Hadoop 5: Assignment 2

Problem Statement:

Imagine that you are working with the Government of India. The government is keen on using Big Data to analyze the rate of suicides in the country and to draw some inference from that. You, being a Data Analyst with expertise on Hive, have been chosen to complete the task.

Link for the dataset and resources: https://intellipaat-course-attachments.s3.ap-south-1.amazonaws.com/Hadoop/Hadoop+Datasets-20200609T120700Z-001.zip

Dataset Description:

The **Suicides.xls** file is the source file.
The dataset consists of seven columns and their respective description as follows:

State: The state where the suicide occurred
Year: Year of death
Type code: One of the following codes representing a reason: Causes, Education Status, Means Adopted, Professional Profile, and Social Status
Type: Sub-category of type code
Gender
Age group
Total: Total cases in the particular age group

Queries to Be Performed:

Participants can use Hive shell to explore the problem and find the solution.
Connect with Hive shell and perform the following analysis:

1. Create a database called Demo and use it

2. Create a table called Suicides in it, matching with the schema of the data
3. Load the given CSV file into the table
4. Find out the most common suicide cause among females in India over the entire period 2001–2012
5. Find out the state-wise most common cause among males over the entire period
6. Find out the age group-wise most common cause among males and females
7. Find out the total number of suicides per year per state