W205 – EXERCISE 1 – SRIRAM RAO

GitHub Repository

https://github.com/sriramrao1/exercise_1/

This is a public repository. So you should be able to make Pull requests.
I have provided 'DwMcclary' with access to contribute as well.

Data Source

In the load_data_lake.sh script, I am downloading the data from the CMS Hospital Compare website every time.

Files Used:

File Name	Description	Renamed file name
Hospital General Information.csv	General Hospital information	hospitals.csv
Timely and Effective Care - Hospital.csv	Procedure data	effective_care.csv
Readmissions and Deaths - Hospital.csv	Procedure data	readmissions.csv
Measure Dates.csv	Mapping of measures to codes	measures.csv
hvbp_hcahps_10_28_2015.csv	Example survey response data	surveys_responses.csv

Please note that one of the files listed in the exercise "**hvbp_hcahps_05_28_2015.csv" has been replaced with "hvbp_hcahps_10_28_2015.csv".** My scripts use the updated file.

Query Processing Engine Used

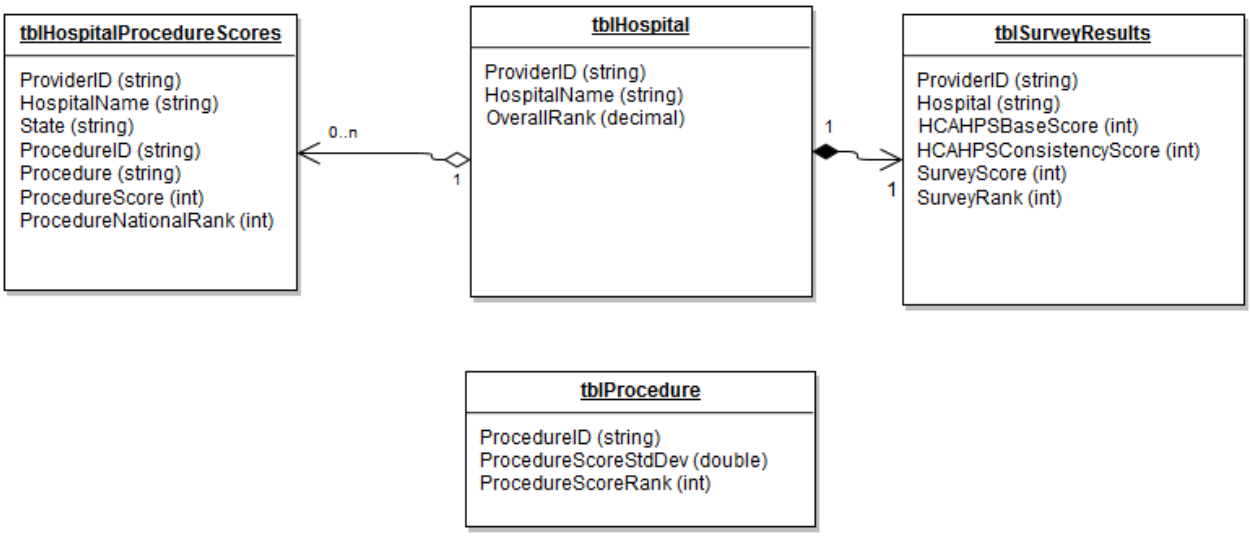
- I have used Hive as the processing engine, due to my familiarity with SQL.
- The queries take longer to execute than PySpark.

Instructions to run the Transformation Scripts

Run the files in this order:
1. tblHospitalProcedureScores.sql
2. tblProcedure.sql
3. tblSurveyResults.sql
4. tblHospital.sql

W205 – EXERCISE 1 – SRIRAM RAO

ERD for the Transformation Tables



Analytical Reasoning Logic

The Hospital Compare project provides several files with a number of measures to assess the quality of care at each of the hospitals assessed.

However for this exercise, I have chosen the "Effective Care" (**Timely and Effective Care - Hospital.csv**) set of measures and the "Readmission Death" (**Readmissions and Deaths - Hospital.csv**) set of measures to rank the hospitals.

However, the logic for scores provided for "Effective Care" measures is totally different from those for the "Readmission Death" measures. For example, an effective care measure is AMI_10 (**Statin at Discharge)** where the best practice is to prescribe statin during discharge of a patient with Heart Attack. So if Statin was prescribed to the patient at the time of discharge by the hospital, a high score is provided to the hospital. On the contrary, a high score for a "Readmission Death" measure such as Acute Myocardial Infarction (AMI) 30-Day Mortality Rate indicates that the hospital has a high mortality rate for patients admitted to the hospital for AMI.

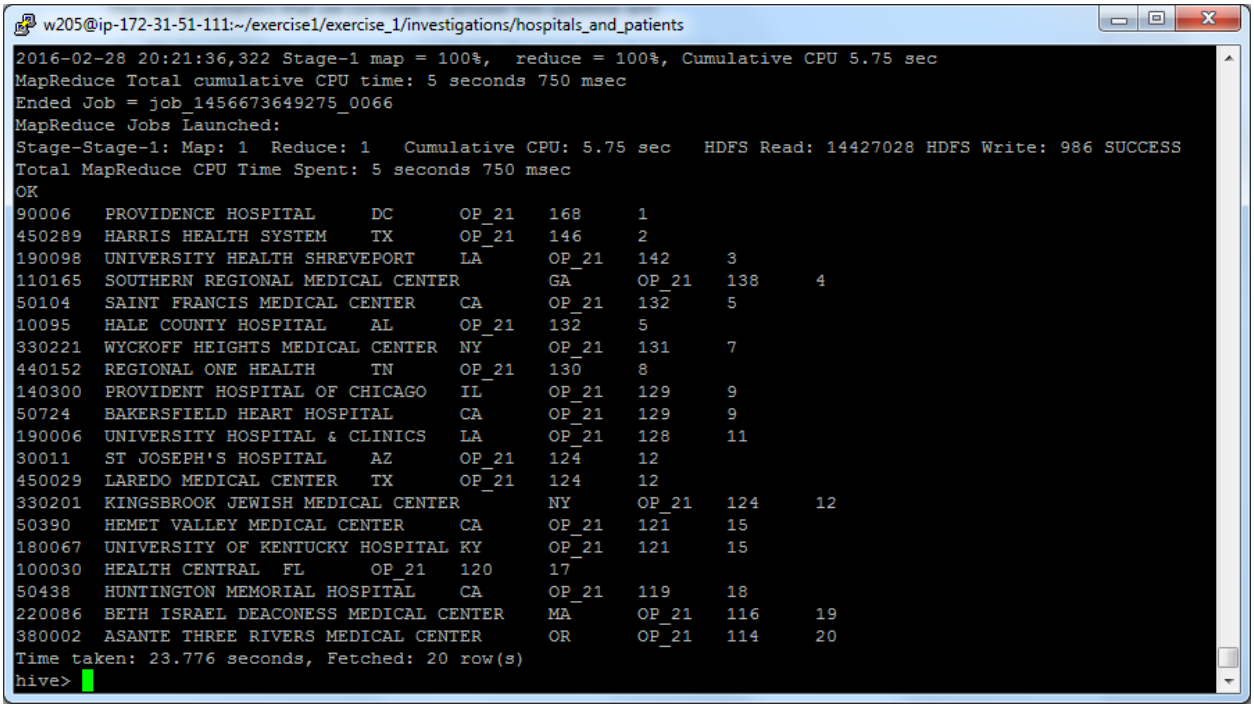
Due to this, we cannot directly sum up scores for the Effective Care and Readmission Death measures and use them for the best hospitals assessment.

Therefore I have attempted to rank the hospitals separately for "Effective Care" measures and "Readmission Death" measures based on high scores for "Effective Care" measures and low scores for "Readmission Death" measures.

Effective Care

Please see attached screenshot. The last two columns provide the raw score and rank of the hospital for the procedure. As you can see as the raw score increases, the hospital receives a better rank

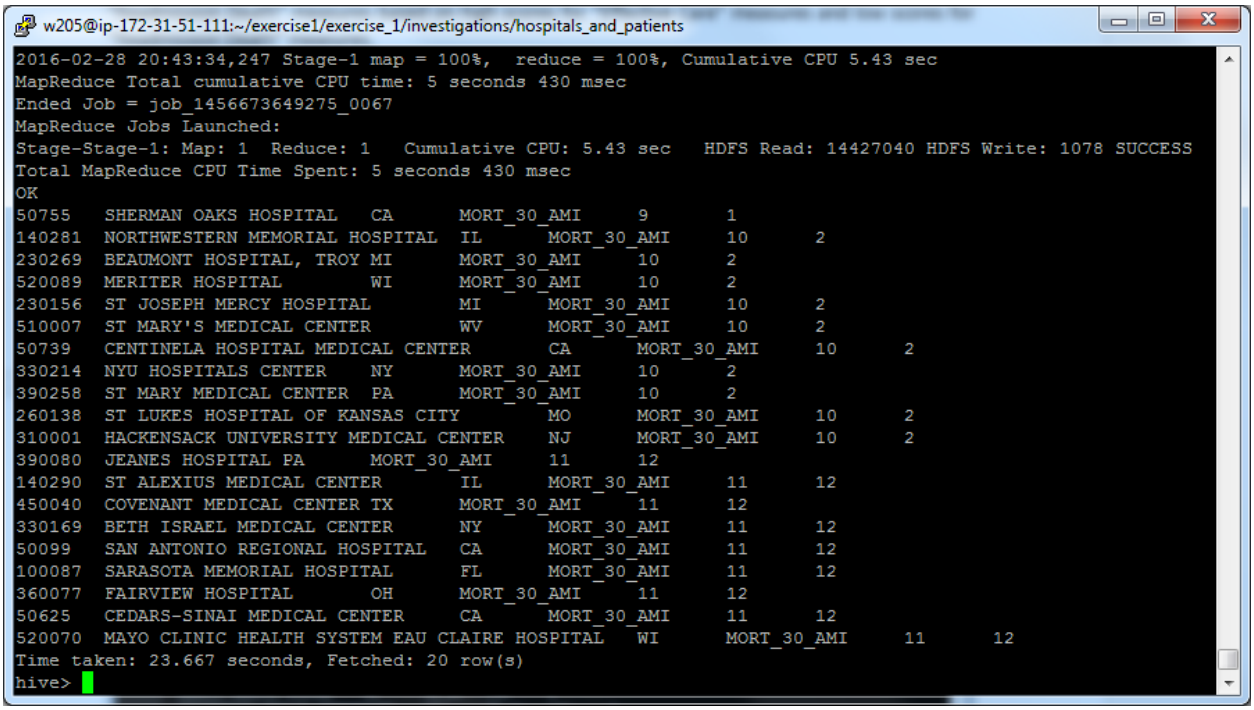
select providerid, hospitalname, state, procedureid, procedurescore, ProcedureNationalRank from tblHospitalProcedureScores where procedureid like '%OP_21%' order by ProcedureNationalRank limit 20;



Readmission Death

Please see attached screenshot. The last two columns provide the raw score and rank of the hospital for the procedure. As you can see as the raw score decreases, the hospital receives a better rank

select providerid, hospitalname, state, procedureid, procedurescore, ProcedureNationalRank from tblHospitalProcedureScores where procedureid like '%MORT_30_AMI%' order by ProcedureNationalRank limit 20;



Best Hospitals - Overall Ranking of the hospital

What hospitals are models of high-quality care? That is, which hospitals have the most consistently high scores for a variety of procedures.

To assess the best quality hospital, we obtain a mean of the ranks obtained by the hospital for each of the procedures it performs. The overall mean rank of the hospitals are stored in the tblHospital table.

There are several items to note with this ranking mechanism:

- All procedures are weighted equally. So a hospital that receives first rank in a complex procedure such as heart surgery, is weighted the same as another hospital that receives first rank for flu treatment
- With the overall mean rank being derived as an average over the ranks of all procedures performed, a specialty hospital that performs one procedure exceptionally well will receive a higher overall rank than a general hospital that provides several procedures. In our analysis, the Children's Hospitals and Surgical Specialty hospitals ranked high overall compared to other general hospitals.
- The ranking just accounts for the raw scores for the procedure and does not account for the complexity and number of procedures being performed. So a hospital that handles more complicated procedures or handles a large number of cases, does not deserve any special treatment.

Best States

What states are models of high-quality care?

I utilize the average of the rank of the hospitals in the State to derive the States that provide the best quality of care in the nation. That is a State with a group of high quality hospitals must provide a higher quality of care overall.

Hospital Variability

Which procedures have the greatest variability between hospitals?

Here I use the Standard Deviation of the raw scores of each measure or procedure to assess the variability in the scores for the same procedures performed at different hospitals.

I could not locate the exact range of the scores in the data dictionary. Hence I made an assumption that the raw scores provided across several procedures are over the same range of scores. The standard deviation values are then comparable across procedures.

Based on ordering the procedures with the top 10 standard deviation values, we get the 10 procedures with the most variability between hospitals.

Hospitals and Patients

Are average scores for hospital quality or procedural variability correlated with patient survey responses?

The two parameters that we correlate to answer this question are:

1. The hospital evaluation scores provided by the evaluators by procedure.

2. The survey result scores provided by the patients.

For the hospital evaluation ranking - we have already ranked the hospitals based on the procedure level scores. This is stored in the tblHospital table.

To derive the survey ranking, we use the HCAHPS Base Score, HCAHPS Consistency Scores. The combined HCAHPS Base Score, HCAHPS Consistency Score is used by CMS to determine the Patient Experience of Care domain portion of the Medicare reimbursement. So we use that as the model to assess the survey responses. We rank the hospitals based on the combined survey score. This is stored in the tblSurveyResults.sql table.

Please note that the survey results file (**hvbp_hcahps_10_28_2015.csv)** only provides data for 3041 hospitals as opposed to 4500+ hospital records in the "Timely and Effective Care - Hospital.csv" and "Readmissions and Deaths - Hospital.csv" files. Hence the correlation coefficient is being derived only for 3041 hospitals.

The correlation coefficient is 0.10767957813957273.

This low value of the correlation coefficient indicates low positive correlation between hospital quality scores and patient survey responses.