Exploratory Data Analysis

Write Up

Prior to conducting Exploratory Data Analysis, data pre-processing steps were executed to assess the integrity of the data from each of the 3 datasets, and to undertake any data cleaning where necessary. Based on the data pre-processing, valuable insights were generated that would subsequently assist in developing the architecture that would define how the data is to be stored for optimal query speed and compression of storage files.

In the case of the *NOAA: Severe Storm Event Details* dataset, it had been found that certain fields pertain to a subset of weather events. For example, the dataset contains columns which record information that is only captured in the event of a tornado. Similarly, the *flood cause* field only captures information for weather events which would naturally lead to flooding events. Given that the *NOAA: Severe Storm Event Details* dataset is comprised of a wide range of weather event types, each with its own distinct set of characteristics and traits to be recorded, the data can be configured following the conventions of a star schema (as discussed in the following paragraph).

With reference to the Entity Relationship Diagram (refer to the following page), the *NOAA: Severe Storm Event Details* dataset can be organized such that attributes that are only applicable to certain weather events may be stored in separate tables (e.g. Tornado, Magnitude and Flood). In addition, information pertaining to the location of a weather event will also be stored separately, as it has been found that a *location index* is only provided if an instance of a weather event belongs to a broader storm system. In configuring the *NOAA: Severe Storm Event Details* dataset in such a manner, the data is organized in a star schema with the *Severe Weather Events* table serving as the fact table, and the following acting as dimensions: *Tornado, Magnitude, Flood, State, County and Location*.

Entity Relationship Diagram