1

> 将输入划分成多个MapReduce任务

(a) 所有整数的平均值：

1. **Map：**
   - 输入：每个整数x。
   - 处理：对每个x，输出键值对 `("avg", (x, 1))`，其中键为固定字符串"avg"，值为元组（数值x，计数1）。
2. **Combiner：**
   - 输入：键"avg"，值列表 `[(x1,1), (x2,1), ...]`。
   - 处理：计算本地总和 `sum_part = Σxi` 和本地计数 `count_part = Σ1`。
   - 输出：键值对 `("avg", (sum_part, count_part))`。
3. **Reduce：**
   - 输入：键"avg"，值列表 `[(sum_part1, count_part1), (sum_part2, count_part2), ...]`。
   - 处理：计算全局总和 `total_sum = Σsum_part` 和全局计数 `total_count = Σcount_part`。
   - 输出：平均值 `total_sum / total_count`。

(b) 整数集合（去重）：

1. **Map：**
   - 输入：每个整数x。
   - 处理：输出键值对 `(x, 1)`，其中键为整数x，值为1。
2. **Combiner：**
   - 输入：键x，值列表 `[1, 1, ...]`。
   - 处理：仅保留一个键x。
   - 输出：`(x, 1)`（每个x输出一次）。
3. **Reduce：**
   - 输入：键x，值列表 `[1, 1, ...]`。
   - 处理：无论值数量，输出键x一次。
   - 输出：`x`。

2

(a) 频繁项：

所有i满足1 ≤ i ≤ 20的项。
当i=20时，支持度=5（购物篮20,40,60,80,100），而i>20时支持度<5。

(b) 频繁项对：

所有i < j且LCM(i,j) ≤ 20的项对。

- (1,2)（LCM=2），支持度=50 ≥5
- (2,3)（LCM=6），支持度=16 ≥5
- (4,5)（LCM=20），支持度=5 ≥5
- (5,7)（LCM=35 >20），支持度=2 <5，故不频繁。

(c) 总项数目之和：

计算所有购物篮b的因数个数之和。结果为 **482**。

$$\sum_{b=1}^{100} d(b) = \sum_{i=1}^{100} \left\lfloor \frac{100}{i} \right\rfloor = 482$$

(d) 可信度：

- 支持度{5,7,2} = 1（仅购物篮70）
- 支持度{5,7} = 2（购物篮35,70）
  可信度 = 1/2 = **50%**

3

计算频繁1项集

支持度阈值：5

> 项i的支持度等于其因数的个数。筛选出因数个数≥5的项。

频繁1项集列表：
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。

生成频繁2项集

> 对任意两个频繁1项i和j，其共同出现次数（即i和j的公因数个数）≥5。
> **等价条件**：i和j的最大公因数d的因数个数≥5（即d在频繁1项集中）。

示例：

- (12,24)：gcd=12（因数6个）
- (16,32)：gcd=16（因数5个）
- (18,36)：gcd=18（因数6个）
- 所有满足gcd(i,j)∈频繁1项集的项对。

生成高阶频繁项集

> 所有k-1项子集必须是频繁的，且k项集的共同支持度（即最大公因数的因数个数）≥5。

示例：

- 3项集{12,24,36}：gcd=12（支持度6）
- 4项集{16,32,48,64}：gcd=16（支持度5）

4

(a) 项及项对的支持度计算

项的支持度:

- 项1: 4
- 项2: 6
- 项3: 8
- 项4: 8
- 项5: 6
- 项6: 4

项对的支持度:

- {1,2}: 2
- {1,3}: 3
- {1,4}: 2
- {1,5}: 1
- {1,6}: 0
- {2,3}: 3
- {2,4}: 4
- {2,5}: 2
- {2,6}: 1
- {3,4}: 4
- {3,5}: 4
- {3,6}: 2
- {4,5}: 3
- {4,6}: 3
- {5,6}: 2

(b) 项对哈希到桶的映射

哈希函数: `i*j mod 11`

- {1,2} → 2
- {1,3} → 3
- {1,4} → 4
- {1,5} → 5
- {1,6} → 6
- {2,3} → 6
- {2,4} → 8
- {2,5} → 10
- {2,6} → 1

- {3,4} → 1
- {3,5} → 4
- {3,6} → 7
- {4,5} → 9
- {4,6} → 2
- {5,6} → 8

(c) 频繁桶

桶号	映射的项对及其支持度	桶总计数
0	无项对	**0**
1	{2,6}(1) + {3,4}(4)	**5** ✔
2	{1,2}(2) + {4,6}(3)	**5** ✔
3	{1,3}(3)	**3**
4	{1,4}(2) + {3,5}(4)	**6** ✔
5	{1,5}(1)	**1**
6	{1,6}(0) + {2,3}(3)	**3**
7	{3,6}(2)	**2**
8	{2,4}(4) + {5,6}(2)	**6** ✔
9	{4,5}(3)	**3**
10	{2,5}(2)	**2**

频繁桶（计数≥4）：**1、2、4、8**

(d) 第二次扫描中计数的项对

> 1.两个项均为频繁项
>
> 2.项对的哈希桶为频繁桶

符合条件的项对：

- {1,2}（桶2）
- {1,4}（桶4）
- {2,4}（桶8）
- {2,6}（桶1）
- {3,4}（桶1）
- {3,5}（桶4）
- {4,6}（桶2）
- {5,6}（桶8）