Big Data Training

Dan Kramer

1. Explore the data in the pyspark shell
   1. /user/natesan.dhanasekar/sparklab/data/vocabulary/
   2. Explore the schema for both concept and concept\_relationship
   3. Load the CSVs into dataframes
      1. Hint: You need to specify that there’s a header and that the delimiter is tab “[\\t](file:///\\t)”
   4. Use .printSchema to look at the columns
2. We’re going to make a mapping of various ICD to SNOMED codes
   1. We’re going to hard code paths – don’t do this in practice!!
   2. Load in concept and filter for valid codes
   3. Load in CONCEPT\_Relationship and filter for valid codes
      1. Filter the reltationship\_id column to when it’s “ICD9P - SNOMED eq” or “Maps to”
   4. Join concept with the concept\_relationship table
      1. What columns do you think we should join on?
      2. Rename a few the columns corresponding to that concept because we’ll be joining with the concept table again
      3. Select only necessary columns
   5. Join concept once again
      1. Rename columns in a similar fashion but for the second concept
   6. Filter the final output so that we have only ICD codes on the left, and we don’t have ICD to ICD mappings
   7. Run a show() on the final output:

