Data Swan Inc: Medicaid and Medicare evaluation of sponsorship

and ratings

Data Procurement and storage plan  
  
Team #11

Lucius Anderson

Chandrika Jones

Anita Pinto

Henry Rivera

Anthony Robnett

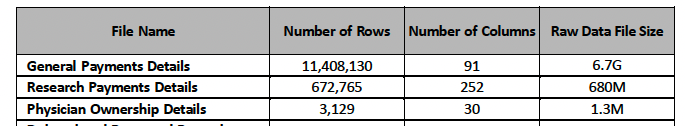
# Data Procurement

**Datasets (subject to change):**

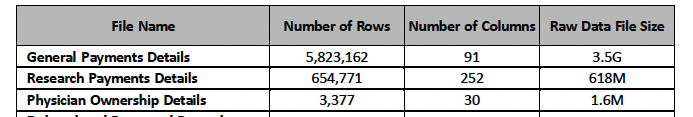
* [Open Payments](https://www.cms.gov/openpayments/resources)
  + Databases:
    - Research Payment
    - Ownership Payment
    - General Payment Data
    - Physician Profile Supplement
    - Hospital Owners
  + Extraction Methods:
    - ZIP Files (CSV) [direct download](https://www.cms.gov/OpenPayments/Data/Dataset-Downloads)
    - [API](https://openpaymentsdata.cms.gov/about/api)
      * API has endpoints for data download, returns JSON object
* Review Data:
  + Databases:
    - Healthgrades
    - ZocDoc
    - WebMD
    - RateMDs
    - RealSelf
    - Vitals
  + Alternative Databases:
    - Govt Ratings Dataset: This file has Merit-Based Incentive Payment System (MIPS) scores for Providers ( Source:CMS)
      * Pro: It has NPI ( unique ID) linking to the payment data providers.
      * Cons: Need to check if all providers have a corresponding Score.
    - [Yelp Review API](https://docs.developer.yelp.com/reference/v3_business_reviews)
    - [Google Reviews](https://developers.google.com/my-business/content/review-data)
    - [Reviewapi](https://reviewapi.com/#pricingSection) has a free tier, but *not sure if the approved sites are supported*
    - [RapidAPI](https://rapidapi.com/category/Medical) showcases some Medical Business APIs that may have review data available
    - [JustLikeAPI](https://www.justlikeapi.io/prices?utm_source=https://www.google.com/) offers a free 14 day trial that has review data, again, *not sure if the approved sites are supported*

**Payment Dataset Description:**

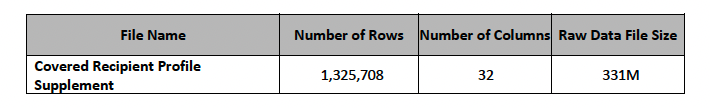
* **Requirements** : Payment Data for 2 years : 2020 and 2021
* **Tools** : API Downloads through Python Scripts
* **Source**: The CMS metadata can be accessed through Python API to obtain identifiers to download datasets for the years 2020 and 2021.
  + *Scalability*: To improve scalability, we can also add automation to expand download of any year’s dataset. For this project, we are focussing on downloading datasets related to 2020 and 2021
* **Dataset Size**:
  + 2021



* + 2020



* + Covered Recipient Profile Supplement File



**Extraction:**

* General Payment Data
  + Fields : There are 91 columns, Based on the dashboard requirement, only the following fields/columns would be extracted.
    - Covered\_Recipient\_Type
    - Teaching\_Hospital\_CCN
    - Teaching\_Hospital\_Name
    - Covered\_Recipient\_NPI
    - Covered\_Recipient\_First\_Name
    - Covered\_Recipient\_Middle\_Name
    - Covered\_Recipient\_Last\_Name
    - Covered\_Recipient\_Name\_Suffix
    - Recipeint\_Primary\_Business\_Street\_Address\_Line1
    - Recipeint\_Primary\_Business\_Street\_Address\_Line2
    - Recipeint\_City
    - Recipeint\_State
    - Recipeint\_Zip\_Code
    - Recipeint\_Country
    - Covered\_Recipient\_Primary\_Type\_1
    - Covered\_Recipient\_Speciality\_1
    - Product\_category\_or\_Therapeutic\_Area\_1
    - Total\_Amount\_of\_Payment\_US\_Dollars
    - Date\_of\_Payment
  + Rows : Extract a subset of 1000 rows which have non-null values in the specified fields
* Research Payment Data
  + Fields : There are 252 columns, Based on the dashboard requirement, only the following fields/columns would be extracted.
    - Covered\_Recipient\_Type
    - Teaching\_Hospital\_CCN
    - Teaching\_Hospital\_Name
    - Covered\_Recipient\_NPI
    - Covered\_Recipient\_First\_Name
    - Covered\_Recipient\_Middle\_Name
    - Covered\_Recipient\_Last\_Name
    - Covered\_Recipient\_Name\_Suffix
    - Recipeint\_Primary\_Business\_Street\_Address\_Line1
    - Recipeint\_Primary\_Business\_Street\_Address\_Line2
    - Recipeint\_City
    - Recipeint\_State
    - Recipeint\_Zip\_Code
    - Recipeint\_Country
    - Covered\_Recipient\_Primary\_Type\_1
    - Covered\_Recipient\_Speciality\_1
    - Product\_category\_or\_Therapeutic\_Area\_1
    - Total\_Amount\_of\_Payment\_US\_Dollars
    - Date\_of\_Payment
  + Rows : Extract a subset of 1000 rows which have non-null values in the specified fields
* Ownership Payment Data
  + Fields : There are 252 columns, Based on the dashboard requirement, only the following fields/columns would be extracted.
    - Physician\_NPI
    - Physician\_First\_Name
    - Physician\_Middle\_Name
    - Physician\_Last\_Name
    - Physician\_Name\_Suffix
    - Recipeint\_Primary\_Business\_Street\_Address\_Line1
    - Recipeint\_Primary\_Business\_Street\_Address\_Line2
    - Recipeint\_City
    - Recipeint\_State
    - Recipeint\_Zip\_Code
    - Recipeint\_Country
    - Total\_Amount\_of\_Payment\_US\_Dollars
    - Date\_of\_Payment
  + Rows : Extract a subset of 500-1000 rows which have non-null values in the specified fields
* Covered Recipient Profile Supplement File
  + Fields: Since this is a physician supplement data, we will use all available fields to create a physician lookup reference table
  + Rows: Extract those rows with “Covered\_Recipient\_NPI” values in the other 6 datasets.

**Intuition behind choosing the datasets:**

Dataset Types:

1. Though there might not be a direct correlation between Patient reviews and payments for Research / Ownership, it will be helpful to know how physician reviews affect payments in those categories also.
2. Recipients receiving payments for Research and Ownership might be practicing physicians and it is good to explore if patient reviews affect their payments in categories like Research and Ownership too.

Field Extraction: It is helpful to use detailed addresses of recipients like street addresses to validate and link the matching recipient info from review data due to common names, nick names and abbreviated names.

# Data storage plan

**Storage Options:**

[AWS Lambda](https://aws.amazon.com/lambda/) with AWS CodeCommit and AWS CodePipeline

* Strengths:
  + AWS Lambda functions can be stored and version controlled in AWS CodeCommit, allowing for collaboration and tracking of code changes.
  + AWS CodePipeline can be used to create a CI/CD pipeline for your Lambda functions, ensuring automated testing, deployment, and monitoring.
  + The serverless nature of AWS Lambda makes it easy to incorporate into CI/CD workflows, as you can deploy changes quickly without worrying about infrastructure provisioning.
* Weaknesses:
  + AWS Lambda functions can become complex to manage and version control, especially as the number of functions and their dependencies grow.
  + Coordinating multiple Lambda functions within a CI/CD pipeline might require additional effort to ensure proper integration and sequencing of the functions..
* What is stored?
  + *From the data procurement section above, we can store subsets of the raw datasets with relevant data to our objective*
  + *At a high-level we still store:*
    - *Open payments general payment and research payment data, only the relevant data*
    - *Review data of only the accepted review publishers*
* Outlines of storage
  + *S3 bucket layout can look similar to this:*
    - *Datasets*
      * */Open Payments*

*/Research Payment*

*/Year*

*/2020*

*{{phyical\_file or data object}}*

*/2021*

*{{phyical\_file or data object}}*

*/Ownership Payment*

*..*

*/General Payment Data*

*..*

*/Physician Profile Supplement*

*..*

*/Hospital Owners*

*..*

* + - * */Review Data*

*/Healthgrades*

*/Year*

*/2020*

*{{phyical\_file or data object}}*

*/2021*

*{{phyical\_file or data object}}*

*/ZocDoc*

*..*

*/WebMD*

*..*

*/RateMDs*

*..*

*/RealSelf*

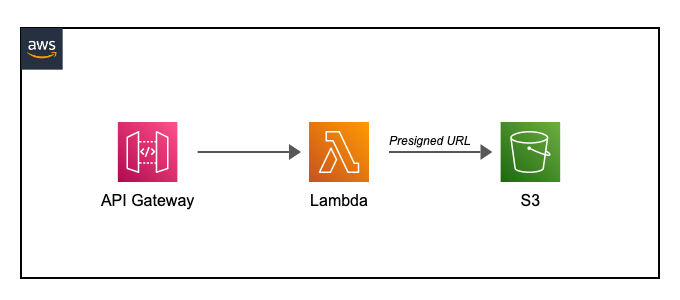
*..*

*/Vitals*

*..*

*CMD MIPS Data*

*..*

* How does storage take place?
  + 
* Uploading raw data to S3
  + *Will upload only the relevant data for completing our objective to S3 bucket, excluding data that is not needed. This is will help reduce volume of data.*

# Resource table

The table below organizes all data resources for easy access.

| **Resource name** | **Type** | **Description** | **Format** | **Size** | **Source link** | **Storage place** |
| --- | --- | --- | --- | --- | --- | --- |
| CMS Payment Data | JSON | These datasets contain consolidated information submitted by reporting entities for active years of Open Payments data. | CSV | 12 GB | https://www.cms.gov/OpenPayments/Data/Dataset-Downloads | AWS S3  ( Only relevant dats of 12GB will be uploaded to AWS) |
| CMS MIPS Review Data | csv | This Dataset contains performance information for Merit-Based Incentive Payment System (MIPS) Quality, Promoting Interoperability, and Improvement Activities performance information submitted by groups. | parquet | 52 MB | https://data.cms.gov/provider-data/search?theme=Doctors%20and%20clinicians | AWS S3 |