参考答案

备注：目前参考答案内容以英文为主，仅供参考。

第一题

In each part, you can assume that the key of each output pair will be ignored or dropped.

Then, （a）Algorithm to calculate the average of all the integers:

Map: for each integer j in the main file emit the key-value pair (m, j)

Reduce: read task pair key as m and store the associated value in list $[j_1, j_2, \cdots, j_n]$.

Output for the reduce task is a pair of $(m, average\{j_1, j_2, ..., j_n\})$.

(b)Algorithm to get the same set of integers, with each integer appear only once :

Map: for each integer j in the main file emit the key-value pair (m, j)

Reduce: read task pair key as m and store the associated value in list $[j,j]$.

It generates exactly one pair for ***m*** key, *(m, j)*.

第二题

Given items are 1,2,3,...,100. Here, 100 baskets are given such that i divides b with remainder 0. Therefore, all 100 baskets are shown below:

basket 1:{1}
basket 2:{1,2}
basket 3:{1,3}
basket 4:{1,2,4}
basket 5:{1,5}
basket 51:{1,3,17,51}
basket 52:{1,2,4,13,26,52}
basket 53:{1,53}
basket 54:{1,2,3,6,9,18,27,54}
basket 55:{1,5,11,55}

basket 6:{1,2,3,6}
basket 7:{1,7}
basket 8:{1,2,4,8}
basket 9:{1,3,9}
basket 10:{1,2,5,10}
basket 11:{1,11}
basket 12:{1,2,3,4,6,12}
basket 13:{1,13}
basket 14:{1,2,7,14}
basket 15:{1,3,5,15}
basket 16:{1,2,4,8,16}
basket 17:{1,17}
basket 18:{1,2,3,6,9,18}
basket 19:{1,19}
basket 20:{1,2,4,5,10,20}
basket 21:{1,3,7,21}
basket 22:{1,2,11,22}
basket 23:{1,23}
basket 24:{1,2,3,4,6,8,12,24}
basket 25:{1,5,25}
basket 26:{1,2,13,26}
basket 27:{1,3,9,27}
basket 28:{1,2,4,7,14,28}
basket 29:{1,29}
basket 30:{1,2,3,5,6,10,15,30}
basket 31:{1,31}
basket 32:{1,2,4,8,16,32}
basket 33:{1,3,11,33}
basket 34:{1,2,17,34}
basket 35:{1,5,7,35}
basket 36:{1,2,3,4,6,9,12,18,36}
basket 37:{1,37}
basket 38:{1,2,19,38}
basket 39:{1,3,13,39}
basket 40:{1,2,4,5,8,10,20,40}
basket 41:{1,41}
basket 42:{1,2,3,6,7,14,21,42}
basket 43:{1,43}
basket 44:{1,2,4,11,22,44}
basket 45:{1,3,5,9,15,45}
basket 46:{1,2,23,46}
basket 47:{1,47}
basket 48:{1,2,3,4,6,8,12,16,24,48}
basket 49:{1,7,49}
basket 50:{1,2,5,10,25,50}
basket 56:{1,2,4,7,8,14,28,56}
basket 57:{1,3,19,57}
basket 58:{1,2,29,58}
basket 59:{1,59}
basket 60:{1,2,3,4,5,6,10,12,15,20,30,60}
basket 61:{1,61}
basket 62:{1,2,31,62}
basket 63:{1,3,7,9,21,63}
basket 64:{1,2,4,8,16,32,64}
basket 65:{1,5,13,65}
basket 66:{1,2,3,6,11,22,33,66}
basket 67:{1,67}
basket 68:{1,2,4,17,34,68}
basket 69:{1,3,23,69}
basket 70:{1,2,5,7,10,14,35,70}
basket 71:{1,71}
basket 72:{1,2,3,4,6,8,9,12,18,24,36,72}
basket 73:{1,73}
basket 74:{1,2,37,74}
basket 75:{1,3,5,15,25,75}
basket 76:{1,2,4,19,38,76}
basket 77:{1,7,11,77}
basket 78:{1,2,3,6,13,26,39,78}
basket 79:{1,79}
basket 80:{1,2,4,5,8,16,20,40,80}
basket 81:{1,3,9,27,81}
basket 82:{1,2,41,82}
basket 83:{1,83}
basket 84:{1,2,3,4,6,7,12,14,21,28,42,84}
basket 85:{1,5,17,85}
basket 86:{1,2,43,86}
basket 87:{1,3,29,87}
basket 88:{1,2,4,8,1122,44,88}
basket 89:{1,89}
basket 90:{1,2,4,23,46,90}
basket 91:{1,7,13,91}
basket 92:{1,2,4,23,46,92}
basket 93:{1,3,31,93}
basket 94:{1,2,47,94}
basket 95:{1,5,19,95}
basket 96:{1,2,3,4,6,8,12,16,24,32,48,96}
basket 97:{1,97}
basket 98:{1,2,7,14,49,98}
basket 99:{1,3,9,11,33,99}
basket 100:{1,2,4,5,10,20,25,50,100}

Then, (a)
Items 1 is in baskets: 1,2,3,…,100 (total 100 baskets)
Items 2 is in baskets:2,4,6,…,100(total 50 baskets)
Items 3 is in baskets:3,6,9,…100(total 33 baskets)
…
Therefore, only 20 items are present in more than 5 baskets. And the list of these most frequent item is: {1,2,3,4,5,…,20}

(b)

For pair of first item 1 and second item 2, {1,2}will occur 50 times, as $\left[\frac{100}{1*2} = 50\right]$

pair {1,3} will occurs 33 times since $\left[\frac{100}{1*2} = 33\right]$ …

and so on till 20 since threshold 5, {1,20}will occur 5 times.
Then,
{2,3} occur 16 times; {2,4} occur 25 times; {2,5} occur 10 times;
{2,6} occur 16 times; {2,7} occur 7 times; {2,8} occur 12 times;
{2,9} occur 5 times; {2,10} occur 10 times; {2,12} occur 8 times;
{2,14} occur 7 times; {2,16} occur 6 times; {2,18} occur 5 times;
{2,20} occur 5 times;
Similarly taking first item 3, pair will form,
{3,4} occur 8 times; {3,5} occur 6 times; {3,6} occur 16 times;
{3,9} occur 11 times; {3,12} occur 8 times; {3,15} occur 6 times;
{3,18} occur 5 times;
Similarly, taking first pair with 4,
{4,5} occur 5 times; {4,6} occur 8 times; {4,8} occur 12 times;
{4,10} occur 5 times; {4,12} occur 8 times; {4,16} occur 6 times;
{4,20} occur 5 times;
Similarly, taking first pair with 5,
{5,10} occur 10 times; {5,15} occur 6 times; {5,20} occur 5 times;
Similarly, taking first pair with 6,
{6,9} occur 5 times; {6,12} occur 8 times; {6,18} occur 5 times;
Similarly, for 7, 8, 9, 10, only one pair exist only,
{7,14}, {8,16},{9,18},{10,20}

(c)

Calculating the sum of length of all baskets: 482
Therefore, sum of length of factors for basket index from 1 to 100 will be 482.

(d)

5 and 7 both are available in basket number 35 and 70 together. So, the total number of entries are 2.
And, 2, 5 and 7 are available in basket number 70 only.
Therefore, here the total number of entries are 1.
Confidence({5,7}->2)=support({5,7}∪{2})/support({5,7})=1/2.

第三题

All 100 baskets are shown below:

basket 1:{1,2,3,..,98,99,100}
basket 2:{2,4,6,…,96,98,100}
basket 3:{3,6,9,…,93,96,99}
basket 4:{4,8,12,…,92,96,100}
basket 5:{5,10,15,…,95,100}
basket 6:{6,12,18,…,90,96}
basket 7:{7,14,21,…,91,98}
basket 8:{8,16,24,…,88,96}
basket 9:{9,18,27,…,90,99}
basket 10:{10,20,30,…,90,100}
basket 51:{51}
basket 52:{52}
basket 53:{53}
basket 54:{54}
basket 55:{55}
basket 56:{56}
basket 57:{57}
basket 58:{58}
basket 59:{59}
basket 60:{60}

basket 11:{11,22,33,44,55,66,77,88,99}
basket 12:{12,24,36,48,60,72,84,96}
basket 13:{13,26,52,65,78,91}
basket 14:{14,28,42,56,70,84,98}
basket 15:{15,30,45,60,75,90}
basket 16:{16,32,48,64,80,96}
basket 17:{17,34,51,68,85}
basket 18:{18,36,54,72,90}
basket 19:{19,38,57,76,95}
basket 20:{20,40,60,80,100}
basket 21:{21,42,63,84}
basket 22:{22,44,66,88}
basket 23:{23,46,69,92}
basket 24:{24,48,72,96}
basket 25:{25,50,75,100}
basket 26:{26,52,78}
basket 27:{27,54,81}
basket 28:{28,56,84}
basket 29:{29,58,87}
basket 30:{30,60,90}
basket 31:{31,62,93}
basket 32:{32,64,96}
basket 33:{333,66,99}
basket 34:{34,68}
basket 35:{35,70}
basket 36:{36,72}
basket 37:{37,74}
basket 38:{38,76}
basket 39:{39,78}
basket 40:{40,80}
basket 41:{41,82}
basket 42:{42,84}
basket 43:{43,86}
basket 44:{44,88}
basket 45:{45,90}
basket 46:{46,92}
basket 47:{47,94}
basket 48:{48,96}
basket 49:{49,98}
basket 50:{50,100}
basket 61:{61}
basket 62:{62}
basket 63:{63}
basket 64:{64}
basket 65:{65}
basket 66:{66}
basket 67:{67}
basket 68:{68}
basket 69:{69}
basket 70:{70}
basket 71:{71}
basket 72:{72}
basket 73:{73}
basket 74:{74}
basket 75:{75}
basket 76:{76}
basket 77:{77}
basket 78:{78}
basket 79:{79}
basket 80:{80}
basket 81:{81}
basket 82:{82}
basket 83:{83}
basket 84:{84}
basket 85:{85}
basket 86:{86}
basket 87:{87}
basket 88:{88}
basket 89:{89}
basket 90:{92}
basket 91:{91}
basket 92:{92}
basket 93:{93}
basket 94:{94}
basket 95:{95}
basket 96:{96}
basket 97:{97}
basket 98:{98}
basket 99:{99}
basket 100:{100}

Then perform the A-Priori algorithm and firstly determine the frequent item sets of size 1 with support threshold 5 as shown below:

12,16,18,20,24,28,30,32,36,40,42,44,45,48,50,52,54,56,60,63,64,66,68,70,72,75,76,78,80,81,84,88,90,92,96,98,99,100

Then again perform the A-Priori algorithm and then determine the frequent item sets of size 2 with support threshold 5 as shown below:

{36,72},{40,80},{42,84},{45,90},{48,96},{50,100}

Then again perform the A-Priori algorithm and then determine the frequent item sets of size 3 with support threshold 5 as shown below:

{28,56,84},{30,60,90},{32,64,96}

Then again perform the A-Priori algorithm and then determine the frequent item sets of size 4 with support threshold 5 as shown below:

{24,48,72,96}

Then again perform the A-Priori algorithm and then determine the frequent item sets of size 5 with support threshold 5 as shown below:

{18,36,54,72,90},{20,40,60,80,100}

Then again perform the A-Priori algorithm and then determine the frequent item sets of size 6 with support threshold 5 as shown below:

{16,32,48,64,80,96}
There are no frequent item sets of size 7. Thus, the A-Priori algorithm stops here.

第四题

(a)每个 item 的支持度如下:

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

每一对 items 的支持度如下:

item	pair	Hash Bucket	count
1	{1,2}	2	2
2	{1,3}	3	3
3	{1,4}	4	2
4	{1,5}	5	1
5	{1,6}	6	3
6	{2,3}	6	3
7	{2,4}	8	4
8	{2,5}	10	2
9	{2,6}	1	1
10	{3,4}	1	4
11	{3,5}	4	4
12	{3,6}	7	2

13	{4,5}	9	3
14	{4,6}	2	3
15	{5,6}	8	2

(b)

桶 0:
桶 1:{3,4},{2,6}
桶 2:{1,2},{4,6}
桶 3:{1,3}
桶 4:{3,5},{1,4}
桶 5:{1,5}
桶 6:{1,6},{2,3}
桶 7:{3,6}
桶 8:{2,4},{5,6}
桶 9:{4,5}
桶 10:{2,5}

(c)

Each bucket and its count is:

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

Thus, the given support threshold is 4 and the frequent buckets are those whose support threshold is more than 4. Thus, the frequent buckets are 1,2,4, and 8.

(d)

We have the support threshold greater than 4 for the buckets 1, 2,4, and 8 and their pairs are: {2,6},{3,4},{1,2},{4,6},{1,4},{3,5},{2,4},{5,6}
Therefore, these are selected for the second pass of the PCY algorithm because bucket numbers have a minimum support threshold of 4.