Hadoop: Spark Assignment - 2

Problem Statement:

Imagine that you are working as an analyst for a retail firm. Your firm collects sensor data from all over the world. You need to analyze the same and provide insights to customers. Since you are collecting huge amounts of sensor data, the analysis platform is chosen as Spark SQL.

Dataset: : :

https://intellipaat-course-attachments.s3.ap-south1.amazonaws.com/Hadoop/Hadoop+Datasets-20200609T120700Z-001.zip

Dataset Description:

The iot_devices.json file is the source file. The dataset has the following attributes:

- Device ID
- Device Name
- IP Address
- Cca2 – Country Code
- Cca3 – Country Name
- Cn – Full Name of the Country
- Latitude
- Longitude
- Scale
- Temperature
- Humidity
- Battery Level
- CO2 Level
- LCD Status
- Timestamp

Tasks 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.