Quiz3

1) **a(hash function: j mod 6)**

frequent bucket #: 0, 3, 4, 5 (1pt)

frequent pairs: {1, 3}, {2, 3}, {2, 4}, {3, 4}, {3, 5}, {4, 5}, {4, 6} (1pt)

b(hash function: j mod 7)

frequent bucket #: 3, 4, 5, 6

frequent pairs: {1, 3}, {2, 3}, {2, 4}, {3, 4}, {3, 5}, {4, 5}, {4, 6}

2) **a(threshold:2)**

{m}, {c}, {b}, {p}, {j}
{m, c}, {m, b}, {m, p}, {b, j}, {m, j}, {c, b}, {c, j}
{m, c, b}, {c, b, j}
(1pt, * lose 0.5 point for each missing/wrong itemset)

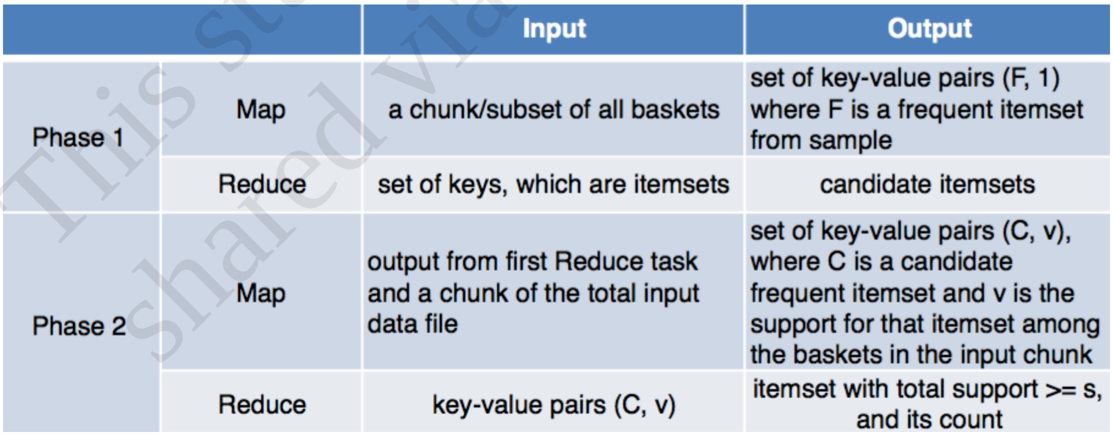
Association rules: (0.5 for association rule, 0.5 for confidence, 0.5 for interest)

1. derived from a frequent pair {m, c}:
   {m}->{c} confidence:3/5 interest : | 3/5-6/8 | = 3/20
2. derived from a frequent triplet{m, c, b}:
   {m, c}->{b} confidence:1 interest: 1-6/8 = 1/4

b(threshold:3)

{m}, {c}, {b}, {j}
{m, c}, {m, b}, {c, b}, {c, j}
{m, c, b}

3) (2pts in total, 1pt for each phase)



		Input	Output
Phase 1	Map	a chunk/subset of all baskets	set of key-value pairs (F, 1) where F is a frequent itemset from sample
	Reduce	set of keys, which are itemsets	candidate itemsets
Phase 2	Map	output from first Reduce task and a chunk of the total input data file	set of key-value pairs (C, v), where C is a candidate frequent itemset and v is the support for that itemset among the baskets in the input chunk
	Reduce	key-value pairs (C, v)	itemset with total support >= s, and its count

4)

Example: singleton in the negative border is not frequent in the sample and its all immediate subsets, which are empty sets, are considered frequent. As for the pair, it is not frequent in the sample, and every single element in the pair is frequent in the sample.
(1pt)

HOW: After we got all the candidate frequent itemsets on the sample and constructed the negative border in the first pass, we count the occurrences of each itemset in both candidate frequent itemsets and negative border. If no itemset from the negative border turns out to be frequent in the whole data set. Then the correct set of frequent itemsets is exactly the itemsets from the sample that were found frequent in the whole data. Or we got no answer this time. Must repeat again with new random sample.
(0.5pt)

WHY: If some member of the negative border is frequent in the whole data set, can't be sure that there are not some even larger itemsets that: 1.Are neither in the negative border nor in the collection of frequent itemsets for the sample. 2. But are frequent in the whole.
(0.5pt)

5) C_n^2 (1pt)