**Phase 3: Database Creation, Table Population, and Business questions**

*Author: Shivani Tayade*

***DATA Conversion:***

The dataset provided contained various pieces of information about museums, their locations, financial details, and their administrative details. To make it manageable and ensure data integrity, I broke down this large dataset into smaller, more focused tables. This process is called **normalization**, and it helps avoid repetition and ensures that each piece of data is stored in the most efficient way.

I have created separate tables as follows-

* A table for basic museum details like MUSEUM\_INFORMATION (museum ID, name, legal name, etc.).
* A separate table for INSTITUTION\_INFORMATION to store institution names associated with each museum.
* Tables like ADMINISTRATIVE\_LOCATION and PHYSICAL\_LOCATION were created to store information related to museum locations (both administrative and physical addresses).
* Contact-related details, such as phone numbers and region codes, were stored in a CONTACT\_AND\_CODES table.
* Financial data, such as income and revenue, was placed in the FINANCIAL\_INFORMATION table.

Each table was connected via a museum\_id field, ensuring that all information about a specific museum could be linked together. This makes querying the data more efficient and helps maintain consistency across the dataset.

***Challenges faced during importing of your data and how did you overcome these data importation challenges:***

While importing the data into MySQL, several challenges came up. Here's how they were addressed:

1. **Data Format Issues** Some fields like Phone number and zipcode were inconsistent, with numeric fields containing non-numeric values (blanks
   1. **Solution**: I adjusted the data types in the tables. For example, Zip Code and Phone Number were stored as VARCHAR to accommodate alphanumeric values.
2. **Large Data Volumes:** The dataset was sizable, which posed a challenge when loading it into MySQL in one go.
   1. Solution: I used LOAD DATA INFILE, which allows loading data in chunks, thus reducing the chances of running out of memory or encountering timeout errors.
   2. Solution: I firstly tried to fetch the data in Sql Workbench by importing the data through the ‘Import wizard’. In this process the data was getting truncated and were not correct. Refer to the below screenshot for the reference:

A screenshot of a phone number

Description automatically generated

Given this, the next attempt was to import the data through the command ‘infile’ in sql where the sql can be fetched and then later tables can be created. In doing so, certain permission and encoding has to be changed to allow the importing data process. Please refer to the SQL code/script for more reference. For an example, the changes I made is to include the ‘opt\_local\_infile =1’ in the coonections methods and setting the privacy setting as ‘Off’. If the privacy setting is kept ‘ON’ then paste the raw data csv file in the mentioned sql directory. The below small piece of code to check the network accessibility.

“ SHOW VARIABLES LIKE 'secure\_file\_priv';

SHOW VARIABLES LIKE 'local\_infile';

SELECT VERSION();

SHOW VARIABLES; “

The below snapshot denotes the changes in setting to get the data imported-

*Before After*

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

1. **Missing Data:** Some crucial fields, such as phone numbers or latitude/longitude, were missing in certain rows, leading to errors during import.
   1. Solution: I cleaned the data using NULLIF() to convert empty strings into NULL values where necessary. This helped ensure the data was consistent and prevented errors during import.
2. **Foreign Key Issues:** Some tables, like INSTITUTION\_INFORMATION, had foreign keys referencing other tables. If there was an issue with the order of insertion or data consistency, foreign key constraints would cause errors.
   1. Solution: I ensured the tables were created in the correct order (parent tables first) and that museum\_id values were consistent across all tables before inserting data.

***A complete data dictionary for every table in your database:***

I have created a comprehensive Data Dictionary for all the tables in the database, detailing the structure of each table, the columns, data types, and descriptions of each field. This dictionary helps to understand the data model and provides context for each table’s role in the overall database. Please refer to the attached excel file.

****

***The list of business questions:***

Here are business questions that you can easily answer using the dataset. These questions are designed to uncover valuable insights that could help museum managers, administrators, and decision-makers improve operations, identify trends, and drive strategic initiatives.

1. Which museums are generating the highest revenue? *We want to know which museums are performing best financially. This insight will help allocate resources and prioritize support for high-performing museums.*
2. What is the total income generated by museums in each state? *By aggregating income across museums in each state, we can identify regional financial performance and perhaps optimize marketing efforts or funding allocation.*
3. Which museums are located in the most cities? *This helps identify which museums have a broader geographic reach or presence*
4. Do any museums lack financial details (income or revenue)? *This question helps pinpoint museums that may have incomplete or missing financial data, which could indicate areas for improvement in record-keeping or reporting.*
5. Which museums are located in the same city and state as others? *This question can highlight clusters of museums in specific cities or states.*
6. What is the average revenue of the top 3 museums by revenue? *Understanding average revenues by region can help tailor strategies for museum management, funding, and marketing specific to each area.*
7. What is the cumulative revenue for each museum by its rank? *To calculate a running total of revenue for each museum*
8. How does the revenue of each museum compare to the overall average revenue? *To find how each museum's revenue compares to the overall average revenue, with the difference calculated.*
9. What is the difference in revenue between each museum and the previous one (ordered by revenue)? *This will help understand significantly outperforming or underperforming museums compared to their peers.*
10. Which museum has the highest latitude and longitude?
11. What is the average income of museums across all states?
12. Counts how many museums are missing phone numbers?
13. State how many museums have same Admin and Physical addresses?
14. What is the average income by states for these museums?