INTRODUCTION TO

PATTERN DISCOVERY

TREY BROOKS

DATA ENGINEER / STUDENT

Healthcare Bluebook™

	CRAZY RONNY'S COLLEGE STUFF EMPORIUM
T1	Pencils, Ramen Noodles, Index Cards, Calculator, Notebook
T2	Ramen Noodles, Index Cards, Cat Food
T3	Pencils, Ramen Noodles, Cat Food, Notebook
T4	Pencils, Ramen Noodles, Index Cards, Cat Food, Calculator, Notebook
T5	Pencils, Index Cards, Cat Food, Calculator
T6	Ramen Noodles, Cat Food, Calculator
T7	Pencils, Index Cards, Cat Food
T8	Pencils, Ramen Noodles, Index Cards, Calculator

In [1]:
```
from itertools import combinations
unq_items = ['Pencils', 'Notebook', 'Index Cards', 'Calculator', 'Ramen Noodles', 'Cat Food']

cand_itemsets = []
for r in range(len(unq_items)+1):
    [cand_itemsets.append({', '.join(itemset)}) for itemset in  combinations(unq_items, r)]
print(cand_itemsets, len(cand_itemsets))
```

```
[{''}, {'Pencils'}, {'Notebook'}, {'Index Cards'}, {'Calculator'}, {'Ramen Noodles'}, {'Cat Food'}, {'Pencils, Notebo
ok'}, {'Pencils, Index Cards'}, {'Pencils, Calculator'}, {'Pencils, Ramen Noodles'}, {'Pencils, Cat Food'}, {'Noteboo
k, Index Cards'}, {'Notebook, Calculator'}, {'Notebook, Ramen Noodles'}, {'Notebook, Cat Food'}, {'Index Cards, Calcu
lator'}, {'Index Cards, Ramen Noodles'}, {'Index Cards, Cat Food'}, {'Calculator, Ramen Noodles'}, {'Calculator, Cat
 Food'}, {'Ramen Noodles, Cat Food'}, {'Pencils, Notebook, Index Cards'}, {'Pencils, Notebook, Calculator'}, {'Pencil
s, Notebook, Ramen Noodles'}, {'Pencils, Notebook, Cat Food'}, {'Pencils, Index Cards, Calculator'}, {'Pencils, Index
 Cards, Ramen Noodles'}, {'Pencils, Index Cards, Cat Food'}, {'Pencils, Calculator, Ramen Noodles'}, {'Pencils, Calcul
ator, Cat Food'}, {'Pencils, Ramen Noodles, Cat Food'}, {'Notebook, Index Cards, Calculator'}, {'Notebook, Index Card
s, Ramen Noodles'}, {'Notebook, Index Cards, Cat Food'}, {'Notebook, Calculator, Ramen Noodles'}, {'Notebook, Calcula
tor, Cat Food'}, {'Notebook, Ramen Noodles, Cat Food'}, {'Index Cards, Calculator, Ramen Noodles'}, {'Index Cards, Ca
lculator, Cat Food'}, {'Index Cards, Ramen Noodles, Cat Food'}, {'Calculator, Ramen Noodles, Cat Food'}, {'Pencils, N
otebook, Index Cards, Calculator'}, {'Pencils, Notebook, Index Cards, Ramen Noodles'}, {'Pencils, Notebook, Index Car
ds, Cat Food'}, {'Pencils, Notebook, Calculator, Ramen Noodles'}, {'Pencils, Notebook, Calculator, Cat Food'}, {'Penc
ils, Notebook, Ramen Noodles, Cat Food'}, {'Pencils, Index Cards, Calculator, Ramen Noodles'}, {'Pencils, Index Card
s, Calculator, Cat Food'}, {'Pencils, Index Cards, Ramen Noodles, Cat Food'}, {'Pencils, Calculator, Ramen Noodles, C
at Food'}, {'Notebook, Index Cards, Calculator, Ramen Noodles'}, {'Notebook, Index Cards, Calculator, Cat Food'}, {'N
otebook, Index Cards, Ramen Noodles, Cat Food'}, {'Notebook, Calculator, Ramen Noodles, Cat Food'}, {'Index Cards, Ca
lculator, Ramen Noodles, Cat Food'}, {'Pencils, Notebook, Index Cards, Calculator, Ramen Noodles'}, {'Pencils, Noteboo
k, Index Cards, Calculator, Cat Food'}, {'Pencils, Notebook, Index Cards, Ramen Noodles, Cat Food'}, {'Pencils, Note
book, Calculator, Ramen Noodles, Cat Food'}, {'Pencils, Index Cards, Calculator, Ramen Noodles, Cat Food'}, {'Noteboo
k, Index Cards, Calculator, Ramen Noodles, Cat Food'}, {'Pencils, Notebook, Index Cards, Calculator, Ramen Noodles, C
at Food'}] 64
```

Unique Item Count	Possible Combinations
6	64
200	1.6069 x 10^60
300	2.0370 x 10 ^ 90
480,000,000	8.29 x 10^144494397

SUPPORT, CONFIDENCE, & ASSOCIATION RULES

CRAZY RONNY'S COLLEGE STUFF EMPORIUM	
T1	Pencils, Ramen Noodles, Index Cards, Calculator, Notebook
T2	Ramen Noodles, Index Cards, Cat Food
T3	Pencils, Ramen Noodles, Cat Food, Notebook
T4	Pencils, Ramen Noodles, Index Cards, Cat Food, Calculator, Notebook
T5	Pencils, Index Cards, Cat Food, Calculator
T6	Ramen Noodles, Cat Food, Calculator
T7	Pencils, Index Cards, Cat Food
T8	Pencils, Ramen Noodles, Index Cards, Calculator

N : 8

items : Pencils, Notebook, Index Cards, Calculator, Ramen Noodles, Cat Food

Support	sum([1 for transaction in TDB if x in transaction])
Relative Support	support / len(TDB)

SUPPORT, CONFIDENCE, & ASSOCIATION RULES

CRAZY RONNY'S COLLEGE STUFF EMPORIUM	
T1	Pencils, Ramen Noodles, Index Cards, Calculator, Notebook
T2	Ramen Noodles, Index Cards, Cat Food
T3	Pencils, Ramen Noodles, Cat Food, Notebook
T4	Pencils, Ramen Noodles, Index Cards, Cat Food, Calculator, Notebook
T5	Pencils, Index Cards, Cat Food, Calculator
T6	Ramen Noodles, Cat Food, Calculator
T7	Pencils, Index Cards, Cat Food
T8	Pencils, Ramen Noodles, Index Cards, Calculator

N : 8

items : Pencils, Notebook, Index Cards, Calculator, Ramen Noodles, Cat Food

Minimum Support : 0.5 => (4/8)

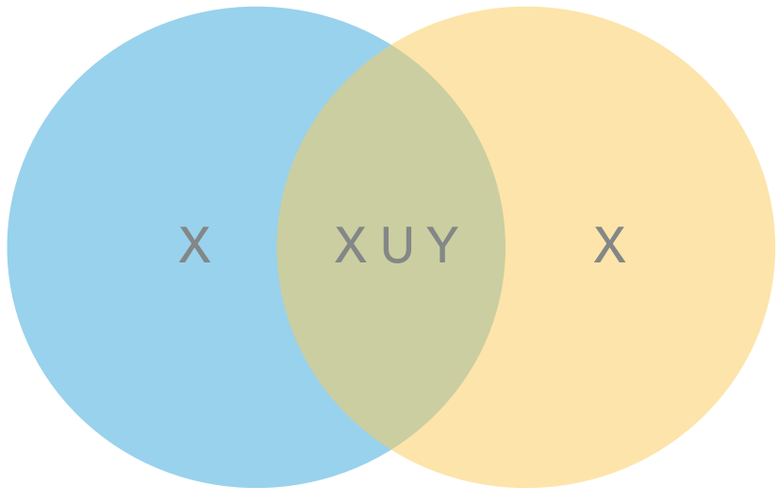
Pencils	6
Ramen Noodles	6
Index Cards	6
Cat Food	6
Calculator	5
~~Notebook~~	~~3~~

Calculator, Index Cards	4
Index Cards, Ramen Noodles	4
Cat Food, Pencils	4
Cat Food, Index Cards	4
Pencils, Ramen Noodles	4
Calculator, Ramen Noodles	4
Index Cards, Pencils	5
Calculator, Pencils	4
Cat Food, Ramen Noodles	4

Calculator, Index Cards, Pencils	4

SUPPORT, CONFIDENCE, & ASSOCIATION RULES

Support:



Confidence:

Support(X & Y) / Support(X)

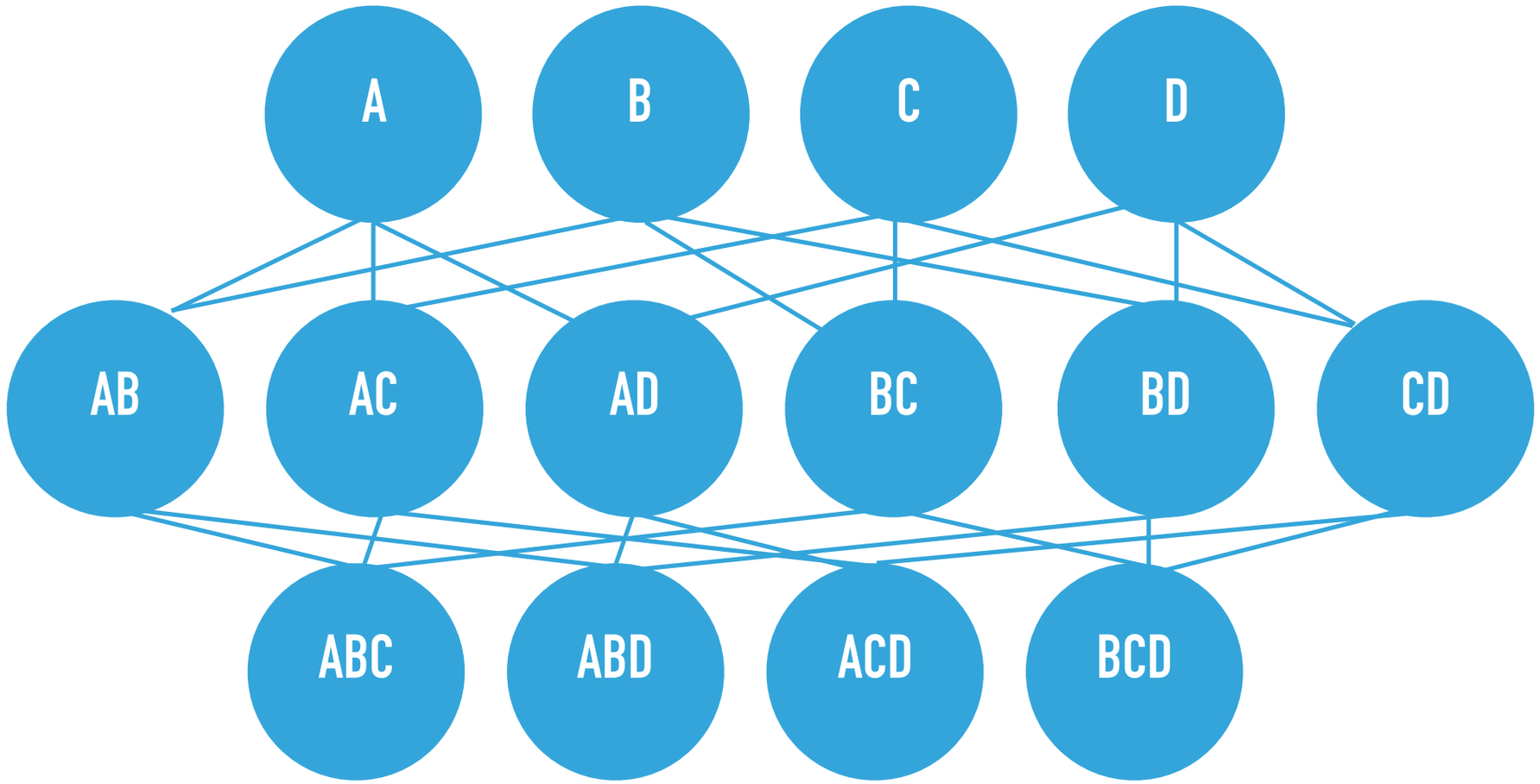
Association Rules	Relative Support	Confidence
{Pencils} -> {Calculator}	0.5	0.66
{Calculator, Index Cards} -> {Pencils}	0.5	1
{Ramen Noodles} -> {Cat Food}	0.5	0.66
{Calculator, Pencils} -> {Index Cards}	0.5	1
Index Cards, Pencils -> Calculator	0.5	0.8

Minimum Support: 50%

Minimum Confidence: 50%

APRIORI

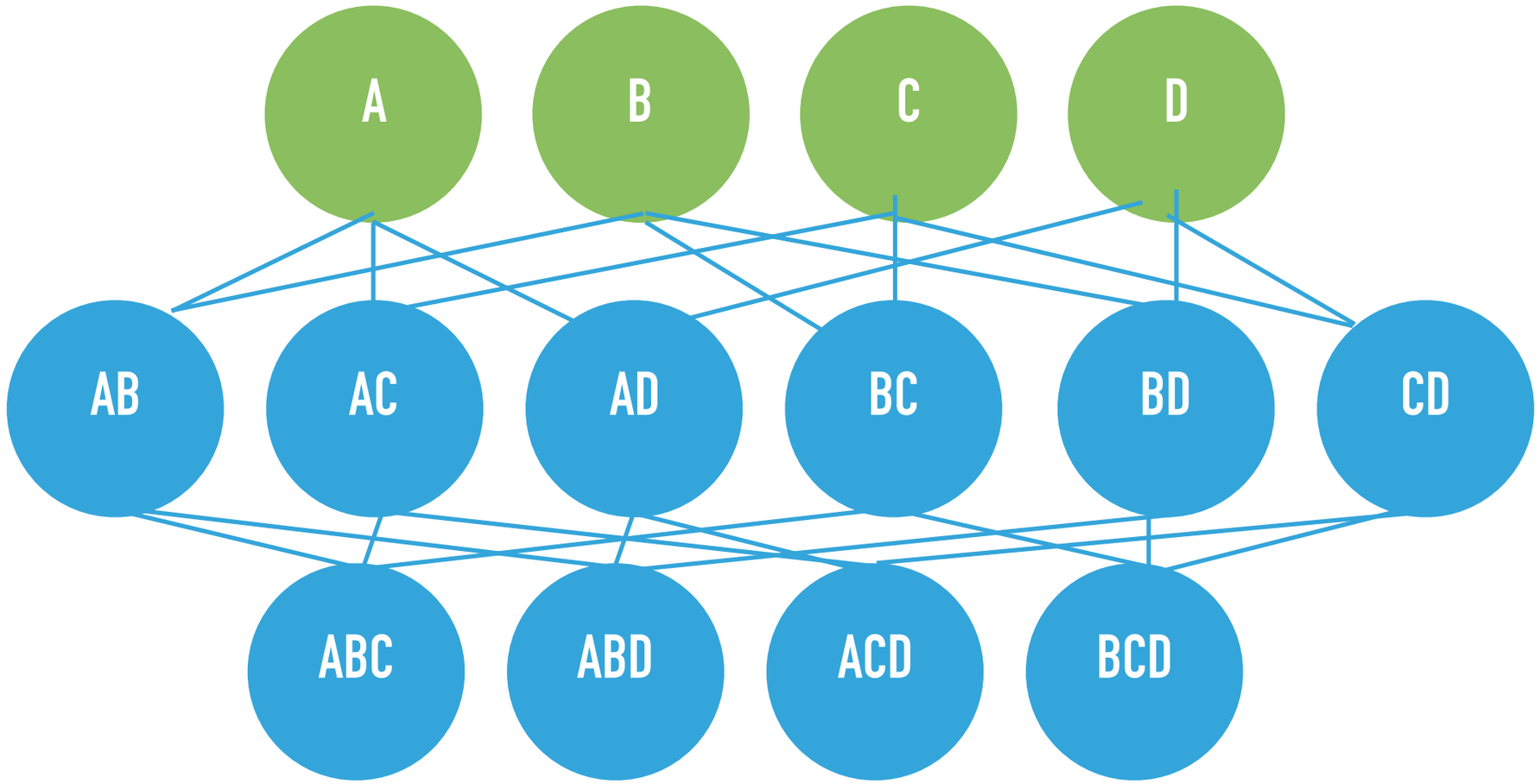
APRIORI RULE



TID	
T1	A, B, C, D
T2	A, B
T3	B, C, D
T4	A, B, C, D
T5	B, C, D
T6	A, B, D

Minimum Support: 60% -> (4)

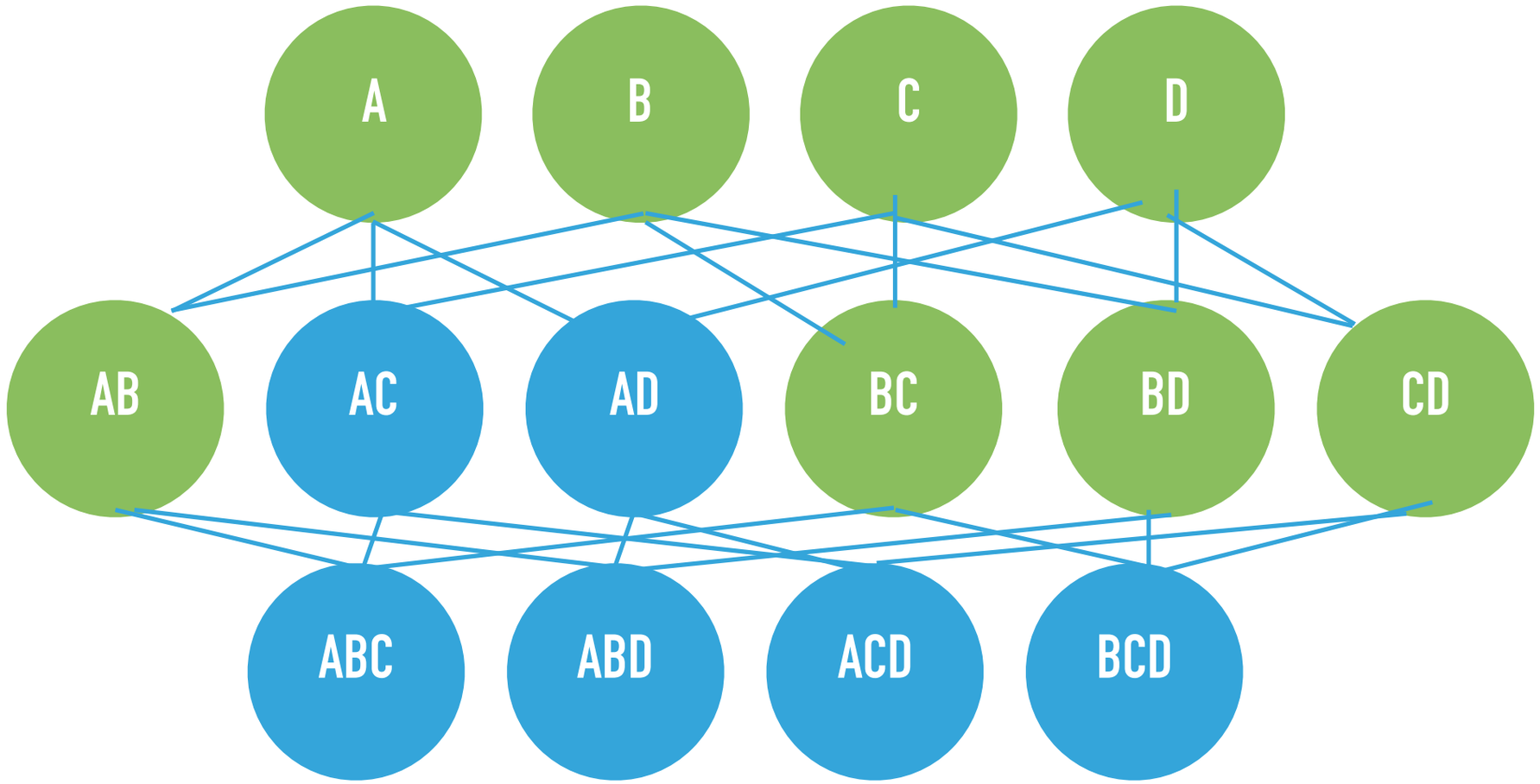
APRIORI RULE



TID	
T1	A, B, C, D
T2	A, B
T3	B, C, D
T4	A, B, C, D
T5	B, C, D
T6	A, B, D

Minimum Support: 60% -> (4)

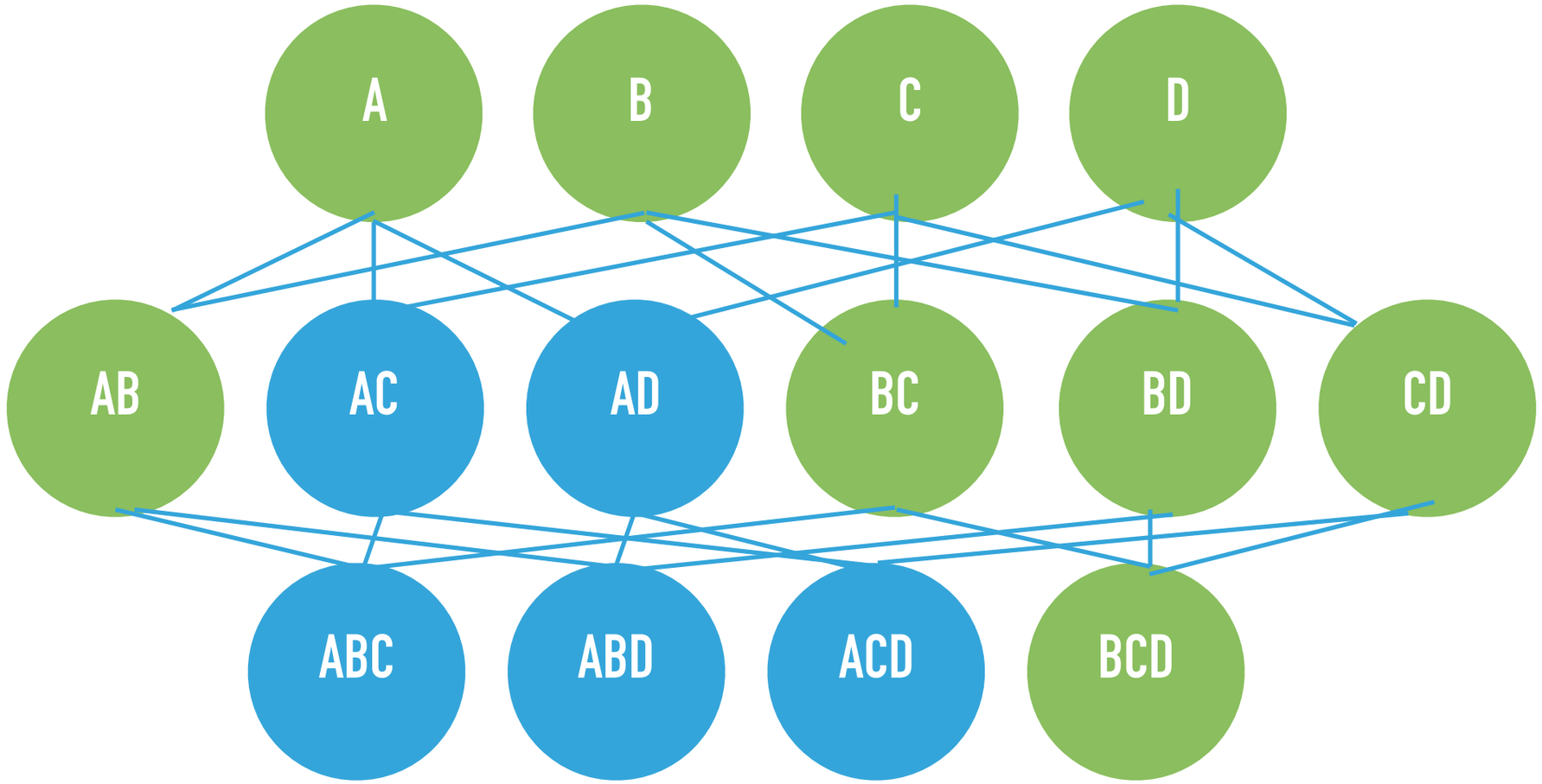
APRIORI RULE



TID	
T1	A, B, C, D
T2	A, B
T3	B, C, D
T4	A, B, C, D
T5	B, C, D
T6	A, B, D

Minimum Support: 60% -> (4)

APRIORI RULE



TID	
T1	A, B, C, D
T2	A, B
T3	B, C, D
T4	A, B, C, D
T5	B, C, D
T6	A, B, D

Minimum Support: 60% -> (4)


60

FP GROWTH

	CRAZY RONNY'S COLLEGE STUFF EMPORIUM
T1	Pencils, Ramen Noodles, Index Cards, Calculator, Notebook
T2	Ramen Noodles, Index Cards, Cat Food
T3	Pencils, Ramen Noodles, Cat Food, Notebook
T4	Pencils, Ramen Noodles, Index Cards, Cat Food, Calculator, Notebook
T5	Pencils, Index Cards, Cat Food, Calculator
T6	Ramen Noodles, Cat Food, Calculator
T7	Pencils, Index Cards, Cat Food
T8	Pencils, Ramen Noodles, Index Cards, Calculator

Item	ID
Pencils	A
Ramen Noodles	B
Index Cards	C
Cat Food	D
Calculator	E
Notebook	F

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E

FP GROWTH (STEP 1, SCAN 1)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E

SORT HERE

TID	
A	6
B	6
C	6
D	6
E	5
F	3

TID	
C	6
D	6
E	6
A	6
B	5
F	3

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E

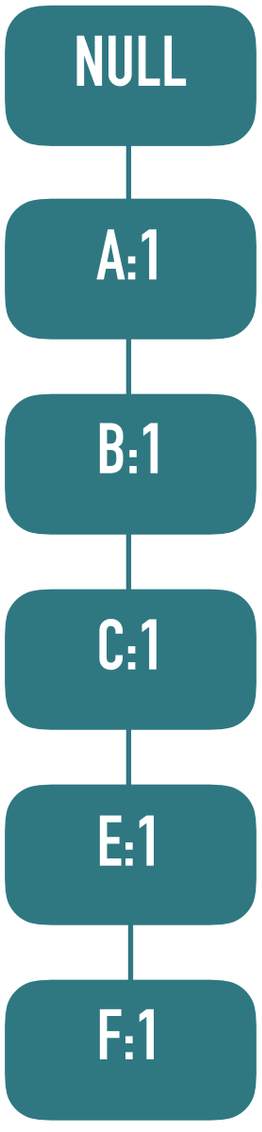
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E

NULL

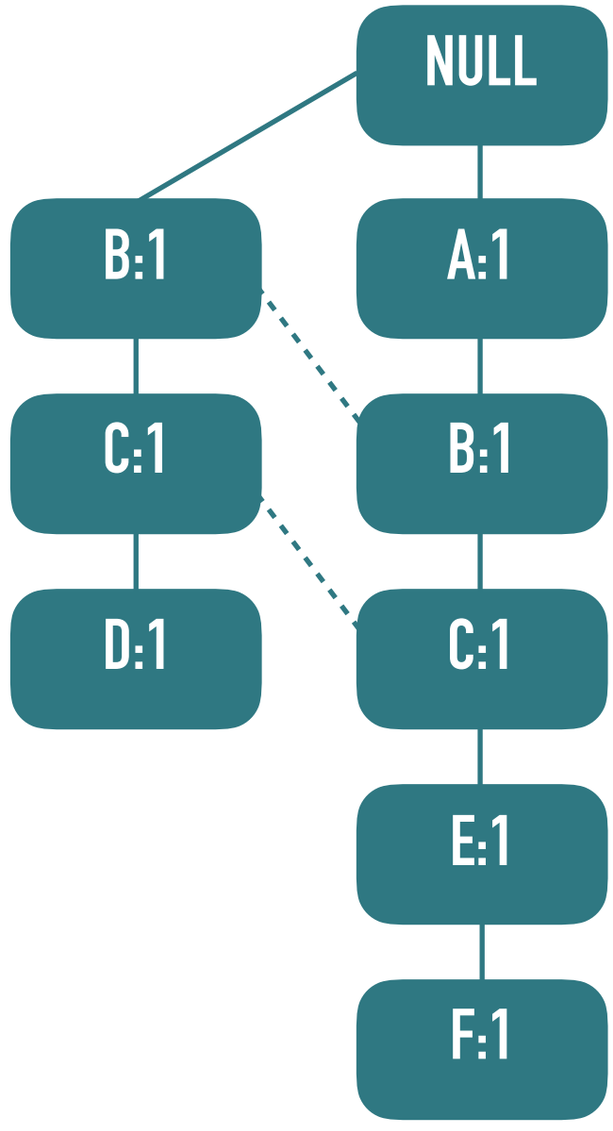
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



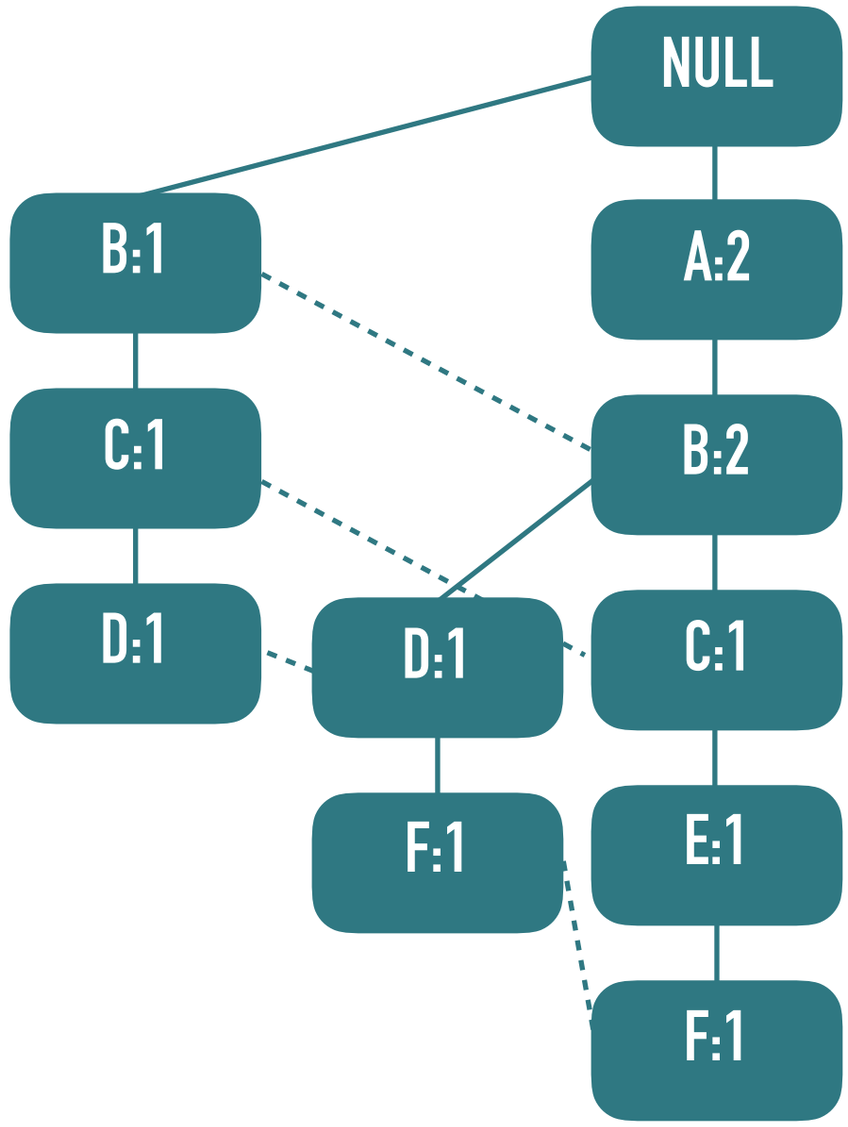
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



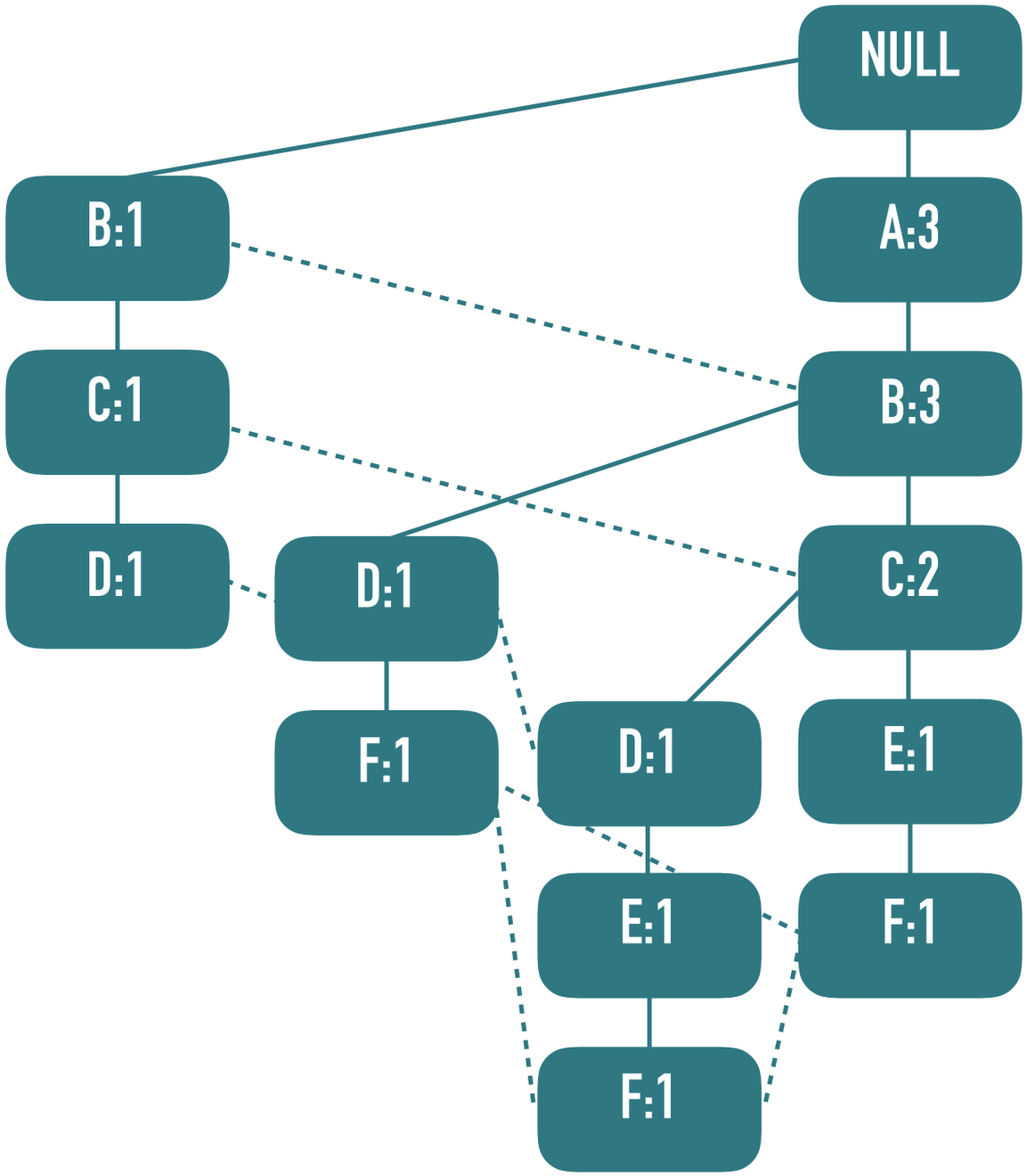
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



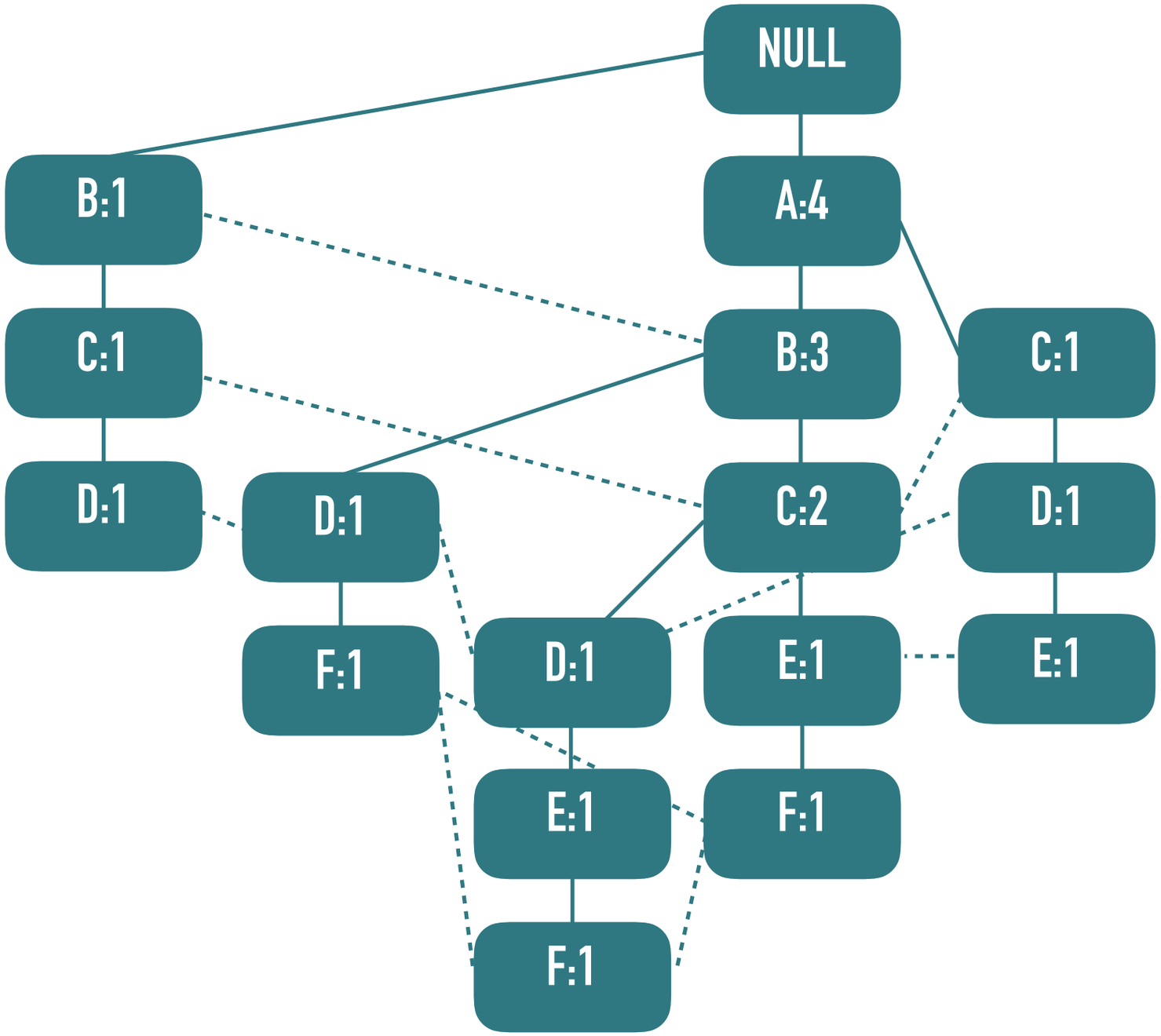
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



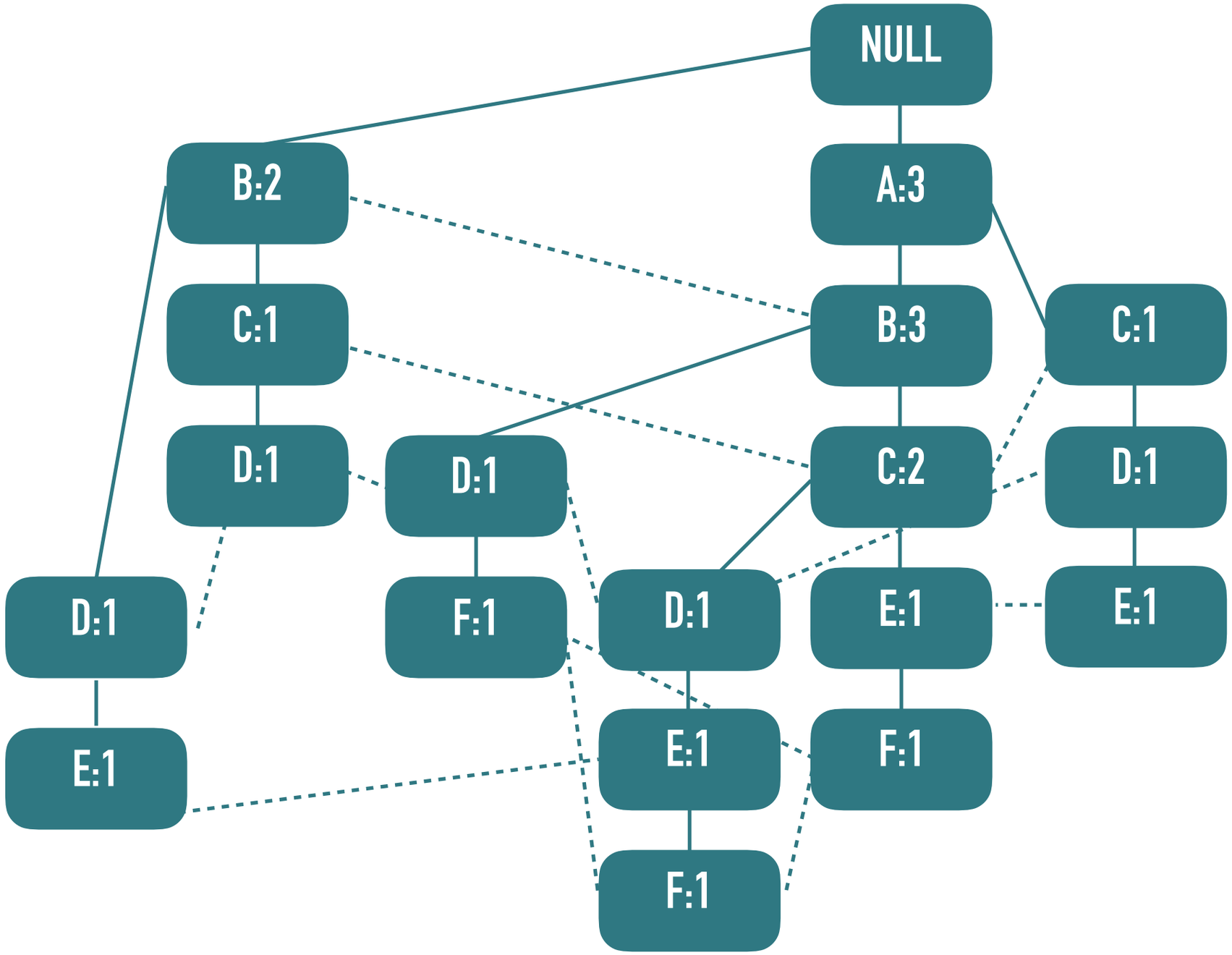
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



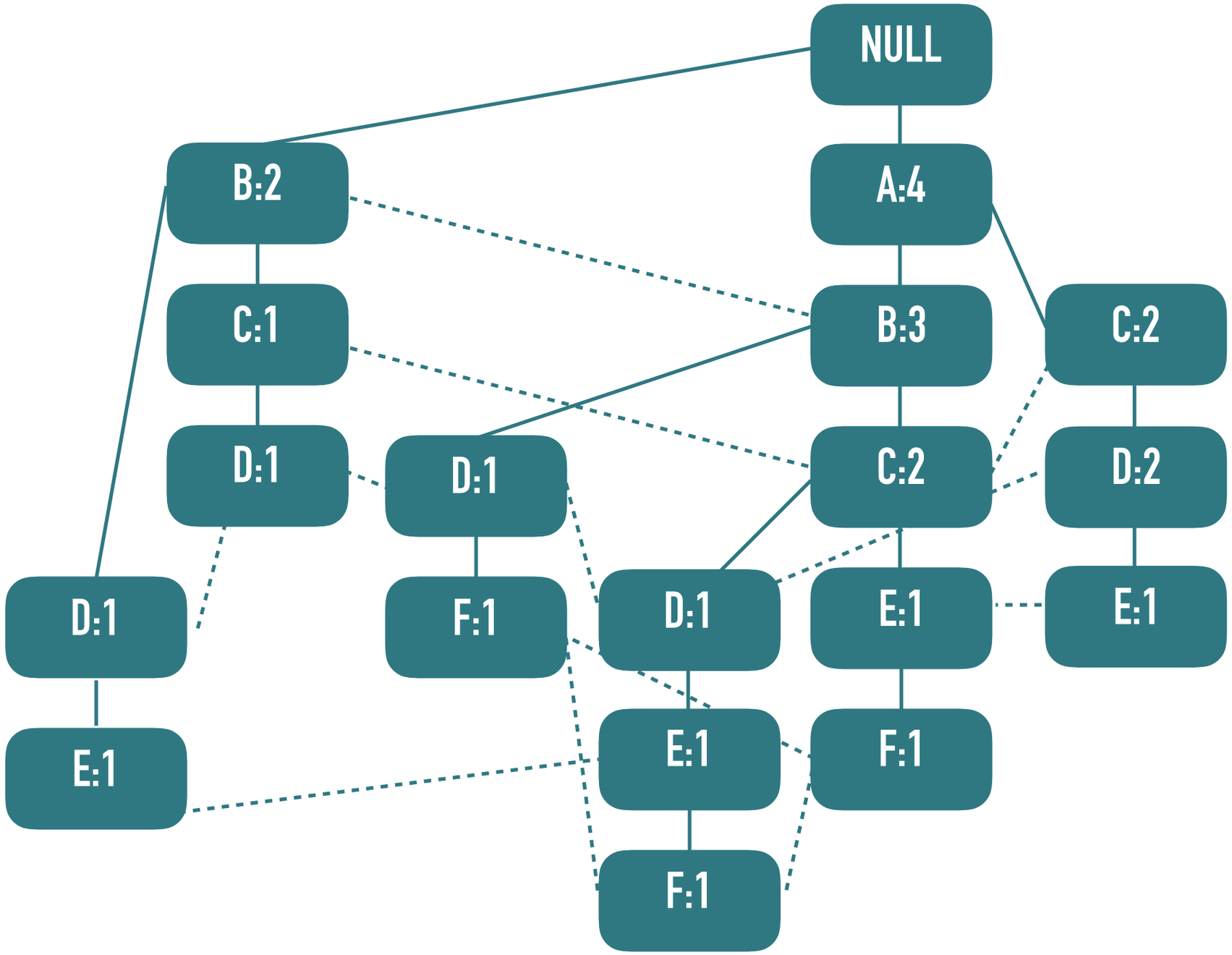
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



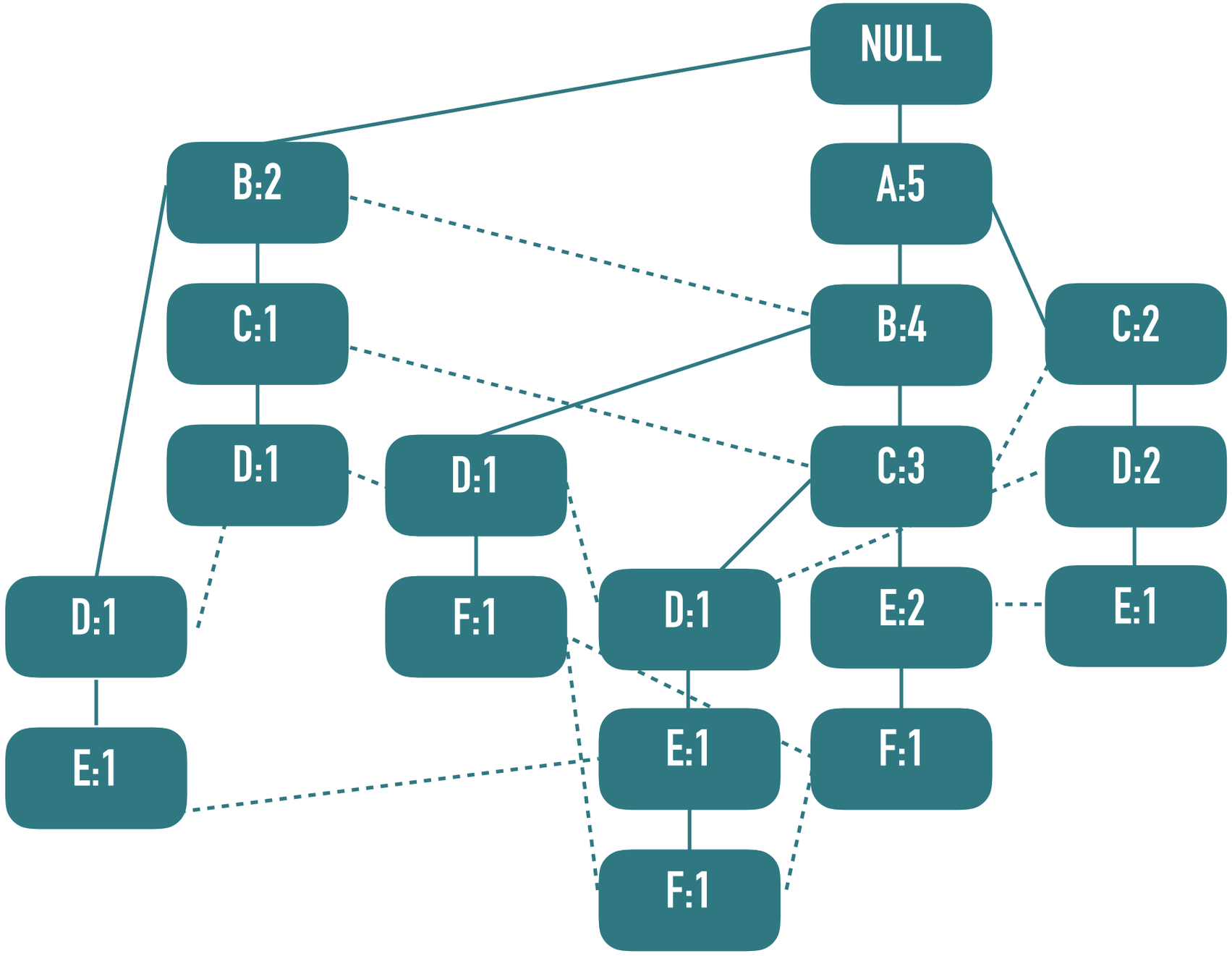
FP GROWTH (STEP 1, SCAN 2)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



FP GROWTH (STEP 1, SCAN 2)

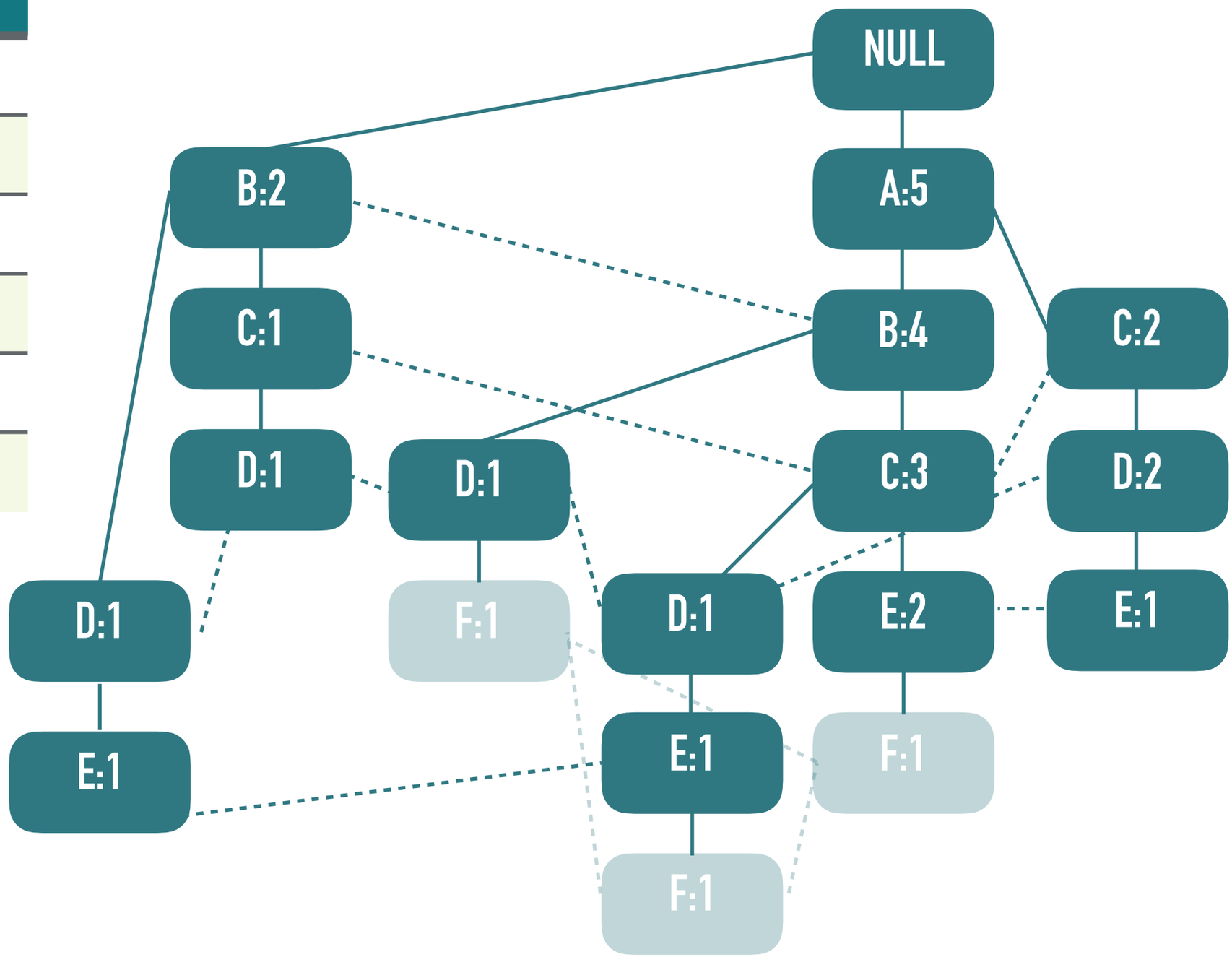
TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



FP GROWTH (STEP 2)

Minimum Support: 4

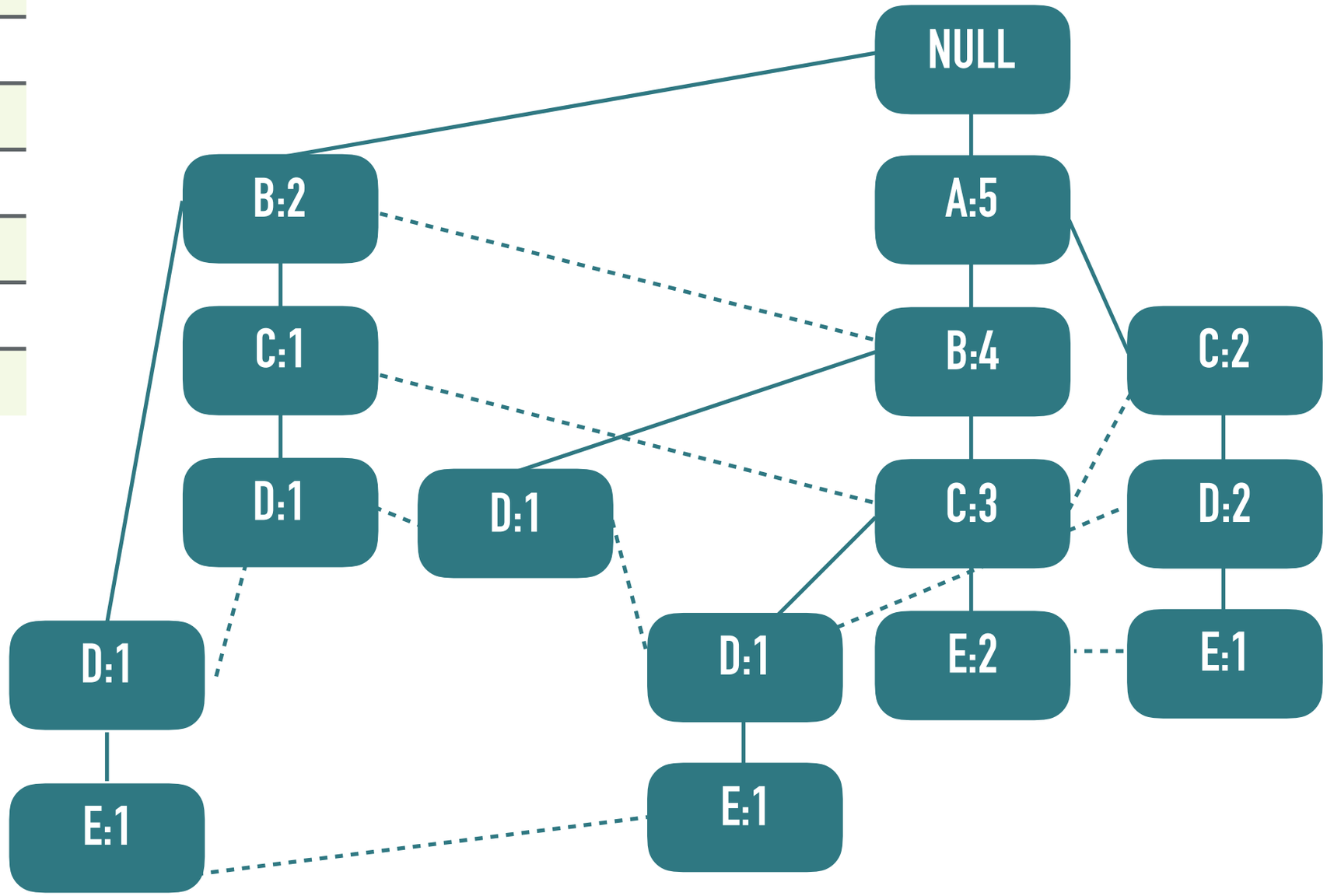
TID	
A	6
B	6
C	6
D	6
E	5
F	3



FP GROWTH (STEP 2)

TID	
T1	A, B, C, E, ~~F~~
T2	B, C, D
T3	A, B, D, ~~E~~
T4	A, B, C, D, E, ~~F~~
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E

Minimum Support: 4

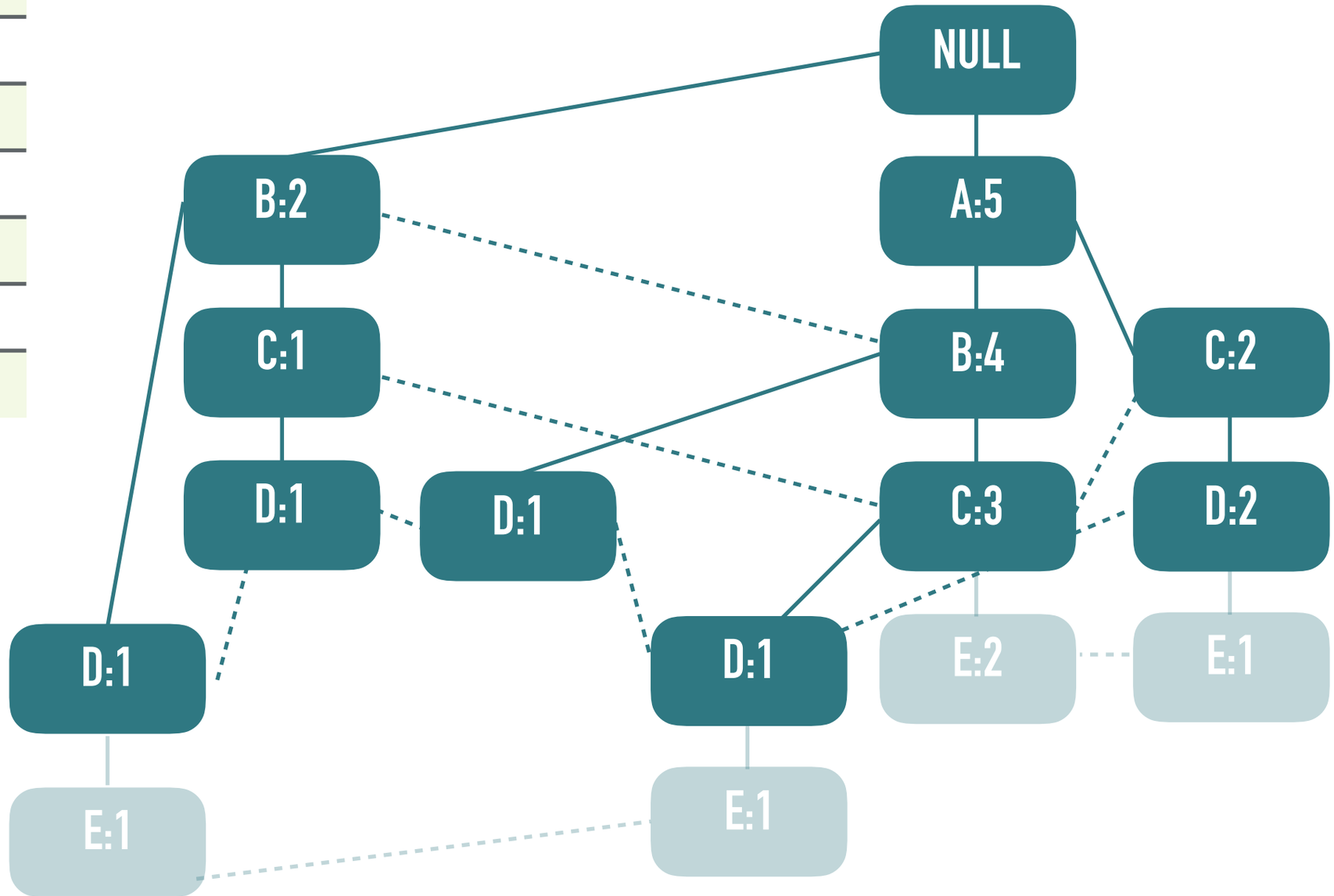


FP GROWTH (STEP 2) E - CONDITIONAL TREE

TID	
T1	A, B, C, E, ~~F~~
T2	B, C, D
T3	A, B, D, ~~E~~
T4	A, B, C, D, E, ~~F~~
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E

Patterns: {E}

Minimum Support: 4

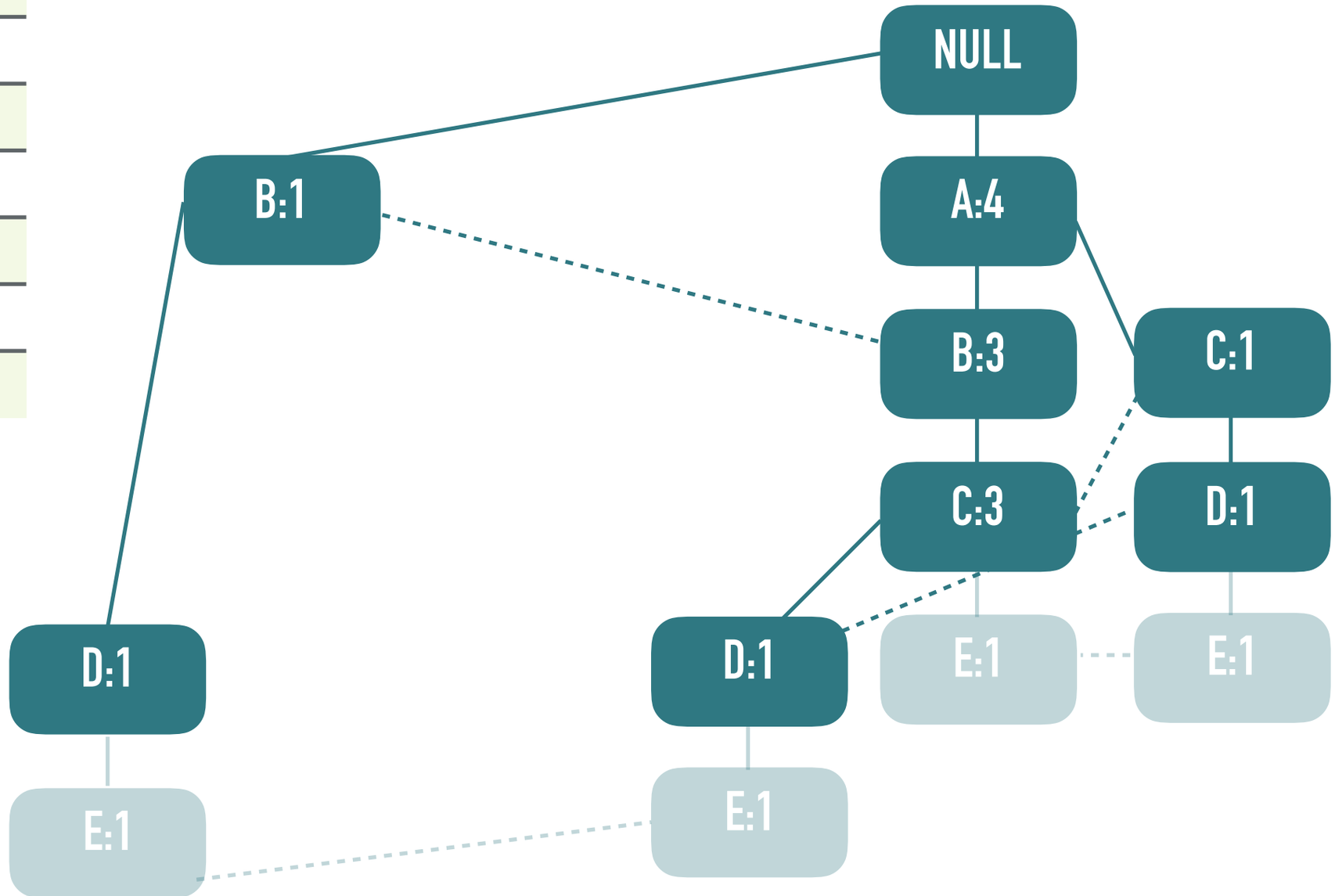


FP GROWTH (STEP 2) E - CONDITIONAL TREE

TID	
T1	A, B, C, ~~E~~, ~~F~~
~~T2~~	~~B, C, D~~
~~T3~~	~~A, B, D, E~~
T4	A, B, C, D, ~~E~~, ~~F~~
T5	A, C, D, ~~E~~
T6	B, D, ~~E~~
~~T7~~	~~A, C, D~~
T8	A, B, C, ~~E~~

Patterns: {E, C}, {E, A}, {E, B}

Minimum Support: 4

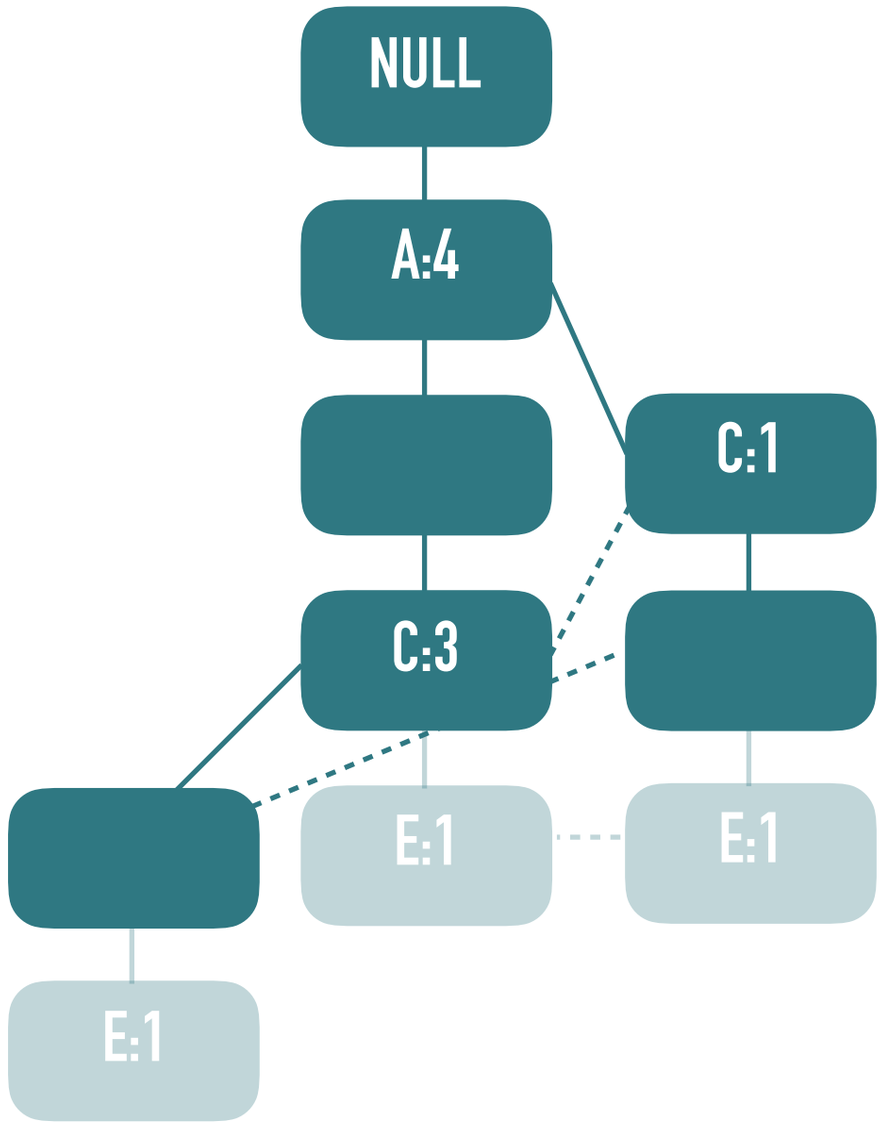


FP GROWTH (STEP 2) EC - CONDITIONAL TREE

TID	
T1	A, B, ~~C~~, ~~E~~, ~~F~~
~~T2~~	~~B, C, D~~
~~T3~~	~~A, B, D, F~~
T4	A, B, ~~C~~, D, ~~E~~, ~~F~~
T5	A, ~~C~~, D, ~~E~~
~~T6~~	~~B, D, E~~
~~T7~~	~~A, C, D~~
T8	A, B, ~~C~~, ~~E~~

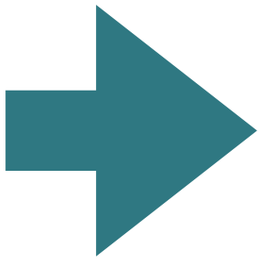
Patterns: {E, C, A}

Minimum Support: 4



FP GROWTH (RESULTS)

TID	
T1	A, B, C, E, F
T2	B, C, D
T3	A, B, D, F
T4	A, B, C, D, E, F
T5	A, C, D, E
T6	B, D, E
T7	A, C, D
T8	A, B, C, E



Pattern	Frequency
A	6
B	6
C	6
D	6
E	5
F	3
E, C	4
C, B	4
D, A	4
D, C	4
A, B	4
E, B	4
C, A	5
E, A	4
D, B	4
E, C, A	4

Pros:

Compresses data

Only 2 passes through database

Way faster than Apriori

Can be partitioned and parallelized

Cons:

Tree might not fit into memory

Expensive to build

INTERSTINGNESS

INTERESTINGNESS (LIFT, MEDICAL)

Lift: Conf(A, B) / Supp(B) == Supp(A & B) / (Supp(A) X Supp(B))

	Knee Replacement	No Knee Replacement	Sum
Knee Xray	13000	3000000	3013000
No Knee Xray	500	10200000	10200500
Sum	14000	13200000	13213500

Support(Xray, Replacement) = (13000/13213500) = 0.00098
Confidence(Xray, Replacement) = (13000/13213500)/(14000/13213500) = 0.93

$$\text{Lift(Xray, Replacement)} = \frac{(13000/13213500)}{((3013000/13213500) \times (14000/13213500))}$$

Lift(X, R)	4.072246456
Lift(X, not R)	0.996703678
Lift(not X, R)	0.04626348848
Lift(not X, not R)	1.00097366

Limits: [0, inf)

INTERESTINGNESS (X^2, MEDICAL)

Chi Squared: ChiSqr(A, B) =Sigma((Obs-Exp)^2 / Exp)

	Knee Replacement	No Knee Replacement	Sum
Knee Xray	13000 (3192)	3000000 (3009921)	3013000
No Knee Xray	500 (10807)	10200000 (10190078)	10200500
Sum	14000	13200000	13213500

Support(Xray, Replacement) = (13000/13213500) = 0.00098
Confidence(Xray, Replacement) = (13000/13213500)/(14000/13213500) = 0.93

ChiSquare =(((13000-3192)^2)/3192) +
(((500-10807)^2)/10807) +
(((3000000-3009921)^2)/3009921) +
(((10200000-10190078)^2)/10190078) = 40005

Limits: [0, inf)

INTERESTINGNESS (NULL INVARIANCE)

	Knee Replacement	No Knee Replacement	Sum
Knee Xray	13000	3000000	3013000
No Knee Xray	500	10200000	10200500
Sum	14000	13200000	13213500

Kulczynski Measure: 1/2 * ((Supp(A & B) / Supp(A)) + (Supp(A & B) / Supp(B)))

Limits: [0, 1] 0.5 * (13000/14000 + 13000/3013000) = 0.466

Imbalance Ratio: |Supp(A) - Supp(B)| / (Supp(A) + Supp(B) - Supp(A & B))

Limits: [0, 1] |14000 - 3013000| / (14000 + 3013000 - 13000) = 0.995

* Some real correlation

* Highly imbalanced

WHERE CAN WE TAKE THIS

Frequency Pattern Mining - Transactional Databases

Downward Closure

ECLAT - exploring vertical data format

FPGrowth - frequent pattern-growth approach

CLOSET+ - mining closed patterns

Graph Pattern Mining

FSG - apriori

gSpan - growth based

CloseGraph -closed graph patterns

SpiderMine - top-k large structure patterns

Sequential Pattern Mining

GSP - Generalized Sequential Patterns

SPADE - vertical format based mining

PrefixSpan - pattern growth method

CloSpan - closed pattern mining

Phrase Mining

Previous Phase mining

TurboTopics: Uses LDA (topicing)
and KERT (postprocessing)

ToP Mine: Mining without training data

SegPhrase: Mining with minimal training data