**Raw data (text format) – 21M**

**Data Preparation and Cleaning involves following steps:**

1. Raw data (text format) – 21M records.
2. Append the current month to Business ID to get B2BuniqueID which is unique ID for the company.
3. Sample record of size 1000 is taken and a quick Quality Check (QC) is done.
4. Roll up the data :
5. Input 21M record and get 2\_SIC code & Description, 4\_SIC code & Description and 6\_SIC code & Description by considering the most occurring values and eliminating repeated ones. So that we get to keep the unique ones. Sec\_SIC, Sec\_SIC\_Desc and Third\_SIC & Third\_SIC\_Desc fields are optional as it depends on the data.
6. The Least value is considered for Get Company size, Company\_sales\_volume, Corporation Emp size and Corporation sales volume fields in getting the unique records.
7. Get NAICS code & Description by referring to step 5.
8. Merge all the files and get the unique records of 14M size.
9. Create metadata for the unique records.

**End up with Unique Records - 14M**