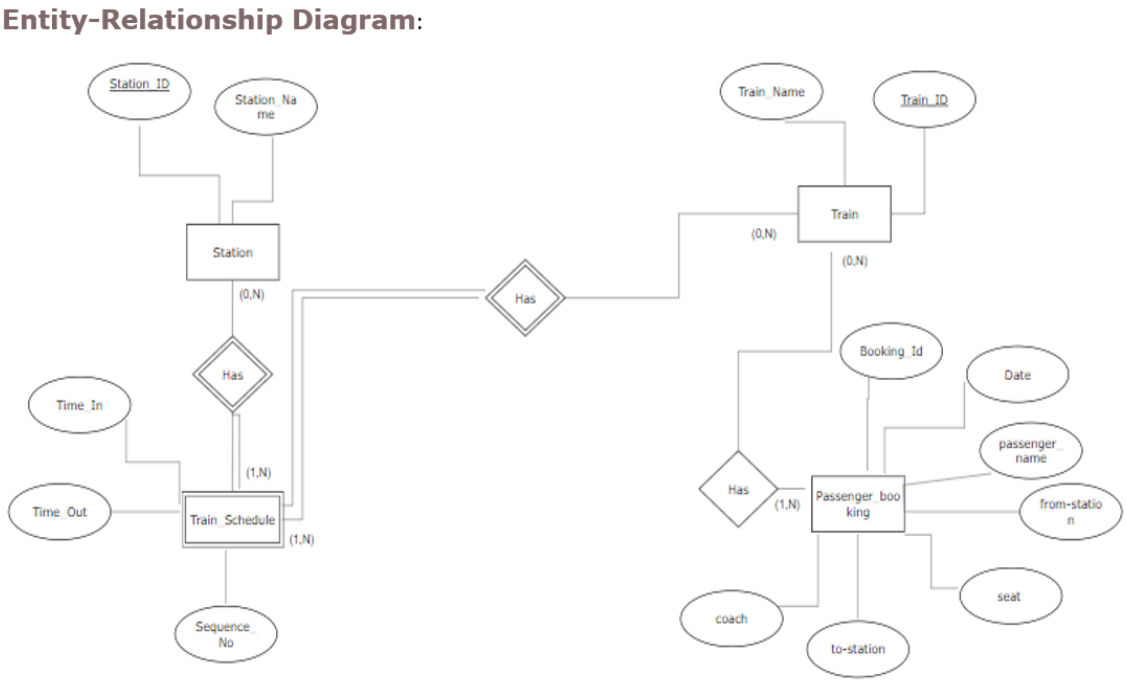

Fig (1): Entity-Relationship Diagram

Then based on this diagram and the cardinality ratios the mapping is done and then normalized based on the guidelines of first, second and third normal forms. The following is the relational diagram in its normal form (after normalization).

NORMALIZED FORM OF THE RELATIONAL DIAGRAM:

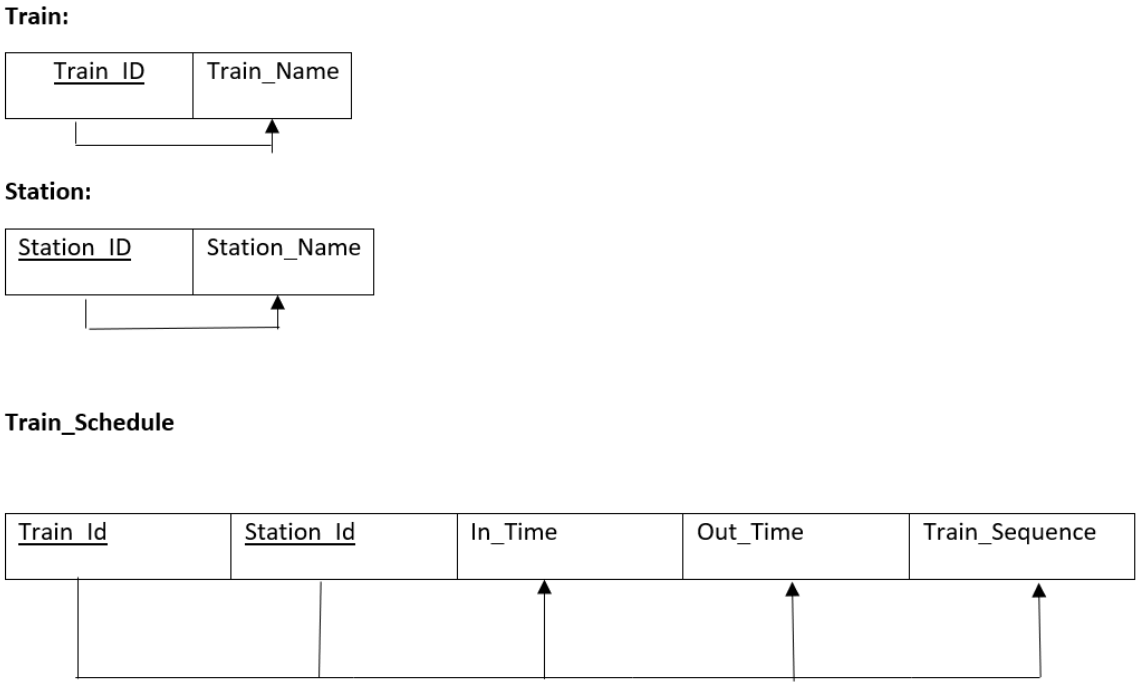

Fig (2): Relationship Diagrams of Entities

This is later passed to the second step of creating and storing a database that is DATA WAREHOUSING. For the data warehousing part, we have considered the addition of two new attributes for a better and clear understanding and at the same time were found to be important in a database that is dedicated to dealing with the railway system and its bookings done by the passenger. The newly considered attributes are as follows.

COST(DOLLARS): This new attribute is used to store the amount that is made by a passenger towards booking and the currency is maintained in dollars.

NO_OF_PASSENGERS: this attribute is used to store the total number of passengers that have booked for one particular train this is used to maintain the record in order to show if the sets are available or not.

DATA WAREHOUSING:

A data warehouse is designed to be used for an informational purpose, analysis and queries to support management decision making processes. In task 2, we designed and developed OLAP (On-Line Analytical Processing) tool. This tool gives multi-dimensional views of the data and lets End Users (railway officials) analyze and study trends on historical data over a specific period of time and predict data which will help in making management decisions.

Like the database, a data warehouse also maintains various schemas. There are three types of schemas, namely, Star schema, snowflake schema, and constellation schema out of which

we chose to use the star schema. The reason behind choosing star schema is that it optimizes ease of use, ease in databases to process and improves data retrieval performance by reducing the number of tables to be joined to materialize a transaction and hence, making the ETL jobs easier. The queries are written with simple joins between a smaller number of dimension tables and the facts where only attributes are filtered, speeding up the aggregations, which reduces the complexity of the schema compared others and hence contributing to improved performance.

STAR SCHEMA:

It is a data mart modeling consist of fact table referencing dimension tables. In the model below, booking_fact is a fact table, it is the information being queried. The tables dimbooking, dimstation and dimtrain are the dimension tables which are carrying the attributes of the railway system.

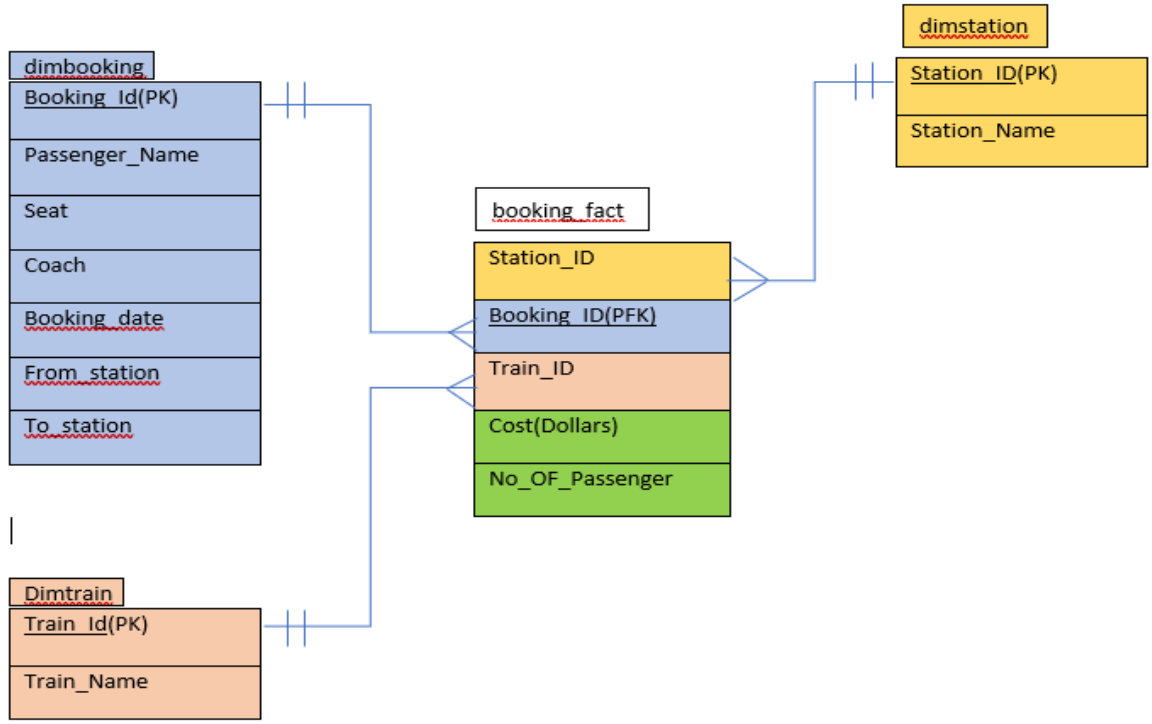

Fig (3): Star Schema

Fact Table:

The fact tables are the data structures which contain factual or quantitative data which is being queried for the particular business process. In the figure below, the query is made to analyze the number of passengers and the cost of the tickets. It consists of primary keys of all the dimension tables as the foreign key and has many to one relationship with the dimension tables. It is generally in denormalized form.

The fact table booking_fact from the model has the following attributes with respective descriptions.

Attributes	**Description**
Station_Id	dimstation table (Dimension table)
Booking_Id	dimbooking table (Dimension table)
Train_Id	dimtrain table (Dimension table)
Cost(Dollors)	Gives the cost of the ticket **(measures)**
Number of passengers	Gives total number of passengers travelled **(measures)**

DIMENSION TABLES:

It holds the descriptive information and reflects the attributes or the dimensions of the business processes. It is in denormalized form. It has a primary key column that uniquely identifies the record/tuple. The dimension table is associated with a fact table using this key. Three-dimension tables have been used in this model as described in the tables below:

1) Dimbooking: reflects the attributes related to the booking details of the passengers.

Attributes	**Descriptions**
Booking_Id(Primary Key)	BookingId of the passenger
Passenger_Name	Full Name of the passenger
Seat	Seat Number
Coach	Coach Number
Booking_date	Booking date
From_station	Source Station of the passenger
To_station	Destination station of the passenger

2) Dimtrain: reflects the details regarding the train.

Attribute	**Description**
Train_Id(Primary Key)	Id number of the train
Train_Name	Name of the train

3) Dimstation: reflects the information regarding the stations.

Attribute	**Description**
Station_ID(Primary Key)	Id number of the station
Station_Name	Name of the station

SQL Script:

The schema of the fact table and the dimension tables are defined in SQL from the relational schema, adding relations with keeping note of the primary keys, foreign key and the measures. Afterward, the SQL script for OLAP cube was created for Extract, load and transfer process. By this, the historical data is populated from the relational database to get the data ready for analysis and reporting.

DATA MINING:

The final assignment of the course was based on data mining. Data mining is the process of extorting useful information from raw data and represent it in an understandable and useful way. The data mining process includes the below steps:

1) Extract, transform and load data into data warehouse.
2) Store and manage the data into multidimensional databases.
3) Provide data access to business analysts using application software
4) Present analyzed data in easily understandable forms, such as graphs

At high-level data mining techniques can be categorized into the below:

1) **Predictive technique:** It is a very important data mining technique as it is used to predict future data based on the available data. In this technique, the historical trends and patterns are recognized to predict future values. For example, customers credit history and purchases could be analyzed to predict the credit risk in the future.

2) **Descriptive technique:** Descriptive as the name implies, they "Describe", or summarize raw data and make it something that is interpretable by humans. They describe the past. Descriptive techniques are useful because they allow us to learn from past behaviors and understand how they might influence future outcomes.

Based on the type of information provided and the type of information sought the data mining process is divided into the following:

1) **Association Analysis:** It is about tracking similar patterns and correlation between the linked variables. For example, when a customer buys a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores like in Amazon.

2) **Classification:** In this technique, the attributes are combined into distinct categories to make draw further conclusions. For example, while evaluating data on customers' financial backgrounds and purchase histories, we might be able to classify them as "low," "medium," or "high" credit risks. We can use the classifications to learn more about the customers.

3) **Clustering:** It is like classification but in clustering, we group data based on their similarities. For example, grouping different social statistics of people based on their trend to shop in the store.

4) **Outlier Detection:** In this technique, the anomalies or outliers in the data are identified to analyze the data. For example, if the credit card is mostly used to make grocery purchases but one strange week it used to make huge payment for electronic devices, we can check credit card thefts.

In the assignment, we did classification analysis using a decision tree. A decision tree is a tree-like classification model which generates a tree and a set of rules representing different classes from a data set. The decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

In the assignment, we used Titanic dataset for classification analysis. We build a decision tree using 'party' library in R. The decision tree for the data set is as below:

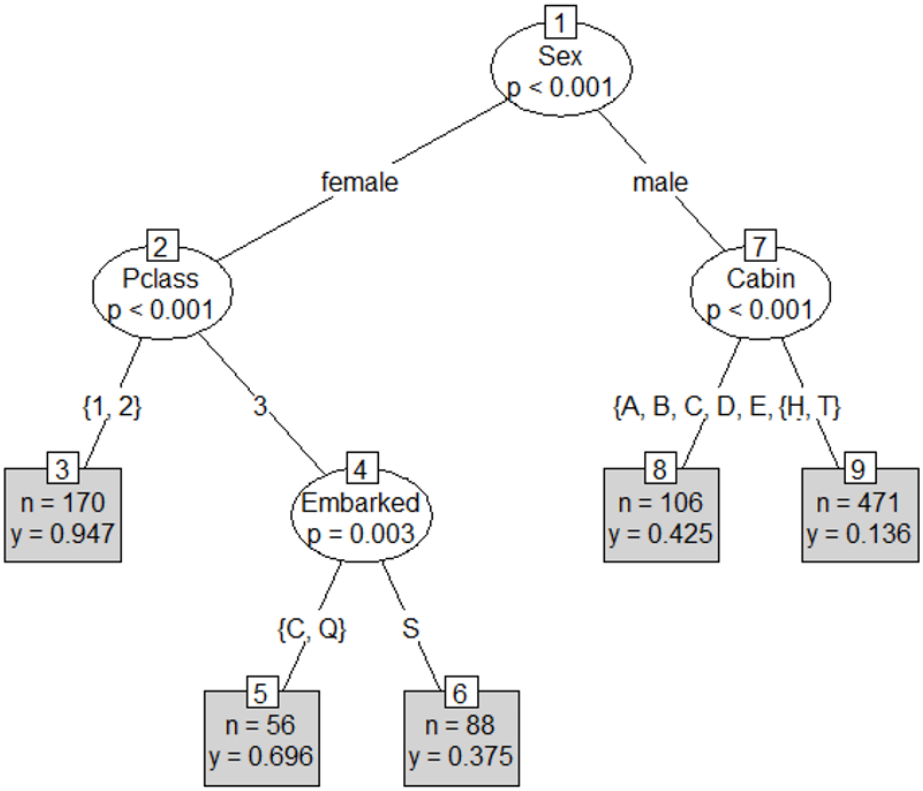

Fig (4): Decision tree of the Titanic dataset

CONCLUSION:

We learned about the concepts of the relational database, Data Warehousing, and Data Mining. We also learned how to create the database in SQL, to collect the data and transform it into entity diagram and to normalize data it and insert into the database using SQL queries. Created Data warehousing using ETL and thereafter did the data mining using decision tree for classification of Titanic Dataset.

REFERENCES:

Class PPTs by Ms. Shaima Ali

https://www.tutorialspoint.com/dwh/dwh_schemas.htm

https://www.teradatapoint.com/data-warehousing-schemas

https://www.folkstalk.com/2010/01/data-warehouse-dimensional-modelling.html

http://www.iasri.res.in/ebook/win_school_aa/notes/Decision_tree.pdf

https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

https://www.britannica.com/technology/data-mining

https://www.ibm.com/developerworks/library/ba-data-mining-techniques/

http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm