Answer：

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.

A1:

Map: for each integer i in the file, emit key-value pair (i, 1) Reduce: turn the value list into 1. Note the result is obtained from the keys of the output.

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?

Definitions:

Support for itemset I: Number of baskets containing **all items in I**

Given a **support threshold *s***, then sets of items that appear in at least *s* baskets are called **frequent itemsets.**

A2.

Item i is in basket b if and only if i divides b with no remainder.

Example: let b=5, i=2, then the remainder should be 1 since 5÷2=2…1.

The baskets with numbers 20, 40(20×2), 60(20×3), 80(20×4) and 100(20×5) should contains item 20.

Frequent items are: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

Q3. For the data of Q2, what is the confidence of the following association rules? (a) {5, 7} → 2. (b) {2, 3, 4} → 5.

Definitions:

If-then rules about the contents of baskets

$\{i_1, i_2, \ldots, i_k\} \rightarrow j$ means: "if a basket contains all of $i_1, \ldots, i_k$ then it is ***likely*** to contain j"

Confidence of this association rule is the probability of j given $I = \{i_1, \ldots, i_k\}$

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

A3.

(a) {5, 7} → 2

The baskets containing both item 5 and item 7 are basket 35 and basket 70, in which only basket 70 also contains item 2. Hence, the confidence of the association rule is

$$\frac{\text{support}(\{2,5,7\})}{\text{support}(\{5,7\})} = \frac{1}{2}.$$

(b){2, 3, 4} → 5

The baskets whose numbers are the multiples of 12 contain item set {2,3,4} as a subset, there are 8 such baskets, while only those whose numbers are the multiples of 60 contain item set {2,3,4,5} as a subset, there are 1 such basket. Hence, the confidence of the association rule is 1/8.

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]?

Definitions:

Triangular Matrix counts pair of items {i, j} where $1 \leqslant i < j \leqslant n$ in lexicographic order: {1, 2}, {1, 3}, ⋯, {1, n}, {2, 3}, {2, 4}, ⋯, {2, n}, {3, 4}, ⋯, {n-1, n}.

Pair {i, j} is at position [n(n -1) -(n -i +1)(n -i)]/2 + (j -i)

A4. Solve the equation:

[20×(20-1)-(20-i+1)(20-i)]/2 + (j -i) = 100

→ [380-(420-41i+i^2)]/2 + j – i =100

→ -i^2+39i+2j = 240 where $1 \leqslant i < j \leqslant 20$

we can get the pair {7,8}.

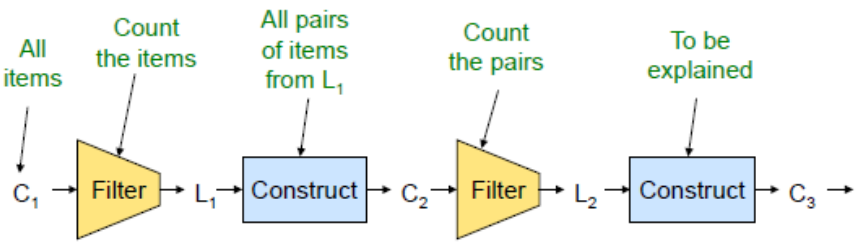
Q5. Apply the A-Priori Algorithm with support threshold 5 to the data of Q2.

Definitions:

A-Priori Algorithm is a two-pass approach with the key idea **monotonicity**.

Monotonicity: If a set of items i appears at least *s* times, so does every subset j of i. In other words, if item i does not appear in at least *s* baskets, then no pair including i can appear in more than *s* baskets.

- **For each *k*, we construct two sets of *k-tuples*** (sets of size *k*):
  - C_k = ***candidate k-tuples*** = those that might be frequent sets (support ≥ s) based on information from the pass for ***k–1***
  - L_k = the set of truly frequent ***k***-tuples



A5. (Based on permutation, many results can be obtained that satisfy the conditions. The answers demonstrated here are just sample results, and there may be missing combinations.)

Candidate itemsets of size 1 (C1): Items with index 1~100

Truly frequent itemsets of size 1 (L1): Items with index 1~20

Candidate itemsets of size 2 (C2): $\forall 1 \leqslant i < j \leqslant 20 \{i,j\}$

Truly frequent itemsets of size 2 (L2): (too many)

Candidate itemsets of size 3 (C3): ∀c, |c| = 3 and ∃ci , cj ∈ $L2$, c = ci ∪ cj

Truly frequent itemsets of size 3 (L3):

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 2, 6}, {1, 2, 7}, {1, 2, 8}, {1, 2, 9}, {1, 2, 10}, {1, 2, 12}, {1, 2, 14}, {1, 2, 16}, {1, 2, 18}, {1, 2, 20}, {1, 3, 4}, {1, 3, 5}, {1, 3, 6}, {1, 3, 9}, {1, 3, 12}, {1, 3, 15}, {1, 3, 18}, {1, 4, 5}, {1, 4, 6}, {1, 4, 8}, {1, 4, 10}, {1, 4, 12}, {1, 4, 16}, {1, 4, 20}, {1, 5, 10}, {1, 5, 15}, {1, 5, 20}, {1, 6, 9}, {1, 6, 12}, {1, 6, 18}, {1, 7, 14}, {1, 8, 16}, {1, 9, 18}, {1, 10, 20}, {2, 3, 4}, {2, 3, 6}, {2, 3, 9}, {2, 3, 12}, {2, 3, 18}, {2, 4, 5}, {2, 4, 6}, {2, 4, 8}, {2, 4, 10}, {2, 4, 12}, {2, 4, 16}, {2, 4, 20}, {2, 5, 10}, {2, 5, 20}, {2, 6, 9}, {2, 6, 12}, {2, 6, 18}, {2, 7, 14}, {2, 8, 16}, {2, 9, 18}, {2, 10, 20},
{3, 4, 12}, {3, 5, 15}, {4, 5, 10}, {4, 5, 20}, {4, 6, 12}, {5, 10, 20}, {6, 9, 18}

Candidate itemsets of size 4 (C4): ∀c, |c| = 4 and ∃ci , cj ∈ $L3$, c = ci ∪ cj

Truly frequent itemsets of size 4 (L4): {1, 2, 3, 4}, {1, 2, 3, 6}, {1, 2, 3, 9}, {1, 2, 3, 12}, {1, 2, 3, 18}, {1, 2, 4, 5}, {1, 2, 4, 6}, {1, 2, 4, 8}, {1, 2, 4, 10}, {1, 2, 4, 12}, {1, 2, 4, 16}, {1, 2, 4, 20}, {1, 2, 5, 10}, {1, 2, 5, 20}, {1, 2, 6, 9}, {1, 2, 6, 12}, {1, 2, 6, 18}, {1, 2, 7, 14}, {1, 2, 8, 16}, {1, 2, 9, 18}, {1, 2, 10, 20}, {1, 3, 4, 12}, {1, 3, 5, 15}, {1, 4, 5, 10}, {1, 4, 5, 20}, {1, 4, 6, 12}, {1, 5, 10, 20}, {1, 6, 9, 18}
{2, 3, 4, 12}, {2, 4, 5, 10}, {2, 4, 5, 20}, {2, 4, 6, 12}, {2, 5, 10, 20}, {2, 6, 9, 18}

{3, 4, 6, 12}, {3, 6, 9, 18}

{4, 5, 10, 20}

Candidate itemsets of size 5 (C5): ∀c, |c| = 5 and ∃ci , $cj \in L4$, c = $ci \cup cj$

Truly frequent itemsets of size 5 (L5):

{1, 2, 3, 4, 6}, {1, 2, 3, 4, 12}, {1, 2, 4, 5, 10}, {1, 2, 4, 5, 20}, {1, 2, 4, 6, 12}, {1, 2, 5, 10, 20}, {1, 2, 6, 9, 18}, {1, 3, 4, 6, 12}, {1, 3, 6, 9, 18}, {1, 4, 5, 10, 20},

{2, 3, 4, 6, 12}, {2, 4, 5, 10, 20}

Candidate itemsets of size 6 (C6): ∀c, |c| = 6 and ∃ci , $cj \in L5$, c = $ci \cup cj$

Truly frequent itemsets of size 6 (L6):

{1, 2, 3, 4, 6, 12}, {1, 2, 3, 6, 9, 18}, {1, 2, 4, 5, 10, 20}

Through combination, no candidate itemset of size 7 could be found, the algorithm stops here.

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?

Definitions:

PCY (Park-Chen-Yu) Algorithm

Pass 1: In addition to item counts, maintain a hash table with as many buckets as fit in memory. Keep a count for each bucket into which pairs of items are hashed

A-priori

```
FOR (each basket) :
    FOR (each item in the basket) :
        add 1 to item's count;
```

New in PCY

```
FOR (each pair of items) :
    hash the pair to a bucket;
    add 1 to the count for that bucket;
```

Pass 2: Only count pairs that hash to frequent buckets

A6.

(a)

item	1	2	3	4	5	6
support	4	6	8	8	6	4

pair	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{2, 3}	{2, 4}	{2, 5}
support	2	3	2	1	0	3	4	2
pair	{2, 6}	{3, 4}	{3, 5}	{3, 6}	{4, 5}	{4, 6}	{5, 6}	
support	1	4	4	2	3	3	2	

(b)

pair	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{2, 3}	{2, 4}	{2, 5}
bucket	2	3	4	5	6	6	8	10
pair	{2, 6}	{3, 4}	{3, 5}	{3, 6}	{4, 5}	{4, 6}	{5, 6}	
bucket	1	1	4	7	9	2	8	

(c)

bucket	0	1	2	3	4	5	6	7	8	9	10
support	0	5	5	3	6	1	3	2	6	3	2

The frequent buckets are those with support above 4, thus 1, 2, 4, 8.

(d) As only pairs in frequent buckets will be counted on the second pass of PCY, they are {1, 2}, {1, 4}, {2, 4}, {2, 6}, {3, 4}, {3, 5}, {4, 6}, {5, 6}