|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Source (s)** | **Field Name** | **Data Type** | **Key** | **DQ Source** | **Data Quality Check** | **Data quality Issues** | **Action Note** |
| DataSet1 | **State** | Varchar2 | No | DataSet1 | Inconsistency | Washington and WA values | Transform data to full name state |
|  |  |  | DataSet1 | Missing value | Missing for customers QZ42725 and ZH19885 | As all customers are for WA, these two customers should also have a value for the state |
| DataSet2 | Varchar2 | No | DataSet2 | Inconsistency | Full state name for all apart from Arizona - AR | Transform data to full name state |
| Definition: | State column represents insurance data on customers in US state Washington | | | | | | |
| Notes: | State column exist in both of our data sets. | | | | | | |
| DataSet1 | **Gender** | Varchar2 | No | DataSet1 | Inconsistency | has no neutral gender but blank field |  |
| DataSet2 | Varchar2 | No | DataSet2 | Inconsistency | Gender column – male and M, female and F  Gender column – has neutral gender N |  |
|  |  |  |  |  |  |  |
| Definition: | Column that holds vales on customer demographics, such as gender | | | | | | |
| Notes: | Gender Column exist in both data sets | | | | | | |
|  |  | | | | | | |

Data Dictionary 1 – To support data sourcing and data integration

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Star schema table** | **Attribute name** | **Data Type** | **Key** | **DQ Source** | **Data source field (mapping)** | **Data quality Issues** | **Transformation** |
| Time\_dim | Time\_id | Number2 | Yes | Automatically generated as a surrogate key | n/a | n/a | Create a sequence time\_seq to support the generation of surrogate keys |
| The\_year | Varchar2 | No | DataSet1 | Effective\_to\_date | Some dates are missing the year part  Is this the date we need here? Effective to? Not date\_taken out? | Will need to extract the year from the date and get distinct values only |
| Definition: | The time\_dimension holds the intervals of time for which the data will be held. It is held at year level meaning now reports are available at a lower granularity. | | | | | | |
| Notes: |  | | | | | | |
| **Star schema table** | **Attribute name** | **Data Type** | **Key** | **DQ Source** | **Data source field (mapping)** | **Data quality Issues** | **Transformation** |
| Time\_dim | Time\_id | Number2 | Yes | Automatically generated as a surrogate key | n/a | n/a | Create a sequence time\_seq to support the generation of surrogate keys |
| The\_year | Varchar2 | No | DataSet1 | Effective\_to\_date | Some dates are missing the year part  Is this the date we need here? Effective to? Not date\_taken out? | Will need to extract the year from the date and get distinct values only |
| Definition: | The time\_dimension holds the intervals of time for which the data will be held. It is held at year level meaning now reports are available at a lower granularity. | | | | | | |
| Notes: |  | | | | | | |

Data Dictionary 2 – To show ETL required for star schema