Sagarkumar .P. Vaghasia
B00878629

CSCI 5408 – Data Management, Warehousing and Analytics

Assignment 1

Problem 3

Report: Ocean Tracking Network

There are total 8 datasets in Ocean Tracking Network. They are:

1. Animals
2. Datacenter
3. Detections
4. Manmade_platform
5. Project_attributes
6. Receivers
7. Recover_offload_details
8. Tag_realeases

1. Animals

The animals dataset contains the data for the animals with small electronic transmitters that are surgically implanted or attached externally.

Attributes:

- animal_project_reference
- datacenter_reference: **Foreign Key**
- animal_reference_id: **Primary Key**
- animal_guid
- vernacularname
- scientificname
- taxonrank
- aphiaid
- tsn
- animal_origin
- stock
- length
- length_type
- weight
- life_stage
- age
- sex

2. Datacenter

The datacenter dataset contains data related to various datacenters such as datacenter names, citation, organization, info url and many more.

Attributes:

- datacenter_reference: **Primary Key**
- datacenter_name
- datacenter_abstract
- datacenter_citation
- datacenter_pi
- datacenter_pi_organization
- datacenter_pi_contact
- datacenter_infourl
- datacenter_keywords
- datacenter_keywords_vocabulary
- datacenter_doi
- datacenter_license
- datacenter_distribution_statement
- datacenter_date_modified
- datacenter_geospatial_lon_min
- datacenter_geospatial_lon_max
- datacenter_geospatial_lat_min
- datacenter_geospatial_lat_max
- time_coverage_start
- time_coverage_end

3. Detections

The detections dataset contains data about different detections of ocean animals which is received by transmitters.

Attributes:

- detection_project_reference
- datacenter_reference: **Foreign Key**
- detection_id: **Primary Key**
- detection_guid
- time
- latitude
- longitude
- tracker_reference
- detection_reference_id
- detection_reference_type
- transmitter_codespace
- transmitter_id

- detection_transmittername
- detection_serial_number
- sensor_data
- sensor_data_units
- receiver_log_id
- deployment_id
- detection_quality
- depth
- position_data_source
- uncertainty_in_latitude
- uncertainty_in_longitude
- depth_data_source
- uncertainty_in_depth
- other_position_data
- dataset_quality

4. Manmade_platform

The manmade_platform dataset contains data about various platforms which is obtained by various receivers placed in the ocean.

- platform_project_reference
- datacenter_reference: **Foreign Key**
- platform_reference_id: **Primary Key**
- platform_guid
- platform_type
- platform_depth
- platform_name
- latitude
- longitude

5. Project_attributes

The project_attributes dataset contains data about various projects as there are many projects currently running whose details are obtained by various transmitters.

- project_reference: **Primary Key**
- datacenter_reference: **Foreign Key**
- project_name
- project_abstract
- project_citation
- project_pi
- project_pi_organization
- project_pi_contact
- project_infourl

- project_keywords
- project_keywords_vocabulary
- project_references
- project_doi
- project_license
- project_distribution_statement
- project_date_modified
- project_datum
- project_geospatial_lon_min
- project_geospatial_lon_max
- project_geospatial_lat_min
- project_geospatial_lat_max
- project_linestring
- geospatial_vertical_min
- geospatial_vertical_max
- geospatial_vertical_positive
- time_coverage_start
- time_coverage_end

6. Receivers

The receivers dataset contains data about various receivers that are placed in the ocean for getting signals.

- deployment_project_reference
- datacenter_reference: **Foreign Key**
- deployment_id: **Primary Key**
- deployment_guid
- receiver_manufacturer
- receiver_model
- frequencies_monitored
- receiver_coding_scheme
- receiver_serial_number
- latitude
- longitude
- time
- recovery_datetime_utc
- array_name
- receiver_reference_type
- receiver_reference_id
- bottom_depth
- depth
- deployment_comments
- deployed_by
- expected_receiver_life

7. **Recover_offload_details**

The recover_offload_details dataset contains data about recovery.

- recovery_project_reference
- datacenter_reference: **Foreign Key**
- recovery_id: **Primary Key**
- deployment_id
- recovery_guid
- recovery_latitude
- recovery_longitude
- recovery_datetime_utc
- recovery_outcome
- data_offloaded
- offload_datetime_utc
- log_filenames
- recovery_comments
- clock_synchronized
- recovered_by

8. **Tag_releases**

Scientists and researchers use transmitters to get details and location of animals beneath water. Those transmitters pass unique code which is stored in receivers known as tag releases. So, data related to those tag are stored in tag_releases.

- release_project_reference
- datacenter_reference: **Foreign Key**
- tag_device_id
- release_guid
- release_reference_id: **Primary Key**
- release_reference_type
- latitude
- longitude
- time
- expected_enddate
- manufacturer
- tag_model
- tag_serial_number
- tag_frequency
- tag_coding_system
- transmitted_id
- transmittername
- transmitter_type
- tag_programming_id

Transformation and Cleaning

1. **Animals**
   - "taxonrank" column has NULL values. So, I removed taxonrank column.
   - In "sex" column, out of 3810 records there were 104 female animals, 41 male animals, 342 as unsexed and 3323 empty. Therefore, those empty records were filled with U: unsexed having highest occurrence.
   - In "age" column, there are only two age values defined. Age 1 of having 100 records and age 14 has 50 records and remaining 3660 values are not a number. This column is of no use as age cannot be 1 or 14 and majority of the records are blank. So, I removed age column.
   - In "life_stage" column, amongst all 4 stages highest and major occurrence is of juvenile so I replaced empty records with juvenile.
   - In "weight" column, the weight is ranging from 0 to 40.46 and there are 736 records having NaN value. As the weight having this many range cannot be average so the I replaced NaN values with 0.
   - In "length_type" column, the maximum occurrence is of Fork. So, I filled 115 blank records with Fork value.
   - In "length" column there were 118 records having value NaN. So, I filled those records with mean value as the column value ranges from 0.104 to 1.8542.
   - In "stock" column, there were 163 records which are blank. So, I assigned UNKNOWN value to those records we don't know the value of it.
   - In "animal_origin" column there are 3 values: W, H, U. The H is having the highest occurrence amongst all three. Thus, I replaced 12 blank records by value H.
   - By observing data, it can be said that column "animal_guide" is a mixture of 3 columns: datacenter_reference, animal_project_reference, animal_reference_id. This column is of no use, so I removed animal_guide column.
   - Shifted animal_reference_id column to the first column as it will become primary key.

2. **Datacenter**
   - There are 4 columns ("time_coverage_start","time_covergae_end"," datacenter_distribution_statement"," datacenter_date_modified") which do not have any values. So, I removed those 4 columns.
   - There is row having NaN value in 4 columns ("datacenter_geospatial_lon_min", "datacenter_geospatial_lon_max", "datacenter_geospatial_lat_min", "datacenter_geospatial_lat_max") which is replaced by 0.
   - In "datacenter_license" and "datacenter_abstract", to increase the readability I removed Â¦ symbol and extra spaces.

3. Detections

- "dataset_quality", "other_position_data", "uncertainty_in_depth", "depth_data_source" and "receiver_log_id" columns has NULL values. So, I removed those columns.
- By observing data, it can be said that column "detection_guid" is a mixture of 3 columns: datacenter_reference, detection_project_reference, detection_id. This column is of no use, so I removed detection_guid column.
- "detection_transmittername" column is the combination of "transmitter_codespace" and "transmitter_id". So, I deleted detection_transmittername column.
- In "sensor_data" column, all the empty cells are replaced by the median of the available values in the cells (14.06).
- Columns "uncertainty_in_longitude", "uncertainty_in_latitude" and "depth" are having NaN value, or they are empty. So, they are of no use. I removed both columns.
- In "sensor_data_unit" column, there are 3 units given having certain records and the records which are blank are assigned as UNKNOWN as we don't know the unit of that value.
- In "detection_quality" column, there are 5054 records having Found Receiver value. I set Receiver not found in the rest of the empty cells.
- Shifted detection_id column to the first column as it will become primary key.

4. Manmade_platform

- Columns "platform_reference_id" and "platform_name" are having same values. So, I must remove any one of those columns. I removed "platform_name" because "platform_reference_id" might become primary key in future.
- By observing data, it can be said that column "platform_guid" is a mixture of 3 columns: datacenter_reference, platform_project_reference, platform_reference_id. This column is of no use, so I removed detection_guid column.
- In "platform_depth" column, 2261 records are having NaN value which I replaced by mean value of the other elements.
- In "latitude" and "longitude" columns there are few values which are NaN. I replaced those values by mean value of the column.
- Shifted platform_reference_id column to the first column as it will become primary key.

5. Project

- Columns "project_references", "project_doi", "project_distribution_statement", "project_date_modified", "project_linestring", "geospatial_vertical_positive", "time_coverage_start" and "time_coverage_end" are empty. So, I deleted those columns.

- In columns “project_abstract”, “project_citation”, “project_pi_contact” and “project_infourl” there are few cells which are empty. So, I replaced those with UNKNOWN.
- In columns “geospatial_vertical_min” and “geospatial_vertical_max”, there are few blank values which I replaced by 0 for identification.

6. **Receivers:**
   - Columns “frequencies_monitored”, “receiver_coding_scheme”, “deployed_by” and “expected_receiver_life” are empty. So, I deleted those columns.
   - Column “deployment_comments” has empty cells and number values which I replaced with “NOT GIVEN” as comments can’t be numbers.
   - By observing data, it can be said that column “deployment_guid” is a mixture of 3 columns: datacenter_reference, deployment_project_reference, deployment_id. This column is of no use, so I removed deployment_guid column.
   - Columns “depth” and “bottom_depth” has many NaN values. So, I replaced those by mean value of the column.
   - Column “receiver_reference_type” has only ManmadePlatform. So, this column is of no use. That’s I deleted that column.
   - Column “receiver_manufacturer” has many empty cells, which I filled with VEMCO as it has highest occurrence.
   - Shifted deployment_id column to the first column as it will become primary key.

7. **Recover_Offload_details:**
   - Columns “clock_synchronized”, “recovered_by” are empty. So, I deleted those columns.
   - Column “recovery_comments” has many empty cells which I filled with NO COMMENTS.
   - By observing data, it can be said that column “recovery_guid” is a mixture of 3 columns: datacenter_reference, deployment_id, recovery_id. This column is of no use, so I removed recovery_guid column.
   - Columns “log_filenames”, “offload_datetime”, “recovery_datetime” have many empty cells which I set to UNKNOWN.
   - Columns “deployment_id” and “recovery_id” are same columns having duplicate values. So, I removed deployment_id.
   - Shifted recovery_id column to the first column as it will become primary key.

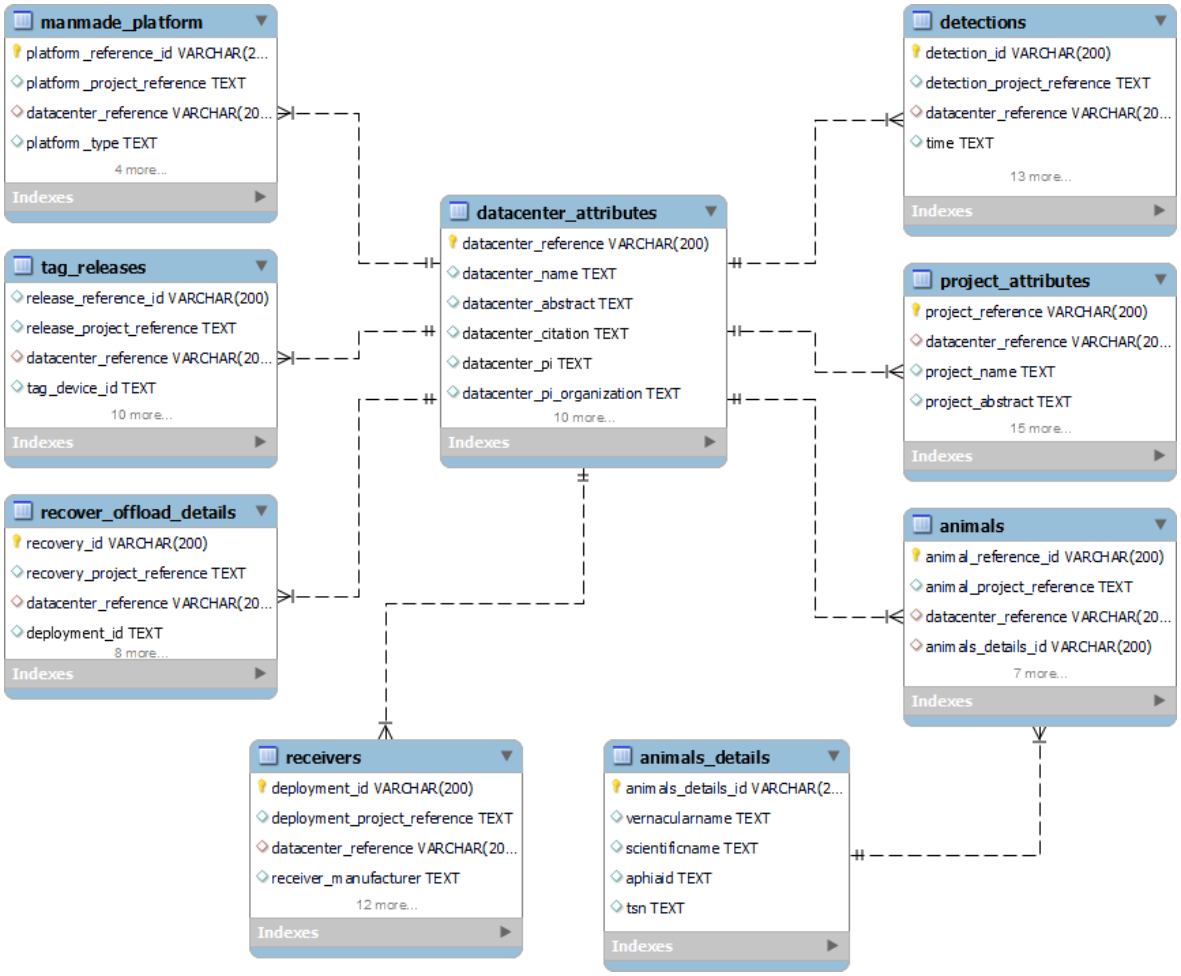
8. **Tag_Releases:**
   - Columns “transmitter_type”, “tag_programming_id”, “tag_frequency” are empty. So, I deleted those columns.
   - By observing data, it can be said that column “release_guid” is a mixture of 3 columns: datacenter_reference, release_project_reference, tag_device_id. This column is of no use, so I removed release_guid column.

- Column “transmittername” is a combination of “tag_coding_system” and “transmitted_id”. So, I removed that column.
- Shifted release_reference_id column to the first column as it will become primary key.

Normalization / Denormalization:

- To increase efficiency, I normalized animals.csv file into two different files as there were 4 columns (vernacularname, scientificname, aphiaid, tsn) whose values will change if any one value is changed. So, I separated those 4 columns and made another file named as animals_details.csv having one primary key column which is linked in animals.csv.

After cleaning and normalization of data, I have created schema in MySQL workbench named as “Ocean Tracking Network” where I imported all the data from the .csv files. I defined primary keys and foreign keys in all tables. When I was defining primary keys at that time in “manmade_platform” table MySQL showed error due to duplicate records in the field of primary key. So, I removed the duplicate rows from that table. Lastly, by applying reverse engineering I generated ERD. I have attached the ERD below:



References :

[1] Ocean Tracking Network : https://oceantrackingnetwork.org/about/#oceanmonitoring

"I Sagarkumar .P. Vaghasia, declare that in assignment 1 of CSCI 5408 course, data scrapping is not done programmatically or using any online or offline tools. However, the webpages or the domain mentioned in this document are visited manually, and some useful information is gathered for education purpose only. Information, such as email, personal contact numbers, or names of people are not extracted. The course instructor or the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data".