# Data Acquisition Performance Assessment

John Foster

Department of Information Technology, Western Governors University

D205: Data Acquisition

Professor William Sewell

March 12, 2023

# Data Acquisition Performance Assessment

## A: Research Question

The research question that I chose to explore in this analysis is as follows: What percentage of total customers in each state have subscribed to each service? I opted to design a query that would return each state as a row, and each service offered as a column. The numerical values in each column reflect the percentage of the total customer base in each state, both active and churned, that have ever subscribed to the services represented by that column.

## A1: Identifying Data

This analysis requires data from both the original data set and the add-on Services.csv file. From the original data set in the *churn*  database, the query will require the *customer\_id* and *location\_id* columns from the *customer* table, as well as the *location\_id* and *state* columns from the *location* table. From the *Services.csv* file, we require the data from all columns except the *multiple* column.

## B: Logical Data Model

The logical data model for this analysis was generated via the ERD tool included in pgAdmin 4. A screen snip of this logical data model can be observed below. In the original data source, the *contract, payment, job,* and *location* tables all have a relationship with the *customer* table via the *contract\_id, payment\_id, job\_id, and location\_id* columns respectively. These keys are all foreign keys of the customer table and primary keys of their respective tables, giving them a cardinality of many-to-one in this direction.

Incorporating the add-on data from the *services* table is unique in that its primary key, the *customer\_id* column, is also the primary key of the *customer* table. As this is the only unique key in the add-on data, it also functions as the foreign key ofther *services* table to create its relationship to the *customer* table. This functionally gives the relationship a one-to-one cardinality, although the ERD tool is not capable of visually representing this (it appears as a many-to-one relationship from the vantage of the *services* table). Based on the uncommon nature of this relationship, an analyst exploring this scenario in the real world should consider incorporating the content of the *services* table into the *customer* table itself. However, I instead elected to preserve the structure original data and import the add-on data as a distinct table for ease of resetting the data for development and presentation purposes, and because a detailed analysis of optimal data structuring falls outside of the scope of this project (Gagner, 2023).

A picture containing graphical user interface

Description automatically generated

## B1: Code for the Physical Data Model

CREATE TABLE services(

customer\_id text PRIMARY KEY NOT NULL,

internet text NOT NULL,

phone text NOT NULL,

multiple text NOT NULL,

online\_security text NOT NULL,

online\_backup text NOT NULL,

device\_protection text NOT NULL,

tech\_support text NOT NULL

);

ALTER TABLE services

ADD CONSTRAINT services\_fk

FOREIGN KEY (customer\_id)

    REFERENCES customer(customer\_id) DEFERRABLE INITIALLY DEFERRED;

## B2: Loading CSV Data

COPY services

FROM 'C:\LabFiles\Services.csv'

DELIMITER ','

CSV HEADER;

## C: SQL Query

SELECT

l.state,

(cast(COUNT(CASE WHEN internet != 'None' THEN 1 ELSE NULL END) AS float) / cast(COUNT(\*) AS float) \* 100) AS internet,

(cast(COUNT(CASE WHEN phone = 'Yes' THEN 1 ELSE NULL END) AS float) / cast(COUNT(\*) AS float) \* 100) AS phone,

(cast(COUNT(CASE WHEN online\_security = 'Yes' THEN 1 ELSE NULL END) AS float) / cast(COUNT(\*) AS float) \* 100) AS online\_security,

(cast(COUNT(CASE WHEN online\_backup = 'Yes' THEN 1 ELSE NULL END) AS float) / cast(COUNT(\*) AS float) \* 100) AS online\_backup,

(cast(COUNT(CASE WHEN device\_protection = 'Yes' THEN 1 ELSE NULL END) AS float) / cast(COUNT(\*) AS float) \* 100) AS device\_protection,

(cast(COUNT(CASE WHEN tech\_support = 'Yes' THEN 1 ELSE NULL END) AS float) / cast(COUNT(\*) AS float) \* 100) AS tech\_support

FROM services AS s

JOIN customer AS c USING(customer\_id)

JOIN location AS l USING(location\_id)

GROUP BY l.state;

## C1: CSV File(s)

The results of the query were exported from PGAdmin4 as a CSV file and submitted for evaluation in conjunction with this word document.

## D: Add-On File

According to the data in the *contract* table, contracts can be month-to-month, one year, or two years in duration. This means that a new or updated contract can be initially signed on any day of the week, month, or year, rendering the relevant service data incomplete after a single day. Therefore, the ideal periodicity for the acquisition and incorporation of current data from *Services.csv* to answer this question would be daily. This schedule ensures that it is always up to date and available for further filtering and analysis based on business needs.

## E: SQL Script

COPY services

FROM 'C:\LabFiles\Services.csv'

DELIMITER ','

CSV HEADER;

## F: Panopto Video

A brief demonstration video of the SQL code, including a description of the software environment used, has been recorded via Panopto and submitted for evaluation in conjunction with this word document in the WGU submission portal.

## G: Web Sources

No web sources were used in the completion of this project to acquire data or segments of code developed by third parties.

# H: Sources

Gagner, D. (2023, March 7). Webinar On D205 Performance Assessment. <https://wgu.webex.com/wbxmjs/joinservice/sites/wgu/meeting/download/45382b396ee45a1caf2381aa8359ec83>