Natasha Estrada

000930497

D205

Data Acquisition

Western Governor’s University

07/14/2023

A. Research Question

1. Summarize a research question that can be answered using both the original database and the add-on CSV data.

How does the customer's tenure with the provider vary based on the type of internet service they have (DSL, fiber optic, None)? Do customers with DSL stay longer than customers who have fiber optic? How does that compare to customers who don’t have any internet service at all?

1. Identify which data from the original data set and the add-on CSV file are needed to answer the research question.

From the original data set (PostgreSQL tables):

*Customer table:* customer\_id, churn, tenure

From the add-on CSV file (customer\_internet.csv):

customer\_id, InternetService

B. Logical Data Model

Create a logical data model for the add-on CSV file by evaluating the data contained in the file and emphasizing the relational constraints.

A screenshot of a computer

Description automatically generated

Based on the provided files, the add-on CSV file includes columns such as customer\_id, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, and TechSupport. To create a logical data model for this CSV file, we can define the following table:

Table: internet\_service

customer\_id (Primary Key, Integer): Represents the unique identifier for each customer.

internet\_service (Text): Indicates the type of internet service (DSL, fiber optic, None) for each customer.

The relational constraint in this case is that the "customer\_id" column in the "internet\_service" table is linked to the "customer\_id" column in the existing "customer" table.

1. Write SQL code that creates a table that accommodates the extension of the logical data model to a physical data model by specifying the field types and relevant keys.

DROP TABLE IF EXISTS internet\_service

CREATE TABLE internet\_service (

internet\_service\_id SERIAL PRIMARY KEY,

customer\_id TEXT,

internet\_service TEXT,

FOREIGN KEY (customer\_id) REFERENCES customer (customer\_id)

);

1. Write SQL code that loads the data from the add-on CSV file into the table created in part B1.

COPY internet\_services (customer\_id,internet\_service)

FROM 'C:\LabFiles\Services.csv'

DELIMITER ',' CSV HEADER;

C. SQL Query

1. Write SQL statement(s) for a query or queries that inform the research question summarized in part A.

CREATE TABLE internet\_services\_churn AS

SELECT i.internet\_service, ROUND(AVG(c.tenure), 2) AS avg\_tenure, COUNT(\*) AS total\_customers,

COUNT(CASE WHEN c.churn = 'Yes' THEN 1 END) AS churned\_customers

FROM customer c

JOIN internet\_services i ON c.customer\_id = i.customer\_id

GROUP BY i.internet\_service;

SELECT \*

FROM internet\_services\_churn;

1. Provide a CSV file or files that capture the results from the query or queries.

Please find attached file internet\_services\_churn.csv submitted with this document.

D. Data Refresh Frequency

1. Determine how often the add-on file should be acquired and refreshed in the database for the data to remain relevant to the business and the research question.  
     
   The add-on file should be acquired and refreshed on a time schedule that matches how often the churn analysis query will be run. It is recommended that this be done weekly to ensure the company has access to timely information.

Ideally this process should be automated via a user created function that can be reused.

E. SQL Script for Loading Add-On Data

1. Create an SQL script that performs the process of loading the add-on data.

CREATE OR REPLACE FUNCTION load\_internet\_services()

RETURNS VOID

LANGUAGE plpgsql

AS $$

BEGIN

EXECUTE 'COPY internet\_services (customer\_id, internet\_service)

FROM ''c:\LabFiles\Services.csv''

DELIMITER '','' CSV HEADER';

END;

$$;

SELECT load\_internet\_services();

F. Panopto Video Recording

[Friday, July 14, 2023 at 9:35:18 PM (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7d78cc41-24ff-4a6f-9927-b040004bf074)

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7d78cc41-24ff-4a6f-9927-b040004bf074

G. Web Sources

1. Record the web sources used to acquire data or segments of third-party code to support the application. Be sure the web sources are reliable.

https://www.enterprisedb.com/blog/create-erd-pgadmin-4#:~:text=pgAdmin%20allows%20you%20to%20create,%E2%80%9CERD%20for%20table%E2%80%9D%20option.

H. Sources and Citations

1. No outside sources used. Code based on skills learned in DataCamp.