Data Warehouse Design Report

Group F
Mariana Narvaez Cubillos
Nisrine Ferahi
Jonas Hellevang
Daniel Bilitewski
Rabiga Shangereyeva
Guillermo Chacon Clericus

Master in Business Analytics and Big Data October 2019 – Section 1
Business Intelligence & Data Warehousing – Professor Josep Curto

A – Dataset Analysis

A.1 – Column Meanings and User Requirements

This is a summary of the column meanings and user requirements. For a detailed table analyzing each column, refer to Appendix A.

The methodology to determine the column meanings and user requirements was a mixture between researching the column meanings and purposes on the FCC website and other online sources as well as determining the relationships between the columns through data exploration using Microsoft Excel and R. Overall, the columns can be grouped into the following categories:

Identifiers. Identifiers can have both technical and legal purposes. Technical purpose refers to the unique IDs required when building tables in databases. Legal purpose refers to numbers or alphanumeric codes used to identify for example the companies owning a license or the license ID itself, which will appear on legal and official documents. For the FCC dataset (license ID, FRN, antenna ID, ASR number, etc.), it can be concluded that all identifiers have more of a legal purpose than a technical purpose, therefore they need to all be included in the data warehouse as they are. In some cases, they can also be used as primary keys in tables, in other cases, a new unique identifier needs to be created.

License Holder Information. Information such as name and address of the entity holding the license, and if it is a legal person, natural person, government entity or other. Used to give additional information beyond the identifier about the entity owning the license.

Contact Person & Company Information. Gives the address and contact information like phone, fax and email which is different to the license holder address. These columns refer to the company which is operating the service under the license. Since license holder and the company operating the service under the license are different, this information is required by the FCC to directly contact the operating entity.

Antenna Location & Technical Information. These columns are used to describe the attributes of the antenna structure through which the service is delivered within a specific zone. This information is required to manage the issuing of frequencies in different markets and to ensure the technical feasibility across the radio frequency spectrum.

Frequency & Emission Information. A major role of the FCC is to manage the range of frequencies it can assign to license holders. These columns inform about the frequencies at which antennas are emitting and the emission on each frequency.

A.2 – Data Quality Issues

This is a summary of the data quality issues. For a detailed table analyzing each column, refer to Appendix B.

The methodology to find data errors was to load columns into an R workspace. Since the dataset has over 18 million rows and a size of over 9 GB, the columns had to be loaded and analyzed individually. The main data issues are summarized in the following:

Missing Values & Empty Strings. All columns have missing values or empty strings, except the license ID. Often times, there is a specific business reason for that. For example, if no service is operated under the license, there will be no antenna, and therefore this kind of fields will be empty. In other cases, like

for the FRN (an identifier within the FCC for the entity owning the license), there shouldn't be empty fields. However, it is not possible to remove rows with empty fields, since in that case rows with active licenses would be removed. It is also not possible to remove columns, since that would remove important information that is available for other rows. Finally, it is not possible to interpolate data, since most of them are strings and the numeric columns don't have the structure required for data interpolating. Therefore, no action will be taken on missing values and empty strings will be converted to NULL, to allow SQL queries to exclude these rows by using the IS NOT NULL command. A special case is the Grant Date, Expiration Date, Cancellation Date and Last Action Date columns, because they are completely empty. These columns are very important, since they determine the active status of the license, therefore they will not be discarded, even though they don't contain data. In the future, users should be forced to enter a value into these fields, or they should be filled automatically to fix this data issue.

Other Data Errors. Many identifiers also have a so-called Sequence ID. For example, if more than one frequency is emitted on the same antenna, they are numbered by the Sequence ID. However, not all sequence IDs follow that structure. Therefore, entries not corresponding to a numbering have to be identified and replaced by the number 1, since in these cases usually there is not more than 1 frequency on that antenna. The same applies to other cases of Sequence IDs. Finally, the contact information data is very unreliable. For example, there are multiple phone numbers which appear hundreds of thousands of times. Since fixing this type of errors like the phone numbers is out of the scope of this project, we will not make corrections on these columns.

A.3 – Detect Duplicates

There are multiple instances of two columns, where one is a short code and the other is a description. Since they contain the same information, the code columns will be discarded. This way, the following columns will be discarded:

- Rollup Category Code
- Radio Service Code
- License Status Code
- Rollup Status Code
- Entity Type Code
- Rollup Entity Code
- Market Code
- Location Type Code
- Loc. County Code
- Antenna Type Code

Also, the column Facility ID has 98% missing values. In the case where values exist, they appear to be the last four digits of the License ID. Since this is also duplicate information, the Facility ID column will be discarded as well.

B – Data Warehouse Approach Selection

In order to select the best data warehouse schema, it's essential to understand the overall purpose and use case of the database at hand. The FCC mission statement is[1]:

> *"The Federal Communications Commission regulates interstate and international communications by radio, television, wire, satellite, and cable in all 50 states, the District of Columbia and U.S. territories. An independent U.S. government agency overseen by Congress, the Commission is the federal agency responsible for implementing and enforcing America's communications law and regulations."*

The FCC states what they do on their website in the following way[2]:

> *"The Federal Communications Commission regulates interstate and international communications by radio, television, wire, satellite and cable in all 50 states, the District of Columbia and U.S. territories. An independent U.S. government agency overseen by Congress, the commission is the United States' primary authority for communications law, regulation and technological innovation. In its work facing economic opportunities and challenges associated with rapidly evolving advances in global communications, the agency capitalizes on its competencies in:*
>
> - *Promoting competition, innovation and investment in broadband services and facilities*
> - *Supporting the nation's economy by ensuring an appropriate competitive framework for the unfolding of the communications revolution*
> - *Encouraging the highest and best use of spectrum domestically and internationally*
> - *Revising media regulations so that new technologies flourish alongside diversity and localism*
> - *Providing leadership in strengthening the defense of the nation's communications infrastructure"*

From this information from the FCC website it can be concluded that the purpose of the data at hand is to help the FCC regulate, especially with a focus on the best use of the spectrum.

To choose the data warehouse schema, the user requirements and data structure of the source data need to be evaluated against the advantages and disadvantages of the snowflake schema and data vault schema. This will be done across two criteria: change of user requirements over time and complexity of queries.

Change of User Requirements over Time. Data vault has the advantage over a snowflake schema that when ingesting data from new sources, the original tables can be retained[3]. This is useful for the regulatory purpose of the FCC, because it makes it easier to retain historical data if user requirements should change. User requirements by the FCC have changed over time, this can be seen by for example by the large percentage of empty fields. This is due to new technologies not requiring certain columns in the table, for example wireless technology doesn't require an antenna registration. With the upcoming transition to 5G, the change in user requirements is unclear at the moment. Another example is that there are five different data sources, which provide data of different services, such as radio, television, cable, etc., suggesting new data sources will appear over time. To tackle this challenge, the data vault schema provides more flexibility.

[1] Source: https://www.fcc.gov/about/overview
[2] Source: https://www.fcc.gov/about-fcc/what-we-do
[3] Source: Schnider et al. "Comparison of Data Modeling Methods for a Core Data Warehouse". Page 18.

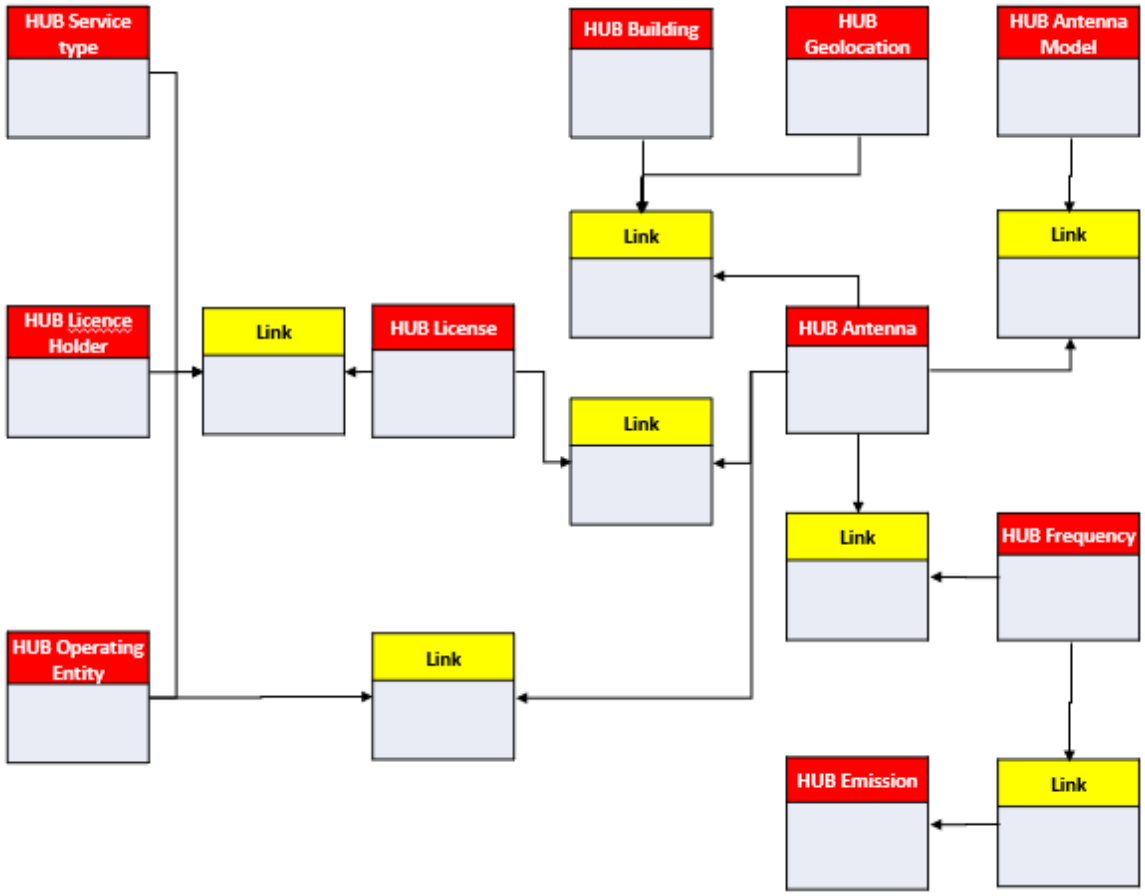
Querying. Making queries with a snowflake schema is simpler as less joins are required. This advantage becomes more significant, the more complex queries and the more often the data warehouse needs to be queried. Based on the user requirement being regulation, the queries will be very standardized and not very complex, since no complex data exploration is required. Furthermore, the number of changes to licenses and new licenses issued will be limited. Due to the low creation of new data, the overall number of queries will be low compared to other use cases. Finally, in a snowflake schema, the fact table would be the licenses. This means that all queries will need to focus around that and therefore, for example joining directly the operating entity and the antenna or frequency, will not be possible in a snowflake schema without linking through license. Through the use of hubs and links, the data vault provides a direct creation of the links between these concepts.

In conclusion, the advantages of the snowflake schema in terms of querying are not significant to this use case. In terms of changing user requirements over time, the data vault provides more flexibility and through the ability to retain the original data structure, it's better suited for regulation, which is the FCC's purpose. Therefore, the data vault will be chosen as the schema for the data warehouse.

C – Data Warehouse Design

This part will start with the conceptual design of the data warehouse, describing hubs and links. Following that is the logical design, adding satellites, primary keys and foreign keys. The physical design is implemented using MySQL Workbench.

C.1 – Conceptual Design



The conceptual design includes all the hubs used in the data warehouse model and the links between them. For the dataset at hand, ten hubs can be identified:

License Hub. One of the two main purposes is the regulation of licenses and whether they are active or not. This hub has one satellite with the information about licenses status, dates and permissions.

License Holder Hub. Since all licenses have a license holder, but one license holder can have multiple licenses, the license holder will be a separate hub. Its satellite contains all the information about the license entity account.

Operating Entity Hub. The entity operating the service under the license is not the same as the license holder, therefore it requires its own hub. The satellite contains the data about the company operating the service under the license and its contact information.

Service Type Hub. The types of services fall in different categories, therefore the type of service delivered under the license will receive its own hub. The corresponding satellite describes the radio service type, rollup service, market and channel block information.

Antenna Hub. The second main purpose of the FCC, aside from regulating licenses, is to regulate the frequency spectrum. A key component of that is the antenna, which connects to all other hubs required to regulate the frequency spectrum. It's a hub because the antenna has various important characteristics that pertain to itself.

Antenna Model. Since the number of antenna models is limited, it justifies giving antenna models its own hub. The correspond satellite contains complementary data for antenna about maker, model and type.

Building Hub. Since multiple antennas can broadcast from the same building, the building will be described by a hub. The satellite corresponding to it describes the station that transmit with the antenna.

Geolocation Hub. The location from which the antenna emits is key to managing the frequency spectrum. Of course, each building only has one location. However, building and geolocation should be separate hubs as they address different concepts that are relevant to FCC regulators. The satellite corresponding to the geolocation hub describes the location data in latitude and longitude variables.

Frequency Hub. This hub is for the key concept around regulating the frequency spectrum, the frequency itself. Multiple frequencies can be emitted from the same antenna. Therefore, the antenna and frequency are separate concepts. The corresponding satellite contains all the frequency attributes.

Emission Hub. Since there can be multiple emissions on the same antenna, the emission information will be considered a separate concept. The corresponding satellite describes the attributes of emissions.

As shown in the conceptual model, there are seven links connecting those hubs.

Link 1. Connects the company, license owner and radio service information with the license. This is because a license has an owner and a service is delivered by an operating entity under the license. The operating entity is usually owned by the license owner. Therefore, all the concepts shown in the four hubs are connected.

Link 2. Connects license with antenna, as it is required to have a license to operate an antenna.

Link 3. Connects company with antenna, this is in order to make specific queries about antennas much easier. Since the company is operating the antenna, because of this link it is not required to connect the two through the license hub.

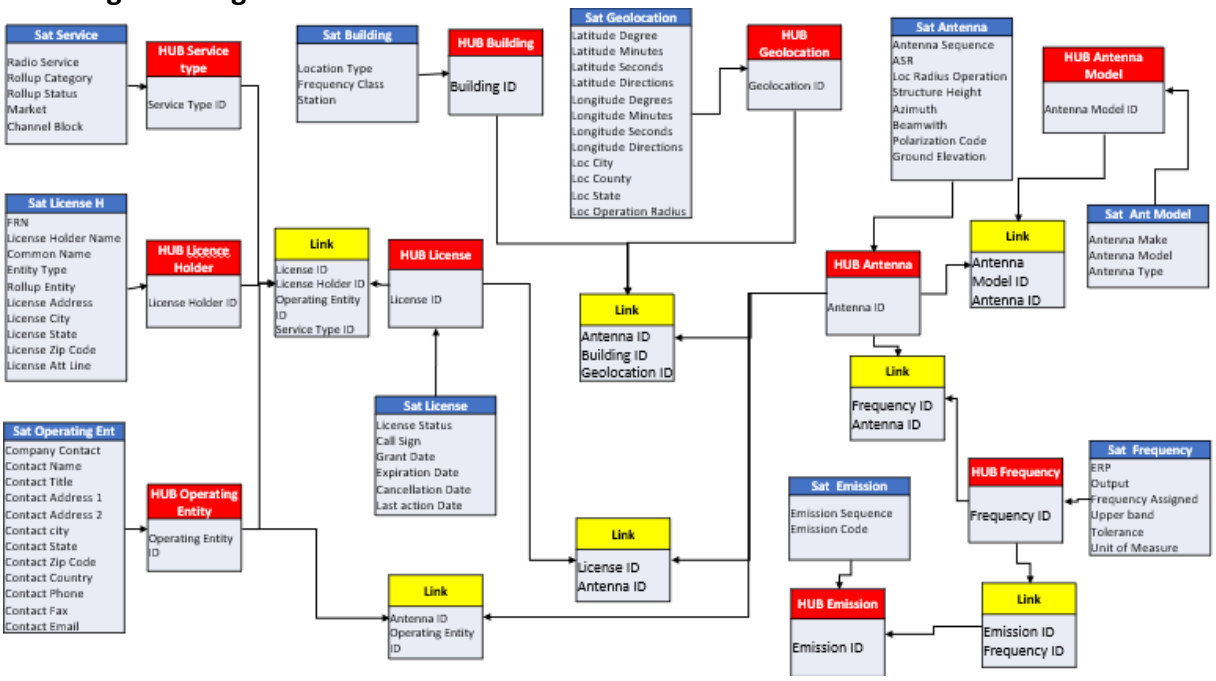
Link 4. Connects the antenna with building and geolocation, because every antenna is on a building, which has a fixed location.

Link 5. Connects antenna with its model information, as all antennas have a certain model.

Link 6. Connects antenna with frequency. There can be multiple frequencies on the same antenna, which is modelled by this link.

Link 7. Connect frequency with emissions. There can be multiple emissions on the same frequency, which is modelled by this link.

C.2 – Logical Design



The following schema is the logical design that shows the satellites with its attributes, the primary keys and connections within variables.

C.3 – Other Considerations

Data Types. Appendix A shows the data types as they are in the original file. Columns which are strings in the original file will be of type VARCHAR(45) in the physical design, except if there is reason to believe that 45 characters is not enough, in such cases VARCHAR(100) was chosen. Strings which are codes with fixed number of letters will have that many characters, e.g. the columns for the two-letter code for US states will be of type CHAR(2). Columns which are of types datetime, integer or decimal in the original file will continue to be of the same type respectively.

Primary Keys. The combination of IDs and Sequence IDs can be used as primary keys for many tables. However, since the data on Sequence IDs is very messy, it is a safer choice to create primary keys from scratch and make them auto-incremental. The information from the IDs and Sequence IDs will be retained, in case they are required for official documents. To show this, these columns will receive the suffix "_official".

Appendix A – Meaning and User Requirements of Columns

Note: Data Type refers to the data type in the original file, not the data type chosen for the MySQL database.

Column Name	Data Type	Meaning	User Requirements
License ID	String	Assigned ID that identifies a license.	Legal and official documents.
Facility ID	String	Assigned ID that identifies a broadcast station. There are rows without facility ID, in which case the license ID is four digits shorter. Was introduced in 1981, so facilities whose license is too old don't have an ID. Before, stations had history cards, they were eliminated, corresponding to a digitalization of the process.	Since it's the last four digits of the License ID, it provides no additional information to the user.
Call Sign	String	Identifier assigned to radio and television stations. Identifies the service, meaning that one company that offers more than one service has more than one call sign.	Legal and official documents.
FRN	String	Identifier for any individual or company doing business with the FCC. A company can have multiple licenses but only one FRN.	Legal and official documents.
License Holder Name	String	Name of the holder of the FRN.	Connects FRN to name of the entity.
Common Name	String	Entity A that has control over the entity B controlling the license. Or name under which the service is marketed to the public.	Connects License ID to an operating entity.
Radio Service	String	Type of radio or radio station, e.g. FM, AM, Digital, etc.	Descriptive information for categorization.
Rollup Category Code	String	High level category for describing the service.	Descriptive information for categorization.
Rollup Category Desc.	String	Description of the rollup category. It has the same information as the rollup category code.	Descriptive information for categorization.
License Status	String	Describes the status of the license.	Temporal information required to find out if the license is active, expired, or cancelled.
Rollup Status	String	Describes the status of the service. Four values: active, expired, cancelled, terminated.	Temporal information required to find out if the service is active, expired, or cancelled.
Entity Type	String	Legal form of the company owning the license.	Descriptive information for categorization.

Rollup Entity	String	Describes if the entity is a business, governmental, individual or other.	Descriptive information for categorization.
License Address	String	Address of the entity holding the license. Contains street or post box.	Contact information of the entity owning the license.
License City	String	Address of the entity holding the license. Contains has city.	As above.
License State	String	Address of the entity holding the license. Contains the state is its two letter code.	As above.
License Zip Code	String	Address of the entity holding the license. Contains the postal code.	As above.
License Attention Line	String	Unclear usage. However, said to be a required field according to FCC website.	As above.
Company Contact	String	Information of the contact person. Contains the contact person's company.	Contact information of the operating entity.
Contact Name	String	Information of the contact person. Contains the contact person's name.	As above.
Contact Title	String	Information of the contact person. Contains the contact person's job title.	As above.
Contact Address 1	String	Information of the contact person. Contains the contact person's street and house number.	As above.
Contact Address 2	String	Information of the contact person. Contains more detailed information on the address, such as the floor or suite in large office buildings.	As above.
Contact City	String	Information of the contact person. Contains the contact person's city.	As above.
Contact State	String	Information of the contact person. Contains the contact person's state.	As above.
Contact Zip	String	Information of the contact person. Contains the contact person's zip code.	As above.
Contact Country	String	Information of the contact person. Contains the contact person's street and house number.	As above.
Contact Phone	String	Information of the contact person. Contains the contact person's phone number.	As above.
Contact Fax	String	Information of the contact person. Contains the contact person's fax number.	As above.
Contact Email	String	Information of the contact person. Contains the contact person's email address.	As above.
Grant Date	Datetime	Date the license becomes effective, which initiates the license term, usually of 10 years.	Temporal information required to inform about the status of the license.
Expiration Date	Datetime	Date the license term ends.	As above.
Cancellation Date	Datetime	Date the license was cancelled.	As above.

Last Action Date	Datetime	Date when a last action was taken on the license.	Temporal information to track historical changes.
Radio Service Code	String	Two-letter code describing the type of service.	Same as *Radio Service.*
License Status Code	String	Short code for License Status	Same as *License Status*.
Roll Status Code	String	Short code for Rollup Status	Same as *Roll Status.*
Entity Type Code	String	Short code for Entity Type	Same as *Entity Type.*
Rollup Entity Code	String	Short code for Rollup Entity	Same as *Rollup Entity.*
Market Code	String	A market is a geographic area where the population receives a similar radio or television offering. This field contains the code.	Same as *Market.*
Market	String	A market is a geographic area where the population receives a similar radio or television offering. This field contains the long name of the market.	Required for regulating the frequency spectrum.
Channel Block	String	The cellular markets have a frequency range of 40 MHz. It is divided into two blocks, block A and block B of 20 MHz each.	As above.
Loc. Type Code	String	Location information contains info about the antenna. The type gives information about the structure holding the antenna. This column contains a short code.	As above.
Loc. Type Description	String	Location information contains info about the antenna. The type gives information about the structure holding the antenna. This column contains a longer description.	As above.
Loc. City	String	Location information contains info about the antenna. This column contains the city of the antenna location.	As above.
Loc. County Code	String	Location information contains info about the antenna. This column contains the county code of the antenna location.	As above.
Loc. County Name	String	Location information contains info about the antenna. This column contains the county name of the antenna location.	As above.
Loc. State	String	Location information contains info about the antenna. This column contains the state of the antenna location.	As above.
Loc. Radius Operations (Km)	Decimal	Location information contains info about the antenna. This column contains radius in kilometers that the antenna broadcasts.	As above.
Loc. Sequence ID	String	The column for the location ID seems to be missing, but from other columns we assume the following:	As above.

		Often times the last five or six digits of the location ID when there is only one location ID per license. If there are multiple locations for the same license, it takes numbers 1, 2, 3, etc.	
Loc. Latitude Degrees	Integer	Location information contains info about the antenna. This column contains the geolocation latitude in degrees.	As above.
Loc. Latitude Minutes	Integer	Location information contains info about the antenna. This column contains the geolocation latitude in minutes. Therefore, it provides an additional level of detail to the latitude in degrees.	As above.
Loc. Latitude Seconds	Decimal	Location information contains info about the antenna. This column contains the geolocation latitude in seconds. Therefore, it provides an additional level of detail to the latitude in degrees and minutes.	As above.
Loc. Latitude Direction	String	Location information contains info about the antenna. This column contains the geolocation latitude direction, meaning west or east of 0° latitude in Greenwich.	As above.
Loc. Longitude Degrees	Integer	Location information contains info about the antenna. This column contains the geolocation longitude in degrees.	As above.
Loc. Longitude Minutes	Integer	Location information contains info about the antenna. This column contains the geolocation longitude in minutes. Therefore, it provides an additional level of detail to the longitude in degrees.	As above.
Loc. Longitude Seconds	Decimal	Location information contains info about the antenna. This column contains the geolocation longitude in seconds. Therefore, it provides an additional level of detail to the longitude in degrees and minutes.	As above.
Loc. Longitude Direction	String	Location information contains info about the antenna. This column contains the geolocation longitude direction, meaning north or south of the equator.	As above.
Structure Height	Decimal	Height of the antenna. Unclear unit of measurement. Most likely centimeters or inches.	Regulatory requirement.
ASR Number	String	All antenna structures are required to register with the FCC. They are assigned the ASR number. ASR is short for Antenna Structure Registration.	Legal and official documents.
Antenna ID	String	Unique, auto-incremental identifier to identify the license.	Identifier.
Antenna Sequence ID	String	Often times the last five or six digits of the antenna ID when there is only one antenna ID per structure. If there are multiple	As above.

		antennas on the same structure, it takes numbers 1, 2, 3, etc.	
Antenna Make	String	Brand of the antenna.	Descriptive information for categorization.
Antenna Model	String	Model of the antenna.	As above.
Antenna Type Code	String	Antenna type describes the type of transmission. For example, a directional antenna transmits a stronger signal in a certain direction. This column contains the code.	As above.
Antenna Type	String	Antenna type describes the type of transmission. For example, a directional antenna transmits a stronger signal in a certain direction. This column contains the description.	As above.
Azimuth	Decimal	Direction in which an antenna is pointing, relative to true north. Expressed in degrees and measured clockwise.	Required for regulating the frequency spectrum.
Beamwidth	Decimal	Width of the broadcasting beam. Measured as an angle in degrees.	As above.
Polarization Code	String	Orientation of the electric field radiated by the antenna.	As above.
Frequency ID	String	Identifier of the frequency.	Identifier.
Frequency Sequence ID	String	Often times the last five or six digits of the frequency ID when there is only one frequency ID per antenna. If there are multiple frequencies on the same frequency, it takes numbers 1, 2, 3, etc.	As above.
Frequency Class Station Code	String	Station class contains the purpose for which a particular station is used. This column contains the code.	Descriptive information for categorization.
Frequency Class Station	Strings	Station class contains the purpose for which a particular station is used. This column contains the description.	As above.
ERP (Watts)	Decimal	Effective Radiated Power (ERP) is the perceived amount of power that the antenna is transmitting.	Required for regulating the frequency spectrum.
Output (Watts)	Decimal	Maximum output power of the antenna.	As above.
Frequency Assigned	Decimal	The actual radio frequency at which the antenna is transmitting.	As above.
Frequency Upper Band	Decimal	Certain types of transmission have a frequency range. In this case, the frequency assigned is the lower bound of the range, and this column specifies the upper bound.	As above.
Unit of Measure	String	Unit of measurement of the frequency in MHz or KHz.	As above.
Tolerance	Decimal	The percentage within the frequency assigned that you must transmit in.	As above.

Emission ID	String	One frequency can have several emissions, which are identified by this ID.	Identifier
Emission Sequence ID	String	Often times the last five or six digits of the emission ID when there is only one emission ID per frequency. If there are multiple emissions on the same frequency, it takes numbers 1, 2, 3, etc.	As above.
Emission Code	String	A code associated with a frequency that gives information about the frequency's bandwidth and the nature of the signal on the frequency.	Required for regulating the frequency spectrum.
Ground Elevation (Feet)	Decimal	Height of the antenna above average terrain.	Same as *Structure Height.*
Source System	String	FCC license system from which the data was obtained	Metadata informing about the source database of the information.
Enigma Serial ID	String	Auto-incremental id of the table in the system.	Auto-incremental id for each column. No real user requirement.

Appendix B – Data Quality Issues

Note: Columns which will be removed because of redundancy (see Part A.3) are not considered in this table.

Column Name	Data Error	Solution
License ID	None.	None.
Facility ID	98% missing entries. Entries show that if Facility ID exists, it's the last 4 digits of the License ID.	Remove column.
Call Sign	0.1% empty strings.	Replace with NULL.
FRN	9.5% missing entries. Of those 9.5%, 14% of licenses are still active.	Since we shouldn't remove rows with active licenses, we take no action.
License Holder Name	None.	None.
Common Name	None.	None.
Radio Service	0.1% empty strings.	Replace with NULL.
Rollup Category Desc.	None.	None.
License Status	18 categories, but four of them (active, expired, cancelled and unknown) make up over 95% of rows.	All other entries should be replaced by one of the four. Since it is unclear how to do that, no action will be taken.
Rollup Status	None.	None.
Entity Type	0.1% data is empty strings.	Replace with NULL.
Rollup Entity	None.	None.
License Address	None.	None.
License City	None.	None.
License State	71 distinct states. 3% of rows correspond to two-letter codes which aren't US states.	Since it is unclear how to fix this issue, no action will be taken.
License Zip Code	1.7% missing values.	None.
License Attention Line	22.4% empty strings.	Replace with NULL.
Company Contact	33.9% empty strings	Replace with NULL.
Contact Name	56.6% empty strings.	Replace with NULL.
Contact Title	97.7% empty strings.	Replace with NULL.
Contact Address 1	37.6% empty strings.	Replace with NULL.
Contact Address 2	98.2% empty strings.	Replace with NULL.
Contact City	33.8% empty strings.	Replace with NULL.
Contact State	33.8% empty strings. Two-letter codes which aren't US states.	Replace empty strings with NULL. Since it is unclear how to fix this other issue, no action will be taken.
Contact Zip	33.8% empty strings.	None.
Contact Country	97.7% empty strings. Unclean data, e.g. data mixed with zip code.	Replace all data that doesn't correspond to a country with US, fill all empty strings with US.
Contact Phone	38.4% missing values. Some numbers are mentioned	Since it is unclear how to fix this issue, no action will be taken. Potentially it

	thousands of times, e.g. 9493488510 is mentioned 169,750 times.	isn't even an issue, if it turns out one entity corresponds to this many rows in the dataset.
Contact Fax	45% missing values. Same problem as above, some fax numbers appear thousands of times.	As above.
Contact Email	53.8% empty strings.	Replace with NULL.
Grant Date	100% missing values.	None.
Expiration Date	100% missing values.	None.
Cancellation Date	100% missing values.	None.
Last Action Date	100% missing values.	None.
Market	96.7% empty strings.	None.
Channel Block	97.1% empty strings. There should only be two categories "A" and "B", some entries don't correspond to those categories.	Since it is unclear how to fix this issue, no action will be taken.
Loc. Type Description	18.85% empty strings	Replace with NULL.
Loc. City	34.9% empty strings.	Replace with NULL.
Loc. County Name	33.2% empty strings.	Replace with NULL.
Loc. State	27.7% of entries are empty strings. Some entries don't correspond to US states.	Replace with NULL. Since it is unclear how to fix the other issue, no action will be taken.
Loc. Radius Operations (Km)	80.5% missing entries.	None.
Loc. Sequence ID	Location ID corresponding to the Sequence ID is missing.	Remove column and create ID from scratch.
Loc. Latitude Degrees	32.2% missing entries.	None.
Loc. Latitude Minutes	32.2% missing entries.	None.
Loc. Latitude Seconds	32.2% missing entries.	None.
Loc. Latitude Direction	32.0% of entries are empty strings. Some entries do not correspond to north or south.	Empty entries should be replaced with NULL. Entries not corresponding to "N" or "S" should be replaced by "N", since that is the dominant category.
Loc. Longitude Degrees	32.2% missing entries.	None.
Loc. Longitude Minutes	32.2% missing entries.	None.
Loc. Longitude Seconds	32.2% missing entries.	None.
Loc. Longitude Direction	32.0% of entries are empty strings. Some entries do not correspond to east or west.	Empty entries should be replaced with NULL. Entries not corresponding to "W" or "E" should be replaced by "W", since that is the dominant category.
Structure Height	46.3% missing entries.	None.

ASR Number	71.1% of entries are empty strings.	Replace with NULL.
Antenna ID	16.1% missing values.	None.
Antenna Sequence ID	16.4% missing entries. Entries which are not part of a sequence, usually they are the last digits of the ID.	Replace missing entries with 1 if there is a corresponding Antenna ID. If there is no Antenna ID, it means no antenna exists, so this field can remain NULL. Replace entries 3 digits or more with 1 if the count of the ID is smaller than 100.
Antenna Make	90.0% empty strings.	Replace with NULL.
Antenna Model	90.1% empty strings.	Replace with NULL.
Antenna Type	26.6% empty strings.	Replace with NULL.
Azimuth	87.5% missing entries.	None.
Beamwidth	87.8% missing entries.	None.
Polarization Code	86.6% of entries are empty strings. The rest of the data is very unclean. Most of the data is a one-letter code, but more codes seem to exist than can be found in the FCC documentation. Additionally, some entries are numbers indicating the polarization angle. Last, some entries are alphanumeric codes with more than one letter.	Best solution: since there are only 64 distinct entries, it is possible to build a table specifying how to replace each instance of a data error. Efficient solution: Since the codes should be H (horizontal), V (vertical), R (45° clockwise) and L (45° counterclockwise), with H supposed to be the most common, all entries that are not one of these four can be replaced with the code H, as long as there is an active antenna ID. Rest of entries should be replaced with NULL.
Frequency ID	15.4% missing entries.	None.
Frequency Sequence ID	17.3% missing entries. Entries which are not part of a sequence, usually they are the last digits of the ID.	Replace missing entries with 1 if a corresponding Frequency ID exists. If not, it means no frequency was assigned and this field can stay NULL. Replace entries 3 digits or more with 1 if the count of the ID is smaller than 100.
Frequency Class Station Code	19% of entries are empty strings.	None.
Frequency Class Station	66.2% of entries are empty strings.	Infer description from the code if possible, else replace with NULL.
ERP (Watts)	32.8% missing entries.	None.
Output (Watts)	31.8% missing entries.	None.
Frequency Assigned	15.4% missing entries.	None.
Frequency Upper Band	92.3% missing entries.	None.
Unit of Measure	15.4% empty strings.	Find a method to infer the unit of measure from the frequency.
Tolerance	90.8% missing entries.	None.
Emission ID	19.7% missing entries.	None.

Emission Sequence ID	20.1% missing entries. Entries which are not part of a sequence, usually they are the last digits of the ID.	Replace missing entries with 1. Replace entries 3 digits or more with 1 if the count of the ID is smaller than 100.
Emission Code	19.7% empty strings.	Replace with NULL.
Ground Elevation (Feet)	41.6% missing entries.	Some types of services don't require an antenna (e.g. Wireless). Therefore, no action taken.
Source System	None.	None.