**Team Members:** Mahesh M. Nair (mmn9), Jigyasa Kohli(jk423)

Database: <https://drive.google.com/open?id=0B27AGhF7KpsVZ2lyOUpVR1FsdE0>

**Step4**

1. The MSHA\_Full.csv file was over 2.5GB with 2.4 million rows
2. Breaking down the main table “MSHA\_Full” into smaller tables for performance. Except for msha\_mine\_inspect and msha\_mine, all other tables query within seconds.
3. This data would be queried by the Government and the public. Splitting it would increase the querying especially in web apps. Have the dataset split by inspection and mine in a whole and then descriptions of codes would be set into separate tables. When data is queried the user would not want all data about the mine to be displayed if it is going to take minutes to load. MSHA\_full took 5 mins plus to query, splitting it would reduce by 70%.
4. Lot of columns had space for data, using case to avoid this value.

