Class of 04/08/2019

Case Study Walkthrough - Data Integration

- Problem: Analyze the H1B Visa Application Dataset.
- Step 1 - Identify entity types associated with the same table, and divide the table by entity types using SQL transforms
  - Issue - joins between the Employer_Temp Table and other tables associated with it are not related via the Employer_Temp's primary key
    - Other tables could potentially be referencing multiple rows in Employer_Temp, because a combination of employer_name, employer_city, and employer_state is not guaranteed unique
- Step 2 - use Beam transforms to normalize names, city, state, and duplicates, as well as changing employer_name, employer_city, and employer_state fields in related tables with employer_id
  - Reduces the foreign key count from 3 potentially non-unique fields to a primary key (guaranteed unique)
  - requires 3 pipelines for each table - normalization and removal of duplicates from Employer_Temp, and fixing foreign keys in Job and Application
- Step 3 - find a secondary dataset and, using transforms, relate it to the first dataset
  - Acknowledge which rows you will need and reduce
  - Issue - corporation names contain punctuation/require normalization
    - Solution? Apply beam transforms to normalize
    - Why is it preferable to match on a numeric ID?