

Question 2

(b)

- **The skewness of data.** The workload is not evenly distributed among different machines.
- **Communication overhead.** The time spent in scheduling and data transfer can not be parallelized.
- **I/O cost.** When more machines are used, the total time spent in I/O operations are higher.
- **CPU utilization.** When the CPU utilization reaches 100%. Adding tasks to the same machine won't help.
- **Task initialization cost.** There are also time spent in initialization for each machine.
- **Variance of machines.** Some machines can be slower.

Question 3

```
member_pos = FlatMap(data, by = lambda (memberId, posList): map(lambda (posId, compId): (memberId, compId), posList)) \
| Distinct()

member_self_join = Join(Jin(member_pos, by = lambda (memberId, compId): memberId), Jin(member_pos, by = lambda (memberId, compId): memberId)) \
| Group(by = lambda ((member1, comp1), (member2, comp2)): (comp1, comp2), reducingTo = ReduceToCount())

res = Join(Jin(member_self_join, by = lambda (pair, cnt): pair), Jin(query, outer = True)) \
| Map(by = lambda (pair_cnt, pair): (pair[0], pair[1], 0) if pair_cnt is None else (pair[0], pair[1], pair_cnt[1]) )
```

Question 4

t3 will contain purely alphabetic tokens, that are present at least twice in the document: Once with just alphabetic characters (from t1, like "large"), and another time with a non-alphanumeric symbol (from t2, like "lar!ge"). Note that two tokens such as "abc" and "abc3", even if present, would not appear in t3.