**Problem Statement: Hive Tables, Partitions, Buckets, and Joins**

**Background:**

A retail company tracks customer purchases. The data is maintained in two datasets:

1. customers – basic customer information
2. transactions – customer purchases over time

You are tasked with:

* Creating Hive tables with appropriate **partitioning and bucketing** strategies
* Populating them with sample data
* Performing **joins** and querying **specific insights**

**customers.csv**

cust\_id, name, state

1, Alice, Karnataka

2, Bob, Tamil Nadu

3, Charlie, Karnataka

4, David, Kerala

5, Eva, Tamil Nadu

**transactions.csv**

txn\_id, cust\_id, amount, txn\_date

1001, 1, 250.75, 2023-06-01

1002, 2, 125.00, 2023-06-01

1003, 1, 75.25, 2023-06-02

1004, 3, 400.00, 2023-06-03

1005, 4, 200.00, 2023-06-04

1006, 5, 500.00, 2023-06-04

### ✅ ****Tasks****

1. **Create Hive Table: Customers**
   * External table
   * Store as **TEXTFILE**
   * Load data from customers.csv
2. **Create Hive Table: Transactions**
   * Partitioned by txn\_date
   * Bucketed by cust\_id into **4 buckets**
   * Stored as **ORC**
3. **Load Data**
   * Load both datasets into their respective Hive tables
4. **Queries**

a. Find total transaction amount per customer  
b. List all transactions for customers in "Karnataka"  
c. What is the total amount transacted on each date?  
d. Identify customers who have done more than one transaction  
e. Verify bucketing: Get the bucket number for each transaction