Sumaiya Gulshan
Polina Polivkina
Most Basari
Sanzida Antora
Professor Jefferson Bien-Aimé (CIS 4400)
July 11th, 2023

SPM Project Paper

Due to a rapid change in technology, terabytes of data are being produced every single second. To be precise, according to the latest estimates, 328.77 million terabytes of data are created each day so it's crucial to design, develop and implement data in a centralized repository that stores and organizes large volumes of data from various sources. The data warehouse does the same thing. In this project, we helped The Centers for Medicare and Medicaid Services to design a database that would show the users the profession and the popularity of all the doctors that are getting paid by the State.

To start this project, we used a non-relational database management system called MongoDB which handles large volumes of unstructured data and makes it suitable for handling diverse data types like documents, graphs, and time-series data. For completing the main task, we referred to OpenPaymentsData.CMS.gov to collect 2021 general, ownership, and research payment data in addition to physician profile supplement data.

After analysis of the data, we used the star schema method to create the dimensional model. In this step, we created three Dimension Tables, including Doctor Dimension, Company Dimension, and State Dimension. Each of these tables include the descriptive attributes or characteristics of the business entities. Also, we have added two Fact Tables, which include measurable numerical data that represent the performance or behavior of the chosen business entities. The two Fact Tables are the Ratings Table which contains CMS_Rating, RateMD_Rating, and the Grants Table, which has the number of grants and amount of grants. Each table has a key that connects that table to another table. After that, we built a logical structure and a DB Schema to provide a framework for organizing and storing data in the database system.

To get the data from the surveys and MongoDB and to filter the data to get specific data, we converted Json data into a data frame and cleaned the data. In the next step, we analyzed the location based on the state, as our goal for the project is to find the profession and the popularity of all the doctors that are getting paid by the State. And finally, we analyzed grants by company and specialties of the doctors with ratings to find out which specialty has the best rating and vice versa. We also came up with a chart that shows the map of all the states showing, the grants that they got, and the average rating for each state. The data can be filtered to display the doctors specialties, ratings, the state they work in, and also the company that sponsors them.