Task 4 (40%). Distributed Join over Apache Spark
Write a Spark application which provides an efficient parallel implementation for the following SQL query:

```
SELECT
     C_CUSTKEY,
     O_COMMENT
FROM
     CUSTOMER join ORDERS on C_CUSTKEY = O_CUSTKEY
```

where relations `CUSTOMER` and `ORDERS` have the following schema:

```
CUSTOMER(C_CUSTKEY     INTEGER NOT NULL,
         C_NAME        VARCHAR(25) NOT NULL,
         C_ADDRESS     VARCHAR(40) NOT NULL,
         C_NATIONKEY   INTEGER NOT NULL,
         C_PHONE       CHAR(15) NOT NULL,
         C_ACCTBAL     DECIMAL(15,2)   NOT NULL,
         C_MKTSEGMENT  CHAR(10) NOT NULL,
         C_COMMENT     VARCHAR(117) NOT NULL)

ORDERS(O_ORDERKEY       INTEGER NOT NULL,
       O_CUSTKEY        INTEGER NOT NULL,
       O_ORDERSTATUS    CHAR(1) NOT NULL,
       O_TOTALPRICE     DECIMAL(15,2) NOT NULL,
       O_ORDERDATE      DATE NOT NULL,
       O_ORDERPRIORITY  CHAR(15) NOT NULL,
       O_CLERK          CHAR(15) NOT NULL,
       O_SHIPPRIORITY   INTEGER NOT NULL,
       O_COMMENT        VARCHAR(79) NOT NULL)
```

C_CUSTKEY and O_ORDERKEY are primary keys and O_CUSTKEY is a foreign key referencing C_CUSTKEY.

The goal of this application is for you to implement Distributed Join.
In your solution, you are **NOT** allowed to use:

- SparkSQL
- Dataframes
- join operation of RDD
- Third party libraries

You are required to implement the queries using standard Spark RDD operations (excluding join).

Output: The format of the result must be CSV (separated by ‘,’).

Data: "customer.tbl" (small, big) and "orders.tbl" (small, big), CSV format where the fields of each tuple are separated by '|'. You can use the small file for testing purposes.

Deliverables:

- Task4.java: Your applications should take as arguments:
  - <inputFile>: denoting the input dataset.
  - <outputFile>: denoting the output file.
- Task4.pdf: A one page description of your design and advantages/disadvantages of the approach.

Grading: We will harshly penalize submissions for which the input, output paths, or other parameters are hardcoded or are not abiding with the required format.