Homework: Frequent Itemset Mining & Association Rules

Part A — Conceptual Questions

1. Define: basket, item, support, confidence, and interest.

2. Explain why a rule can have high confidence but still be uninteresting.

3. Why is finding frequent pairs harder than frequent triples?

4. Distinguish between basket and bucket.

5. Explain the Apriori monotonicity principle.

Part B — Support & Confidence Calculations

Given dataset of 8 baskets (provided separately):

1. Find all frequent 1-itemsets and 2-itemsets with s=3.

2. Compute support({m,b}), confidence({m,b}→c), and interest.

3. Generate all association rules from any frequent 3-itemsets.

Part C — Apriori Algorithm Tasks

1. Perform Apriori Pass 1 and identify frequent items.

2. Generate C2, L2, C3 using pruning.

3. Explain why Apriori requires multiple passes.

Part D — PCY, Multistage, Multihash

1. Simulate PCY Pass 1 with h(i,j) = (i*j) mod 5 for basket {1,3,4}.

2. Determine frequent buckets for s=3 and candidate pairs.

3. Compare multistage vs multihash.

Part E — Programming Tasks

1. Use PySpark FP-Growth to compute frequent itemsets.

2. Implement simple Apriori in Python.

3. Simulate PCY with hashing and bitmap.

Part F — Essay Questions

1. Explain the main-memory bottleneck.

2. When to prefer Apriori, PCY, random sampling, or SON?