**Part 1: The Star Schema**

**What are some practical advantages to using a star schema data model over a relational model? A few are mentioned in the exercise, but challenge yourself to identify others**.

It is extremely simple to understand and build.

No need for complex joins when querying data.

Accessing data is faster (because the engine doesn't have to join various tables to generate results).

Simpler to derive business insights.

**What are some of the disadvantages of using a star schema versus a relational model?**

Data integrity is not enforced well since in a highly de-normalized schema state.

Not flexible in terms if analytical needs as a normalized data model.

Star schemas don't reinforce many-to-many relationships within business entities – at least not frequently.

**What criteria might you use to evaluate which approach to use in a given situation**?

Data Quality

**For Part 2: Examining Administrative Data**

**In 5-6 sentences, describe the data: Why was it originally collected? What were the sources of the data? Who collected it, ie. Doctors, nurses, pharmacy, self-reported, etc. Where was it collected? How is the context identified?**

The source of the CMS Beneficiary Summary DE-SynPUF is a 5% sample of 2008 Medicare beneficiaries and their claims from 2008, 2009, and 2010.

**What sorts of things could this data measure? How might these measures identify opportunities to improve patient care or health care operations? Who would benefit from this information?**

Allow data entrepreneurs to develop and create software and applications that may eventually be applied to actual CMS claims data;

Train researchers on the use and complexity of conducting analyses with CMS claims data prior to initiating the process to obtain access to actual CMS data; and,

Support safe data mining innovations that may reveal unanticipated knowledge gains while preserving beneficiary privacy**.**

**What are some of the limits to the use of this data?**

**How might you verify and validate this data?**

**How difficult would this data be to map for comparison to other data?**

**What might be some possible issues with these data in contrast to the more rich clinical type of data?**