Data Swan Inc: Medicaid and Medicare evaluation of sponsorship

and ratings

**Data Modeling**Team #11

Lucius Anderson

Chandrika Jones

Anita Pinto

Henry Rivera

Anthony Robnett

# Data Overview

The objective is to create a Data Warehouse to determine if there is a correlation between the payments made by CMS to physicians and the ratings generated by patients.

Data Sources and Information:

1. **Payments Data : CMS Payment Data ( Source: CMS Open Payment)**

These datasets contain consolidated information submitted by reporting entities for active years of Open Payments data

Link: https://openpaymentsdata.cms.gov/about/api

Years : 2020 and 2021

Types: General

Download: API

Size: Fetching only relevant columns the total size for 2020 and 2021 is 1.29 GB and 2.69 GB.

For the initial ETL pipeline, due to AWS’s Free Tier limitations and to avoid billing for large storage usage, the team has decided to begin the ETL implementation with a smaller subset of data ( 1000 - 5000 rows).

1. **Review Data : Government Ratings for Physicians ( Source CMS)**

Link: <https://data.cms.gov/provider-data/archived-data/doctors-clinicians>

Years : 2020 and 2021

Type : Annual Public Reporting Data

**Download: Direct csv download ( can use pandas read\_csv to directly download)**

**Size: 50 MB**

# Facts & Dimensions

**Payments Dataset:**

Dimensions:

1. Covered Hospital
   1. Hospital CCN (Pkey)
   2. Hospital Name
2. Covered Physician
   1. Physician NPI ( PKey)
   2. Physician Last Name
   3. Physician First Name
   4. Physician Middle Name
   5. Physician Suffix Name
   6. Physician Primary Type
   7. Physician Speciality
3. Address DImension
   1. Address Key ( PKey)
   2. Business\_Street\_Address\_Line1
   3. Business\_Street\_Address\_Line2
   4. City
   5. State
   6. zip
   7. Country

Facts:

Payment Fact

* 1. Payment ID ( PKey )
  2. Physician NPI
  3. Hospital CCN
  4. Address Key
  5. Payment Date
  6. Payment Quantity
  7. Prod Category

**Review Dataset:**

Dimensions:

1. Covered Physician
   1. Physician NPI (Pkey)
   2. Physician Last Name
   3. Physician First Name

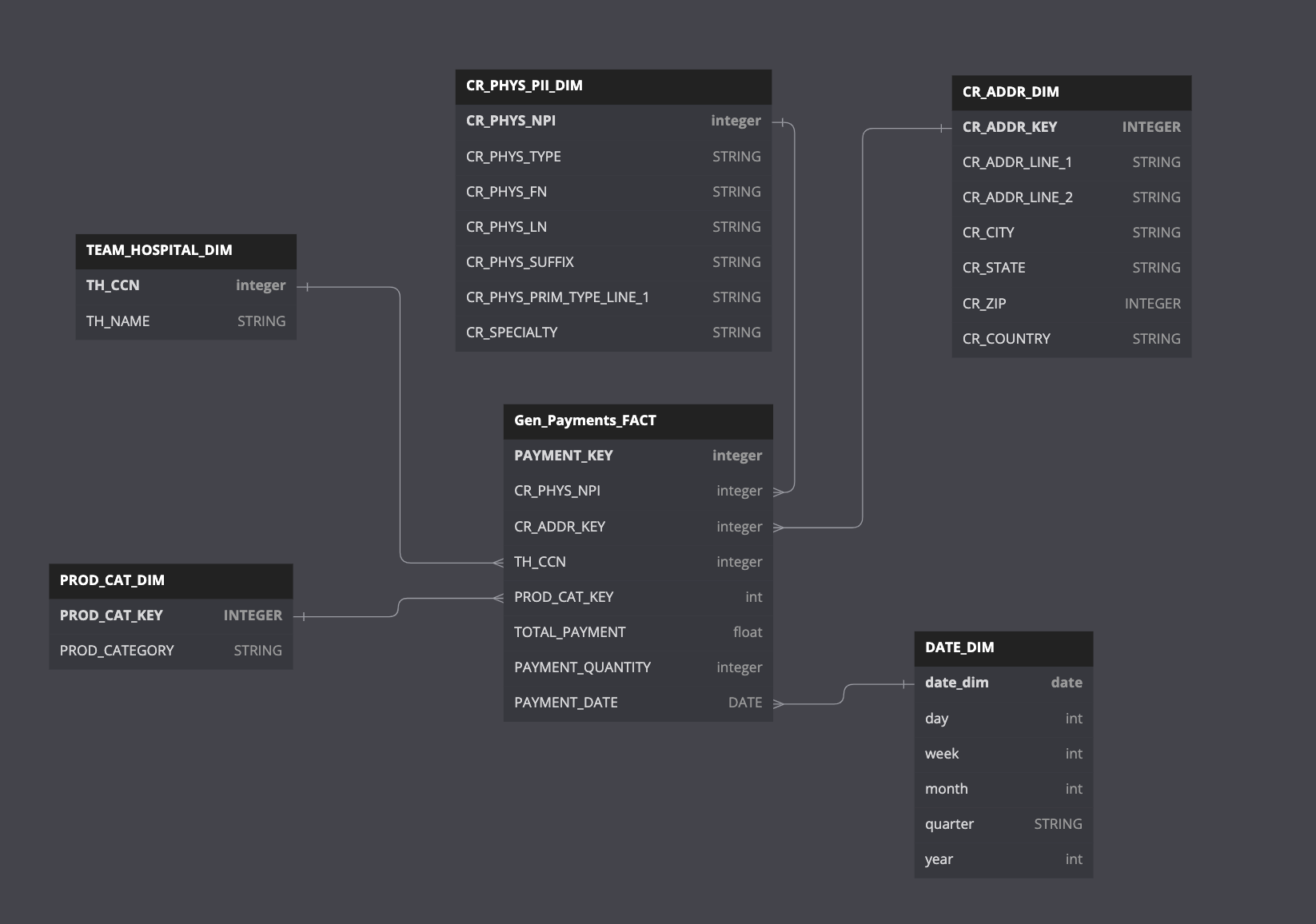
Facts:

1. Review Fact
   1. Review ID ( Pkey)
   2. NPI
   3. Possible Date Field( Month/Year)
   4. Quality\_Score
   5. IA\_Score
   6. Cost\_Category\_Score
   7. final\_MIPS\_score\_without\_CBD
   8. final\_MIPS\_score

# Schema

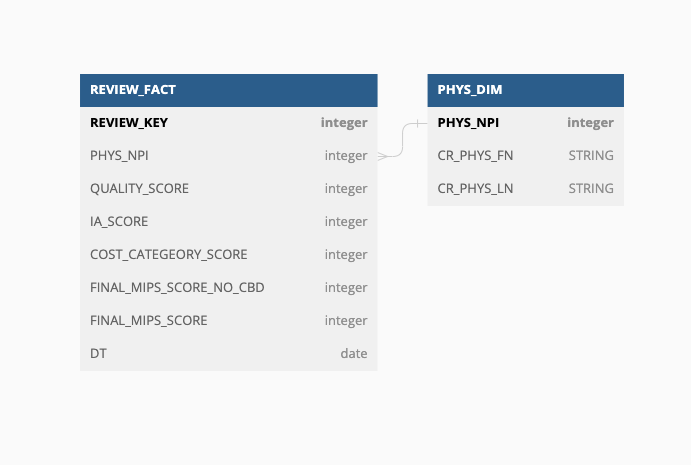
Created with: https://dbdiagram.io/home

Payment Schema:



Link: <https://dbdiagram.io/d/647a43d9722eb774944de470>

Review Schema:



Link - <https://dbdiagram.io/d/647bfb9f722eb77494585707>

# Justification

1. We currently have separate schemas for both Payment and Review, both payments and reviews are Facts. Star schema methodology limits 1 fact table per schema. One option we are considering is using a DIM\_NPI table which will connect both FACT tables.
2. The Current Data for Review provided to us does not have a Date field. However the source <https://data.cms.gov/provider-data/archived-data/doctors-clinicians> has annual public ratings from 2017. We can use available individual reporting data to recreate the date timeline.
3. The team also decided to continue only using the General Payment Dataset and drop the “Research” and “Ownership” Datasets. We believe the General Payment Dataset is a larger super set of all payments and includes payments already in Research and Ownership categories.