Q1. Design MapReduce algorithms to take a very large file of integers and produce the output as same set of integers, but with each integer appearing only once.

Q2. Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items {1, 2, 3, 4, 6, 12}. If the support threshold is 5, which items are frequent?

Q3. For the data of Q2, what is the confidence of the following association rules?

(a) {5, 7} → 2.

(b) {2, 3, 4} → 5.

Q4. If we use a triangular matrix to count pairs, and n, the number of items, is 20, what pair's count is in a[100]?

Q5. Apply the A-Priori Algorithm with support threshold 5 to the data of Q2.

Q6. Here is a collection of twelve baskets. Each contains three of the six items 1 through 6. {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6} {1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5} {3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6} Suppose the support threshold is 4. On the first pass of the PCY Algorithm we use a hash table with 11 buckets, and the set {i, j} is hashed to bucket i × j mod 11.

(a) By any method, compute the support for each item and each pair of items

(b) Which pairs hash to which buckets?

(c) Which buckets are frequent?

(d) Which pairs are counted on the second pass of the PCY Algorithm?