Performance Assessment

11/13/2023

Data Acquisition - D205

Student: Darian Gurrola

Course Instructor: Dr. William Sewell

1. **Question**

Businesses often use customer segmentation as a strategy to understand their customers and increase profitability (Raitaluoto, 2023). One example of this is family size segmentation, in which customers are divided into different groups based on the number of people in their household. The question we will ask is “What is the average amount of data used by fiber optic customer with 2 or more children?”.

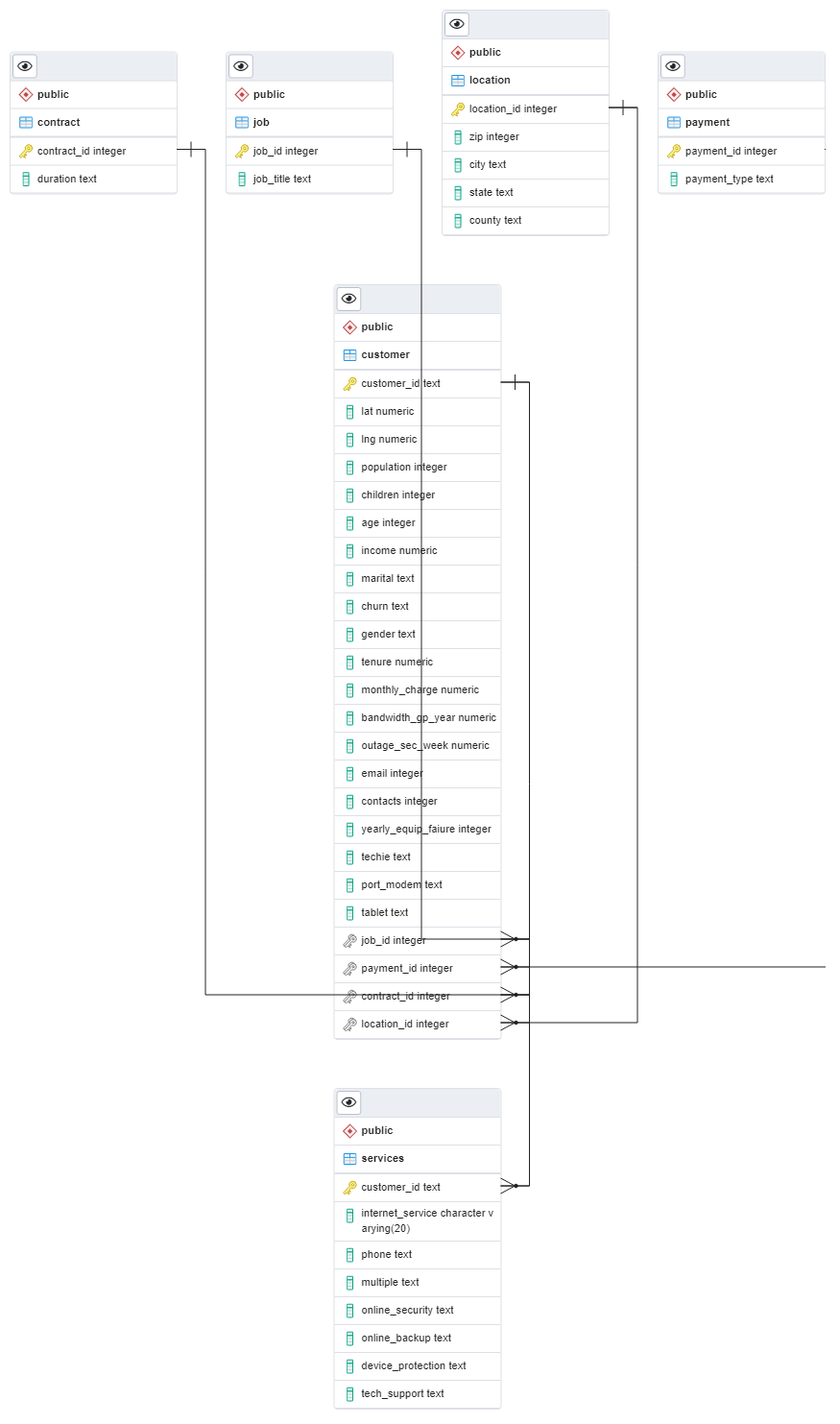
**A1. Identifying Data**

The first step in acquiring the necessary data is to join the customer and services table. A join query will need to be performed to access data across multiple tables (PostgreSQL, 2023). The customer and services table are joined on the customer\_id field. The data type for the customer\_id column in both tables is text.

The relevant columns in the churn table are ‘children’ and ‘bandwidth\_GB\_Year’. The ‘children’ column These columns provide the amount of children in a customer’s household. The ‘bandwidth\_gp\_year’ provides the average amount of data used by the customer in a year. The ‘children’ column uses an integer datatype, while ‘bandwidth\_gp\_year’ column uses a numeric datatype.

Lastly, the ‘internet\_service’ column from the services table is used to identify customers with fiber optic internet service. This column uses a varchar data type.

1. **Entity Relationship Diagram**



The entity relationship diagram above maps the relationships between the tables in the churn database. The services table has been created and added to the database. The customer\_id primary key in the services table references the same primary key in the customer table. There is a 1:1 relationship between the customer and services table. This is denoted by the constraint line between the customer\_id columns in both tables.

The contract, job, location, and payment tables are also connected to the customer table using their respective id columns. There is a constraint line between the customer and payment tables on the payment\_id columns, however, this was cut off when exporting the ERD image from pgAdmin.

**B1. Code for the ERD**

CREATE TABLE public.services (

customer\_id text PRIMARY KEY REFERENCES public.customer (customer\_id),

internet\_service varchar(20),

phone text,

multiple text,

online\_security text,

online\_backup text,

device\_protection text,

tech\_support text

);

The code sample above shows the services table being created. The customer\_id column acts as the primary key for the table. This column also serves as a foreign key by referencing the customer\_id column in the customer table. Referential integrity requires a foreign key to depend on a primary key in a parent table (IBM, 2022). All columns in the services table are assigned the appropriate type before add-on csv data is imported.

**B2. Loading CSV data**

"\\copy public.services (customer\_id, internet\_service, phone, multiple, online\_security, online\_backup, device\_protection, tech\_support) FROM 'C:/LabFiles/Services.csv' DELIMITER ',' CSV HEADER QUOTE '\"' ESCAPE '''';""

After creating the services table, the next step was to import add-on data from the services.csv file. The data was successfully loaded into the appropriate columns and can now be joined with other tables in the churn database. The code above was generated in pgAdmin software after the data was imported into the services table.

1. **SQL Query**

SELECT

      AVG(bandwidth\_gp\_year)

FROM customer

JOIN services

ON customer.customer\_id = services.customer\_id

WHERE children >= 2

AND internet\_service = 'Fiber Optic';

The query above was used to answer the question in section A. In the select statement, an aggregate function is used to find the average bandwidth used by customers in the past year. The data is obtained from the customer table which is joined with the services table on the customer\_id column in both tables. Lastly, filters are used to identify those customers with both fiber optic internet service, and 2 or more children in the household.

**C1. CSV Files**

The CSV file containing the result of the query has been submitted alongside this document. The average amount of data used in the past year by households with two or more children is 3293.77 gigabytes.

1. **Add-on Files**

The services table should be acquired and refreshed monthly in order to remain relevant to the business question.

**D1. Explanation of time period**

The database should be refreshed monthly to allow the business to adjust prices more frequently. Internet service providers typically adjust prices on every monthly billing cycle (Seitz, 2021). The information from the query can also be used to identify customers who may need to upgrade to a more suitable plan. Reviewing customer segmentation data every month will allow the business to optimize its prices.

1. **Panopto video of code**

**E1. Panopto video of programs**

Please use the URL below to view the Panopto video recording.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=83a51edc-7186-4557-8440-b0b900378aaa>

1. **Web Sources**

No web sources were used to obtain data or develop code.

1. **Sources**

Raitaluoto, T. (2023, May 11). *Segmenting customers by family size*. Markettailor. https://www.markettailor.io/blog/segmenting-customers-by-family-size

Referential Integrity (2022, October 10). *IBM*. Retrieved November 12, 2023, from[https://www.ibm.com/docs/en/informix-servers/14.10?topic=integrity-referential\](https://www.ibm.com/docs/en/informix-servers/14.10?topic=integrity-referential%5C)

Seitz, L. (2021, October 8). Hidden Costs: Why Internet Service Providers Regularly Raise Their Prices. *BroadbandSearch*. Retrieved November 12, 2023, from<https://www.broadbandsearch.net/blog/internet-providers-raising-prices>

PostgreSQL. (2023, November 9). *2.6. joins between tables*. PostgreSQL Documentation. https://www.postgresql.org/docs/current/tutorial-join.html