Architecture Design

Explaining the high-level design of the program, assumptions made, possible use of external libraries, how to run the program, requirements and related issues that need to be considered by the instructor

The design of this program is spread across the several scripts that load the data into the data lake, create tables, and transform the data for analysis. The final resulting tables are a flat dimensional model design with dimensions for hospital and measure information, and facts for readmissions, care, and survey results.

The following assumptions were made:

There are several null values and scores that are Not Available that must be accounted for. There might be several reasons that a score would not be available—the hospital might not have the resources for that measure, the survey taker could not get the measurements, or the number of cases is too small. In the case that the hospital does not have those resources, presenting the 0 in the calculations makes sense because we want to get hospitals with the highest scores and the most variety. Ignoring null values will result in hospitals with high scores and fewer measures ranking the same as hospitals with high scores and many measures. For this reason, in any case, when a score is not available, the number of cases is too small, or null it will be recorded as 0.0 in calculations.

Specifically, in the readmissionsDim table, the scores are assessed compared to the national rate. When the lower estimate is greater than the national rate, then the lower estimate is the value that is compared to the national rate. When the higher estimate is less than the national rate, then the higher estimate is the number used. Otherwise, it is the regular score. There is an assumption here that the comparison of the national rate is what determines whether a score is good or bad. For this reason, both straight score and comparison score are recorded. For the analysis, the comparison score and rank are used. Additional comments are outlined in each script.

To run the program, each script must be run in the order as described in the Exercise 1 project document, as follows: load_data_lake.sh, hive_base_ddl.sql, transforming.sql, best_hospitals.sql, best_states.sql, hospital_variability.sql, hospitals_and_patients.sql. All of the sql files were run through hive or spark-sql.