# Extract Transform and Load Tutorial – ETL Data Quality and cleaning

# ETL – 1

This tutorial works with the insurance data sets S\_0\_DataSet1 and S\_0\_DataSet2 to support the reports such as: No of claims per year per location.

Each dataset represents data from insurance companies, one operates across multiple states in America the other is just from Washington (S\_0\_DataSet1 and S\_0\_DataSet2).

# ETL 1 – overview

In this first tutorial we will implement the data dictionary design. There is a template for this Data Dictionary 1.

This tutorial:

1. Load the raw data (data sets S\_0\_DataSet1 and S\_0\_DataSet2) into Apex
2. Explore the datasets and consider the data quality issues in each data set.
3. Identify the columns in each dataset that are required to support the report: No of claims per year per location.
4. EXTRACT: Create a dataset that contains JUST the data you need to work out how many claims per year per location from both data sets (call this S1\_StageArea). You could have a number of these ‘staging’ datasets)
5. Use code or tools, to find data quality issues with the data sets S1\_StageArea*. Document these in the Data Dictionary using the Data Dictionary 1 template.*
6. Use code to clean (quality checks) S1\_StageArea.
7. Consider the use of a table to log the changes made to the data.
8. Create the star schema model for the exercise in QSEE and create the DDL for the SS. It is good practice to prefix the star schema tables with SS.