# **Requirements**

**Executive Summary**

Our data warehouse stores data on global wildlife trading statistics, including trading countries, trading quantity, trading species, trading prices, and trading purpose, etc. Our data warehouse could be used for learning the wildlife trading trend in the world and helping animal protection organizations to know their work concentration. Potential users of our data warehouse include World Animal Protection and World Wildlife Fund.

**List of Business Questions**

* Which country imports the most quantity and the most value of animals each year?
* Find the average quantity of transaction and obtain which year has the highest wildlife transactions?
* Top 3 species in quantities imported by the United States from 2015 to 2019?
* Top 3 countries which imported the largest quantity of wildlife each year?
* Which month has the most value of exports and imports each year?
* Flora or fauna are imported on which purpose most by each country?

**Business Terms and Definitions**

Term: The good being exported or imported. (Live for live specimens, skin for animal skins, and so on.)

Appendix: the appendix the species belongs to. Appendix I species are the most protected, and Appendix III species are the least.

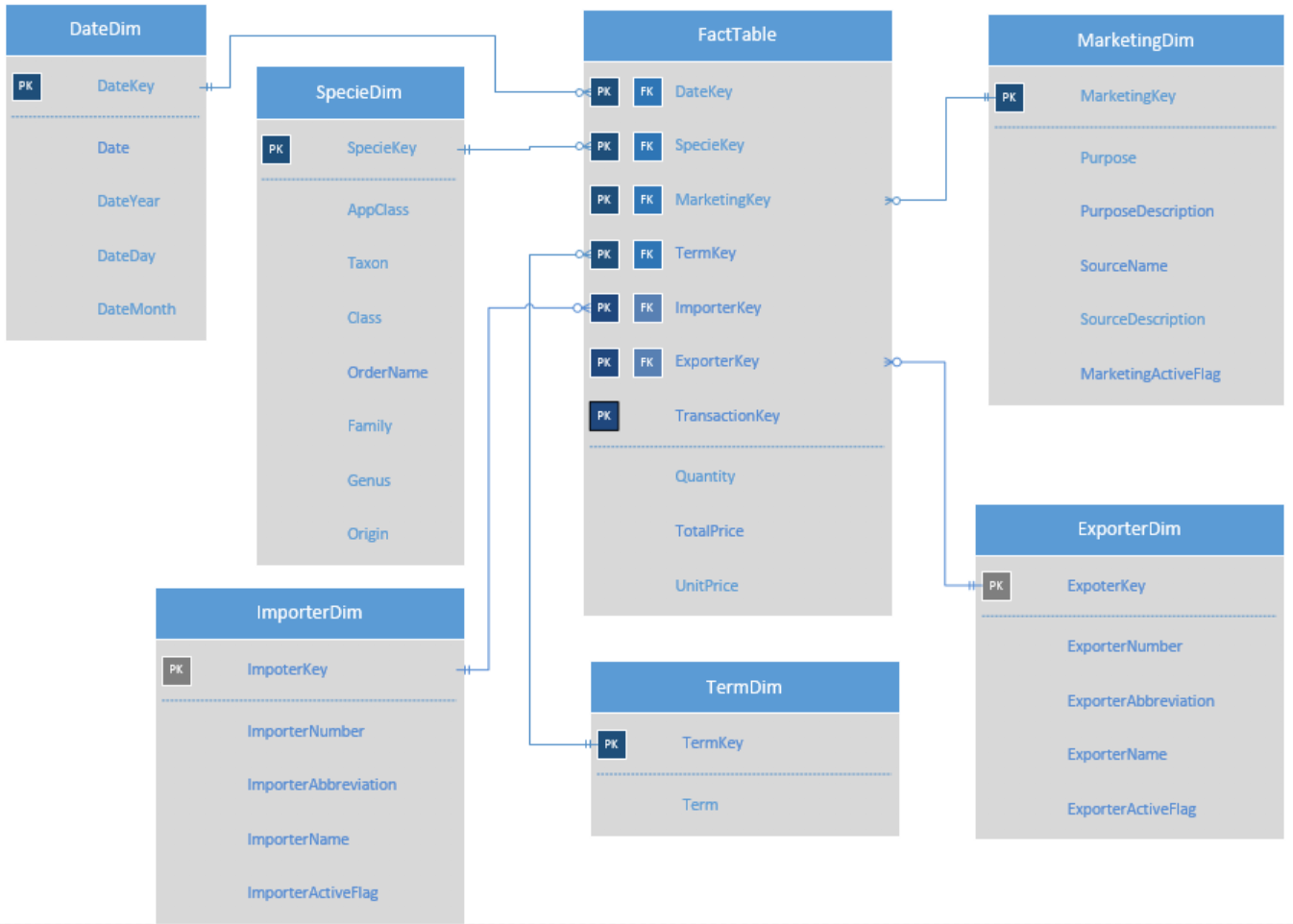
Taxon, class, order, family, and genus are all taxonomic terms; the genus is the scientific name of the species.

**Data Source**

Our data source is downloaded online via this link: <https://www.kaggle.com/cites/cites-wildlife-trade-database>. Each row of the dataset refers to a transaction of import or export. The attributes of this dataset include taxon, class, order, family, genus etc. that describing the biological classification of the species, date information pointing out the transaction date, purpose of this transaction, source of the transacted flora or fauna, their term of transaction, import and export country, quantity of the transacted species and price. Every transaction involves two countries to participate, one species of flora or fauna to be transacted, a particular purpose and a source, at a specified price, of a quantity on a stated day.

# **Design**

**Detailed Data Model**



|  |  |
| --- | --- |
| **Table** | **Number of records** |
| FactTable | 365000 |
| DateDim | 3650 |
| SpecieDim | 9900 |
| ImporterDim | 220 |
| ExporterDim | 220 |
| MarketingDim | 120 |
| TermDim | 84 |

**Fact Table Strategy**

The trading information in the data warehouse is the annual trading summary of different countries. The frequency of data collection is once per year, but different countries may submit their data at a different date. So we will reload the fact table once per season.

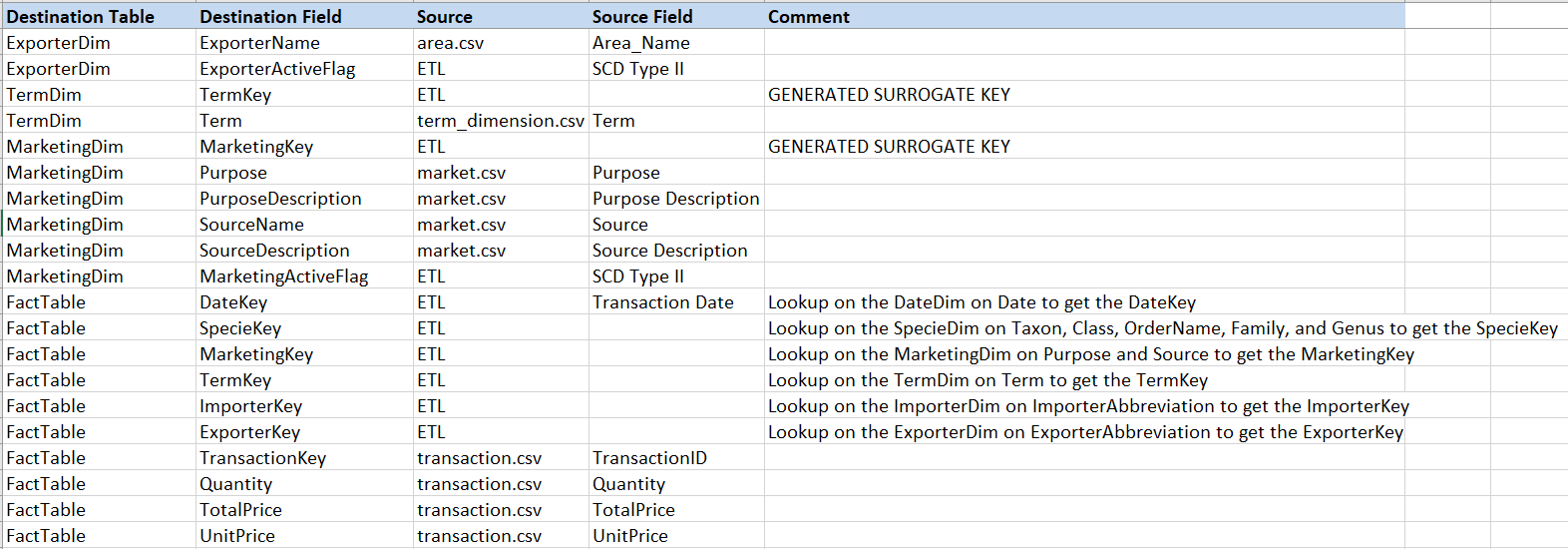
The grain of fact table: Transaction, date, species, marketing, term, importer, exporter

**Dimensional Load Strategy**

Our original dataset is a whole spreadsheet of data, means that we need to extract data from it and load the data into the data warehouse. Therefore, we plan to separate the whole dataset into several tables, and tables are related via foreign keys in OLTP schema. Then, we are able to load data from those tables to the data warehouse.

For Marketing Dimension, we plan to apply Type II SCD on these following attributes: PurposeDescription and SourceDescription. Each purpose code and source code can have different meanings over time, such as “B” indicates “Breeding in captivity or artificial propagation” now, but it used to represent “Craft making”. For Importer Dimension, we plan to apply Type II SCD on ImporterName, since a country may modify its name but it rarely happens. Similarly, we also plan to apply Type II SCD on ExporterName.

**Technical Metadata Mapping**



**Physical Capacity Plan**

Dimension:

* Date Dimension: 3650 dates (10 years) \* 1 KB per row = 3.65 MB
* Specie Dimension: 9900 species \* 1 KB per row = 9.90 MB
* Importer Dimension: 220 countries \* 1 KB per row = 0.22 MB
* Exporter Dimension: 220 countries \* 1 KB per row = 0.22 MB
* Term Dimension: 84 terms \* 1 KB per row = 0.08 MB
* Marketing Dimension: 120 types \* 1 KB per row = 0.12 MB

Total Dimension = 14 MB

Fact:

quantity of record = 3650 dates \* 10 trade per day \* 10 species \* 1 importer per trade \* 1 exporter per trade \* 2 terms \* 5 marketing types = 365000 records

Fact table size = 3650000 records \* 1 KB per record = 3650 MB

Total size = Fact + Dimension = 3650 MB + 14 MB = 3.7 GB

**Miscellaneous**

To meet requirements, we add the following four attributes into the dataset: unit price, total price, importer name, exporter name, source description, and purpose description.

|  |  |
| --- | --- |
| Attribute name | comment |
| unit price | The original dataset doesn't have a price, in order to make our data warehouse more informative, we decide to add a price attribute. |
| total price | The total price is calculated by unit price multiplied by quantity. |
| importer name | The original dataset only includes importer country’s abbreviation, in order to identify the country more conveniently, we add the country’s full name into our dataset. |
| exporter name | The original dataset only includes importer country’s abbreviation, in order to identify the country more conveniently, we add the country’s full name into our dataset. |
| source description | The original dataset only has an abbreviation of source; we add descriptions of the sources. |
| purpose description | The original dataset only has an abbreviation of purpose; we add descriptions of the purposes. |

# **Appendix**

Metadata mapping table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Destination Table** | **Destination Field** | **Source** | **Source Field** | **Comment** |
| DateDim | DateKey | ETL |  | GENERATED SURROGATE KEY |
| DateDim | Date | date\_dimension.csv | Date |  |
| DateDim | DateYear | date\_dimension.csv | DateYear |  |
| DateDim | DateDay | date\_dimension.csv | DateDay |  |
| DateDim | DateMonth | date\_dimension.csv | DateMonth |  |
| SpecieDim | SpecieKey | ETL |  | GENERATED SURROGATE KEY |
| SpecieDim | AppClass | specie\_dimension.csv | AppClass |  |
| SpecieDim | Taxon | specie\_dimension.csv | Taxon |  |
| SpecieDim | Class | specie\_dimension.csv | Class |  |
| SpecieDim | OrderName | specie\_dimension.csv | Order |  |
| SpecieDim | Family | specie\_dimension.csv | Family |  |
| SpecieDim | Genus | specie\_dimension.csv | Genus |  |
| SpecieDim | Origin | specie\_dimension.csv | Origin |  |
| ImporterDim | ImporterKey | ETL |  | GENERATED SURROGATE KEY |
| ImporterDim | ImporterNumber | area.csv | Area\_Number |  |
| ImporterDim | ImporterAbbreviation | area.csv | AreaAbbreviation |  |
| ImporterDim | ImporterName | area.csv | Area\_Name |  |
| ImporterDim | ImporterActiveFlag | ETL | SCD Type II |  |
| ExporterDim | ExporterKey | ETL |  | GENERATED SURROGATE KEY |
| ExporterDim | ExporterNumber | area.csv | Area\_Number |  |
| ExporterDim | ExporterAbbreviation | area.csv | AreaAbbreviation |  |
| ExporterDim | ExporterName | area.csv | Area\_Name |  |
| ExporterDim | ExporterActiveFlag | ETL | SCD Type II |  |
| TermDim | TermKey | ETL |  | GENERATED SURROGATE KEY |
| TermDim | Term | term\_dimension.csv | Term |  |
| MarketingDim | MarketingKey | ETL |  | GENERATED SURROGATE KEY |
| MarketingDim | Purpose | market.csv | Purpose |  |
| MarketingDim | PurposeDescription | market.csv | Purpose Description |  |
| MarketingDim | SourceName | market.csv | Source |  |
| MarketingDim | SourceDescription | market.csv | Source Description |  |
| MarketingDim | MarketingActiveFlag | ETL | SCD Type II |  |
| FactTable | DateKey | ETL | Transaction Date | Lookup on the DateDim on Date to get the DateKey |
| FactTable | SpecieKey | ETL |  | Lookup on the SpecieDim on Taxon, Class, OrderName, Family, and Genus to get the SpecieKey |
| FactTable | MarketingKey | ETL |  | Lookup on the MarketingDim on Purpose and Source to get the MarketingKey |
| FactTable | TermKey | ETL |  | Lookup on the TermDim on Term to get the TermKey |
| FactTable | ImporterKey | ETL |  | Lookup on the ImporterDim on ImporterAbbreviation to get the ImporterKey |
| FactTable | ExporterKey | ETL |  | Lookup on the ExporterDim on ExporterAbbreviation to get the ExporterKey |
| FactTable | TransactionKey | transaction.csv | TransactionID |  |
| FactTable | Quantity | transaction.csv | Quantity |  |
| FactTable | TotalPrice | transaction.csv | TotalPrice |  |
| FactTable | UnitPrice | transaction.csv | UnitPrice |  |