Assignment 3: Relational Join with MapReduce Simulation

Topic: Algorithms in MapReduce (Lecture 10)
Time: 30-40 minutes
Objective: Perform a relational join between two datasets using a MapReduce-like process.

Problem Statement:
You are given two small datasets: an "Employee" table (Name, SSN) and an "Assigned Departments" table (EmpSSN, DepName). Simulate a MapReduce process to join these tables on SSN = EmpSSN and output the joined results (Name, SSN, DepName).

Dataset:
Provided as Python lists in the notebook:

python
CollapseWrapCopy

```
employees = [("Sue", "999999999"), ("Tony", "777777777")]
departments = [("999999999", "Accounts"), ("777777777", "Sales"),
("777777777", "Marketing")]
```

Requirements:

- Implement map_function to emit (SSN, (table_type, data)) pairs.
- Implement reduce_function to join matching records.
- Output the joined results as a list of tuples.
- Use pandas to display the results as a DataFrame.

Solution Outline:

1. map_function: For each tuple in both datasets, emit (SSN, (table_type, tuple_data)).
2. Group mapped data by SSN (simulate shuffle).
3. reduce_function: For each SSN, combine Employee and Department data where SSN matches.
4. Collect and display results.

Sample Code Starter:

python

```
import pandas as pd


employees = [("Sue", "999999999"), ("Tony", "777777777")]
departments = [("999999999", "Accounts"), ("777777777", "Sales"),
("777777777", "Marketing")]
```

```
def map_function(record, table_type):
    # Students implement
    pass

def reduce_function(key, values):
    # Students implement
    pass

# Students complete the rest
```