GDELT GKG EDA Insights

The GDELT Global Knowledge Graph (GKG) Dataset updates every fifteen minutes as a .CSV file appended to the master file list URL.
Preliminary exploratory data analysis on files within the dataset have shown that each file consists of 27 tab delimited columns.
Each column corresponds to a code outlined within the GDELT v2 Codebook. Some fields are references to external sources while others are flattened data, represented by other tables.
Upon inspection, GDELT GKG files reveal themselves as a typical star schema with a single fact table referencing multiple dimension tables.
Among the 27 columns are multiple fields containing denormalized models - nested structures which are collapsed dimensional models.

For example, looking at the V2GCAM field, in the documentation it is outlined as a series of comma-delimited blocks containing colon-delimited key-value pairs. A subset of data found within this field may look like this:

wc:125,c2.21:4,c10.1:40,v10.1:3.21111111

Referencing the GCAM Master Codebook at: http://data.gdeltproject.org/documentation/GCAM-MASTER-CODEBOOK.TXT

This would be inflated to a table resembling:

Type	Count
WordCount:	125
General Inquirer Bodypt	4
SentiWordNet	40
SentiWordNet average	3.21111111

Multiple fields work in a similar way such as V2Locations, V2Persons, V2Organizations, etc.

As such, to be able to gain insights from the data contained within these fields, they will at some point need to be joined to their respective reference table.

The proposed storage solution is contained within ‘storage_plan.rft’