ASSIGNMENT 1
ANSWER 2

CSCI 5408 – DATA MANAGEMENT, WAREHOUSING & ANALYTICS

PRASHIT PATEL

B00896717

Ocean Tracking Network (OTN) Datasets and Attributes Report:

OTN collects data with the help of technology and stores data in different datasets as follows:

- **Project-Attributes** – It captures data about different projects carried out under OTN. It contains attributes such as project_reference for unique project identification, datacenter_reference for specifying the datacenter used for the project, project details such as name, abstract, citation, principal investigator details like name, organization, contact, project URL, license, project keywords, project vocabulary, project references, project doi, license, project distribution statement, date modified, project datum, project geospatial lon min, project geospatial lon max, project geospatial lat min, project geospatial lat max, project line string, geo spatial vertical min, geo spatial vertical max, geo spatial vertical positive, time coverage start, time coverage end.

- **Datacenter-Attributes –** It captures data about different datacenters used by OTN to store data. It contains attributes such as datacenter_reference for unique datacenter identification, name, abstract, citation, principal investigator name, organization and contact information, URL, datacenter keywords, datacenter vocabulary, digital object identifier (DOI), license, statement, date modified, geospatial lon max, geospatial lon min, geospatial lat min, geospatial lat max and time coverage start and time coverage end.

- **Receivers –** It captures data about receivers placed by OTN at different locations. It contains attributes project_reference and datacenter_reference for mentioning project to which receiver is assigned to and datacenter used by that receiver respectively. Other receiver attributes are deployment id and deployment global id for unique identification, device information attributes such as manufacturer, model, frequencies monitored, coding scheme and serial number, latitude, longitude, time, recovery datetime, array name, receiver reference type and id, bottom depth, depth, deployment comments, deployed by and expected receiver life.

- **Detections –** It captures detection information and contains attributes project_reference and datacenter_reference for mentioning project to which detection is related to and datacenter used for that detection respectively. Other detection attributes are detection id and detection global id for unique identification, detection time, latitude, longitude, tracker reference, detection reference id and type, transmitter name, id, codespace, detection serial number, sensor data and units, receiver log id, deployment id to capture receiver information, detection quality, depth, position data source, uncertainty in latitude and longitude, depth data source, uncertainty in depth, other position data and dataset quality.

- **Animals –** It captures information about animals. It contains attributes such as project reference and datacenter_reference for linking animal to project and datacenter respectively. Other animal attributes are animal id and animal global id for unique identification, vernacular name, scientific name, taxonomy rank, aphia id, taxonomy serial number (TSN), animal origin, stock, length, length type, life stage, weight, age and sex.

- **Tag releases –** It captures tag release information and contains attributes such as project reference and datacenter_reference for linking tags to project and datacenter respectively. Other tag_release attributes are tag device id, release global id for unique identification, release reference type and its id, tag locations - latitude and longitude, time and expected end data, tag manufacturer, model, serial number, frequency, coding system, transmitter id, name and type and tag programming id.

- **Recover offload details –** It captures offloaded information from the receivers and contains attributes such as project reference and datacenter_reference for linking recovery details to project and datacenter respectively. Other recover offload attributes are recovery id and global id for unique identification, deployment id of respective receiver, recovery location - latitude and longitude, recovery date and time, recovery outcome, data offloaded flag, offloaded date time, log filenames, recovery comments, clock synchronized and recovered by.

- **Manmade platform –** It captures manmade platform related information and contains attributes such as project reference and datacenter_reference for linking platforms to project and datacenter respectively. Other manmade platform attributes are platform id and global id for unique identification, type, depth, name, and location - latitude and longitude.

Cleaning and Transforming Data:

1. Columns with all null values were removed from datasets.
2. For columns with data type as text, null values were replaced with NA.
3. For columns with data type as date, null values were replaced with 0001-00-00T00:00:00Z which is the first date for date data type.
4. For columns with data type as float (only positive values), null values were replaced with -1.
5. For columns with data type as float (positive & negative values except 0), null values were replaced with 0.
6. For columns containing latitude and longitude information, null values were replaced with 1000 (1000 value out of range for latitude and longitude).

Normalization:

1. Project_attributes dataset contained multiple values for project_citation attribute, so new dataset is created named project_citation containing project_reference as primary key and author & reference as attribute which will contain all the values previously present in project_citation, thus removing multiple values in project_citation attribute (1NF).

2. Tag_releases dataset contains partial dependency where manufacturer, tag_model, tag_serial_number and tag_coding_system depends on tag_device_id only and not release id and release global id, thus partially dependent. A new dataset named tag_details containing tag information such as manufacturer, model, serial number, coding system is created with tag_device_id as primary key and thus removing partial dependency (2NF).

3. Receivers dataset contains partial dependency where manufacturer, model, receiver_serial_number and receiver_reference_type depends on receiver_reference_id only and not deployment id and deployment global id, thus partially dependent. A new dataset named receiver_details containing receiver information such as manufacturer, model, serial number, and reference type is created with receiver_reference_id as primary key and thus removing partial dependency (2NF).

EERD using reverse engineering in MySql workbench