

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

An average of all scores across a cross section of different procedures have been conducted using SQL on hive. And below is a screen-shot of the top ten hospitals along with scores:

```
2017-02-27 04:41:46,216 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 2.87
sec
MapReduce Total cumulative CPU time: 2 seconds 870 msec
Ended Job = job_1488136239136_0038
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.11 sec HDFS Read: 334091
HDFS Write: 3960 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 2.87 sec HDFS Read: 8401 HD
FS Write: 415 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 980 msec
OK
h.hospitalname avg_score
HOLY CROSS HOSPITAL 0.6190599394542257
MARSHALL MEDICAL CENTER 0.6120756992560855
ST JOSEPHS HOSPITAL 0.5956062772814308
MONROE COUNTY HOSPITAL 0.5890794255400374
ST CLARE HOSPITAL 0.5881941138088279
TAYLOR REGIONAL HOSPITAL 0.5872436048256955
MERCY HOSPITAL 0.5843495015057304
SUMNER REGIONAL MEDICAL CENTER 0.5788979149702126
NEWTON MEDICAL CENTER 0.5772440395946228
STEPHENS MEMORIAL HOSPITAL 0.5763270907334845
Time taken: 57.248 seconds, Fetched: 10 row(s)
hive>
```

Clearly, HOLY CROSS HOSPITAL is #1 in terms of quality of care score, with MARSHALL MEDICAL being a very close second; note the scores differ only in the third place of decimal (=so, it may be inferred that the top two positions are practically tied). It is instructive to note that the top five hospitals differ among them in average scores by about 5.3% roughly, so competition for the top five slots is very steep

Finally, a word or two about the scoring methodology used. I excluded two items:

- NA's and non-integer values
- Hospitals which had less than 30 procedures

In order to normalize the scores, I used the descriptions in the data dictionary for the raw data (https://data.medicare.gov/views/bg9kemty/files/o846xclMrfQkBsWynXrdeftJouyrOYf_EUT8yLi7T4c?filename=Hospital.pdf&content_type=application%2Fpdf%3B%20charset%3Dbinary), as discussed in Kevin Crook video #1 for Exercise 1.

The approach I adopted (please also see the References section in the ReadMe file) is normalizing by the maximum of each procedure across all the eligible hospitals. Then I proceeded to further adjust the scores so obtained for which a higher score is worse by subtracting them from unity (for example, a number of procedures 'ED_1b', 'ED_2b', 'EDV', etc in the timely and effective care data gets affected)