Khai khoáng dữ liệu

Hà Thị Thúy Vi

NHÓM 9 : THUẬT TOÁN FP-GROWTH

1. Võ Minh Châu 110117045
2. Nguyễn Anh Khoa 110116035

FP-GROWTH

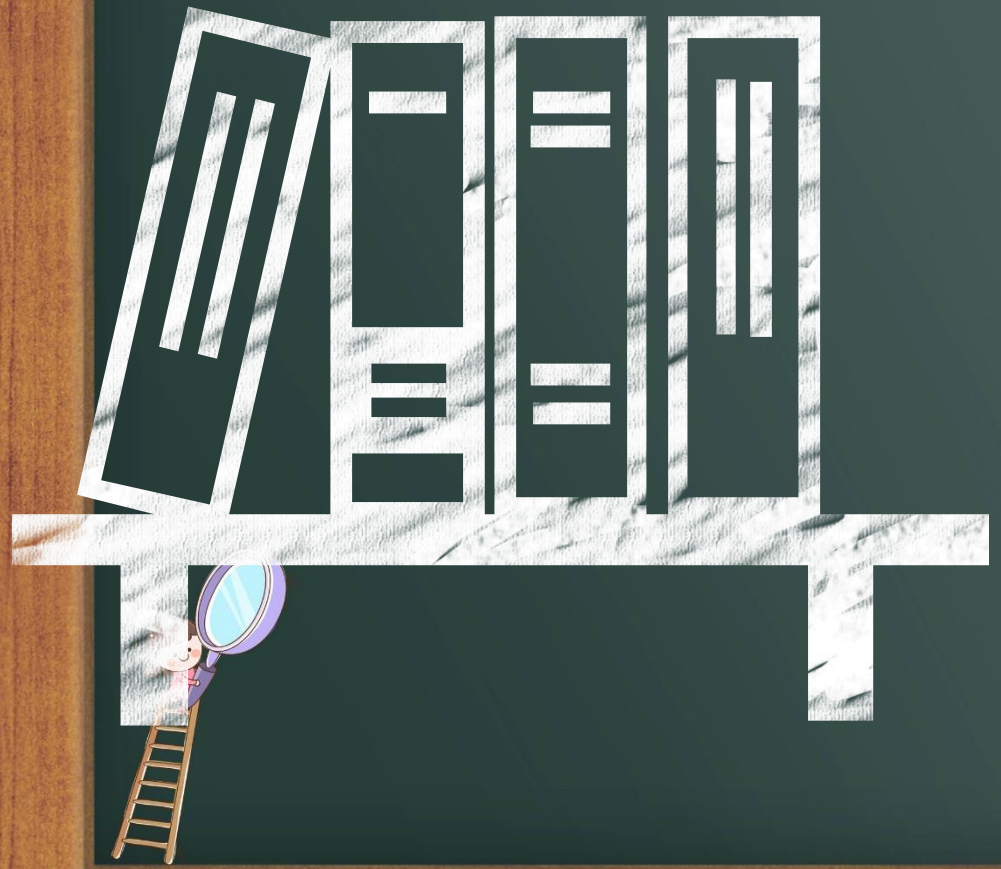
Đặc trưng

Ưu/Nhược điểm

Ứng dụng


01
LÝ THUYẾT VỀ FP-GROWTH
FP-GROWTH là gì?

FP-GROWTH



Lý Thuyết

FP-Growth biểu diễn dữ liệu các giao dịch bằng một cấu trúc dữ liệu gọi là FP-Tree

FP-Growth sử dụng FP-Tree để xác định trực tiếp các tập hạng mục phổ biến (không sinh các tập hạng mục ứng viên từ các tập hạng mục ứng viên trước).

Đặc trưng

Khai thác tập phổ biến
không sử dụng hàm tạo ứng viên

Nén CSDL thành cấu trúc cây FP

Duyệt đệ qui cây để tạo tập phổ biến

FP-GROWTH

Khi một FP-Tree đã được xây dựng, FP-Growth sử dụng cách tiếp cận chia để trị đệ quy để khai thác các tập phổ biến.

- **Với mỗi giao dịch, FP-Tree xây dựng một đường đi (path) trong cây.**
- **Hai giao dịch có chứa cùng một số các mục, thì đường đi của chúng sẽ có phần (đoạn) chung.**

FP-GROWTH

B0

Thiết lập cây FP

B1

Thiết lập cơ sở mẫu điều kiện cho mỗi hạng mục phổ biến (mỗi nút trên cây FP

B2

Thiết lập cây FP điều kiện từ mỗi cơ sở mẫu điều kiện

B3

Khai thác đệ qui cho cây FP – điều kiện

FP-GROWTH

B1

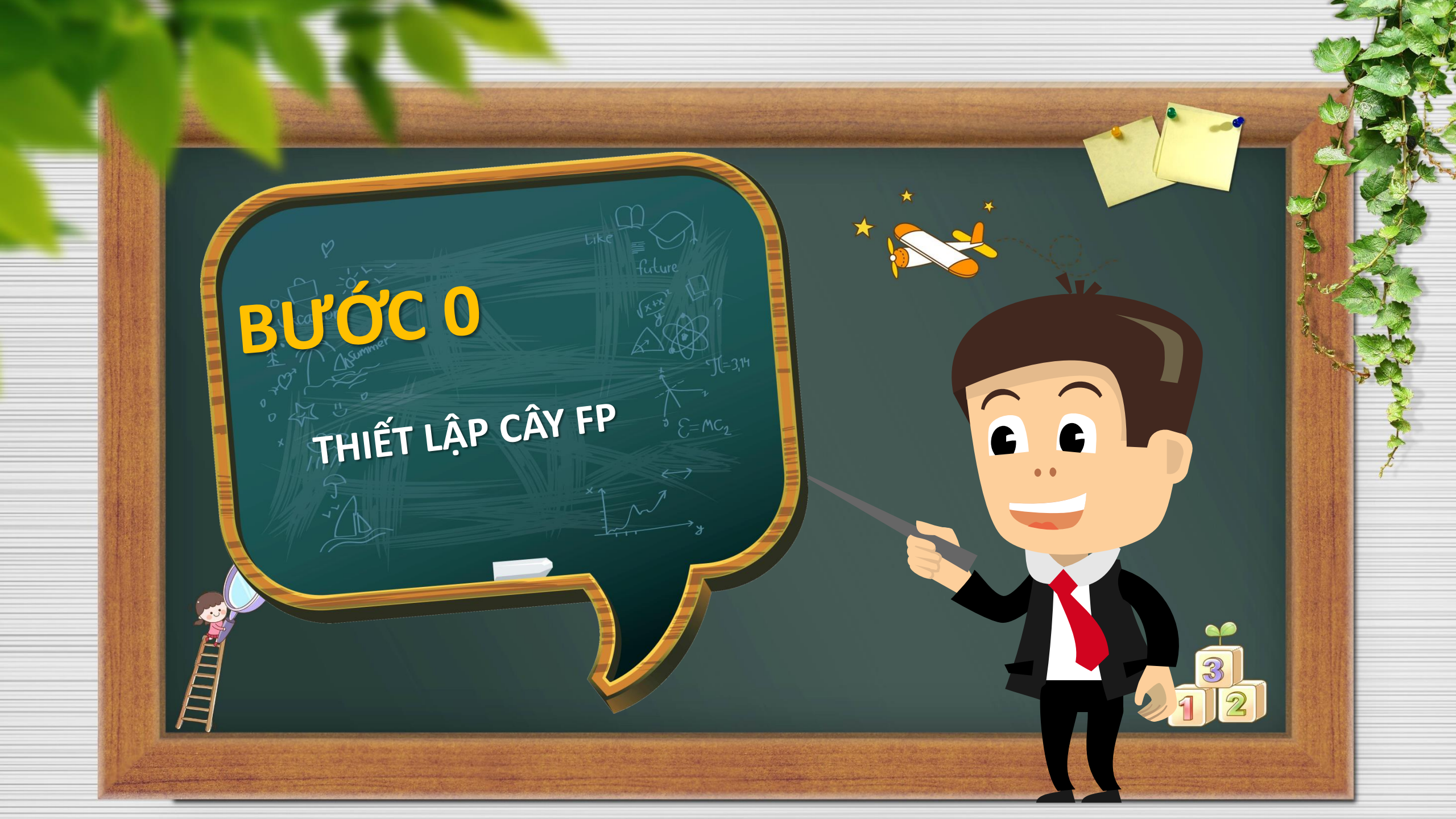
Ban đầu, FP-Tree chỉ chứa duy nhất nút gốc (được biểu diễn bởi ký hiệu null). C

B2

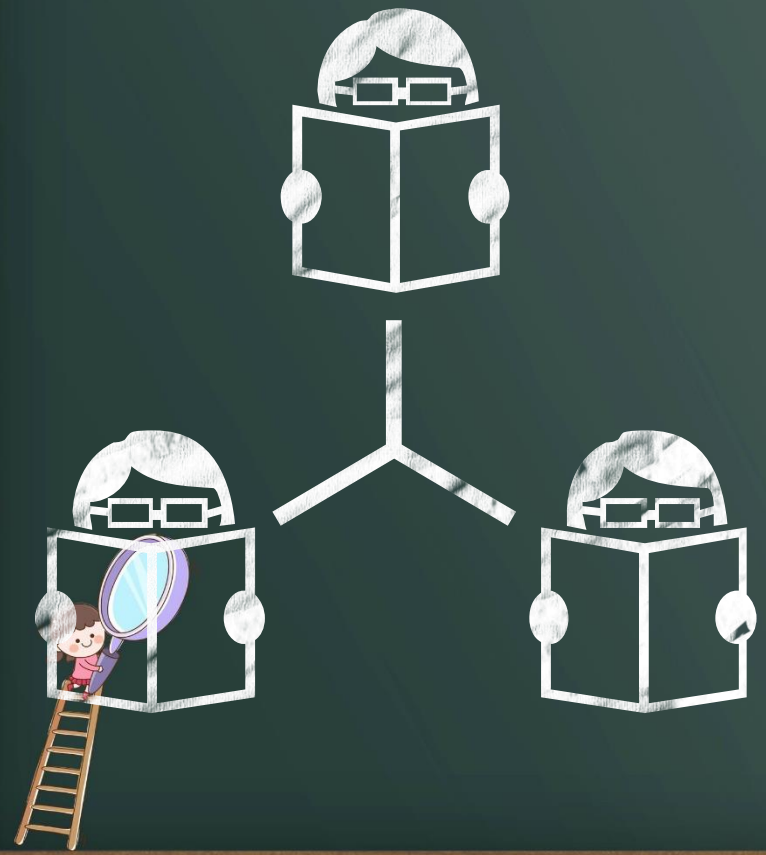
Các mục không thường xuyên *(infrequent items)* bị loại bỏ.

B3

Các mục thường xuyên *(frequent items)* được sắp xếp theo thứ tự giảm dần về độ hỗ trợ.


BƯỚC 0
THIẾT LẬP CÂY FP

QUI TRÌNH THIẾT LẬP



1

Tìm tập phổ biến 1- hạng mục (duyệt csdl lần 1)

2

Sắp xếp tập phổ biến giảm dần vào trong ds F-list

3

Sắp xếp lại từng giao dịch của csdl theo danh sách F-list *(loại bỏ các hạng mục không có trong F-list)*

4

Duyệt csdl lần nữa và thiết lập cây Fp

FP GROWTH

Minsupp=60%

(60x5)/100 = 3

TID	Items	
1	f, a, c, d, g, i, m, p	
2	a, b, c, f, l, m, o	
3	b, f, h, j, o	
4	b, c, k, s, p	
5	a, f, c, e, l, p, m, n	

LỌC TẦN SUẤT

Items tần suất xuất hiện >=3

SẮP XẾP

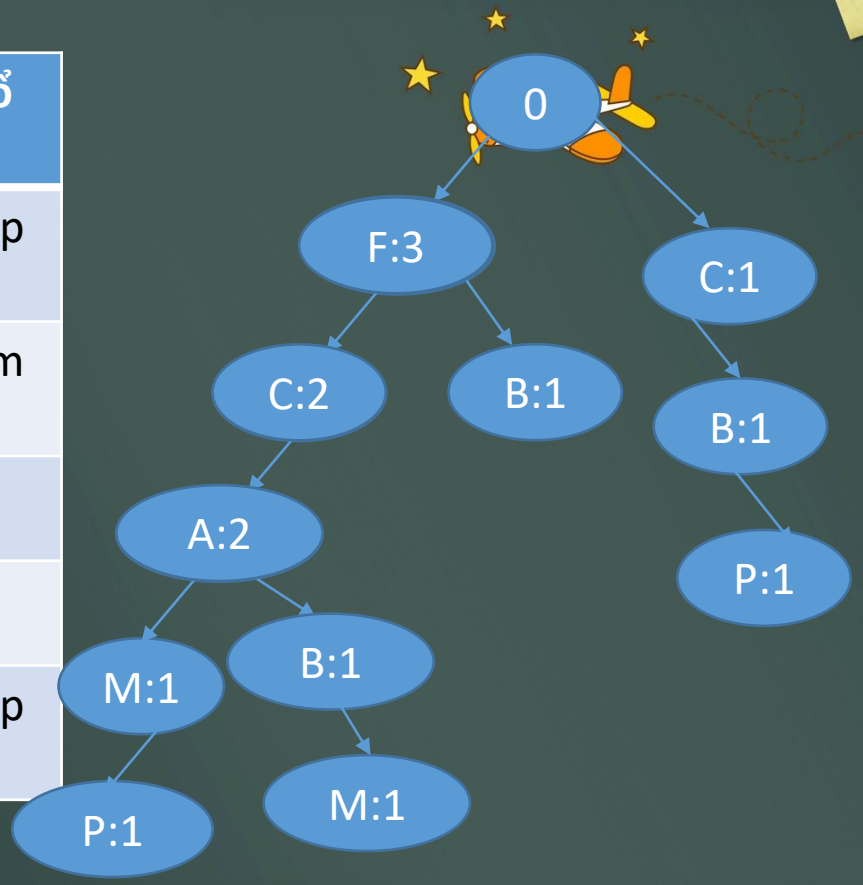
Sắp xếp giảm dần tần xuất xuất hiện

ĐƯA RA F-List CỦA GIAO DỊCH

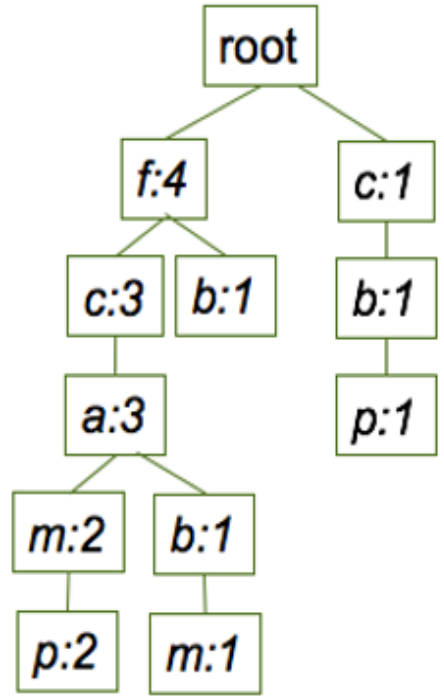
VD : f->c->a->b->m->p

FP GROWTH

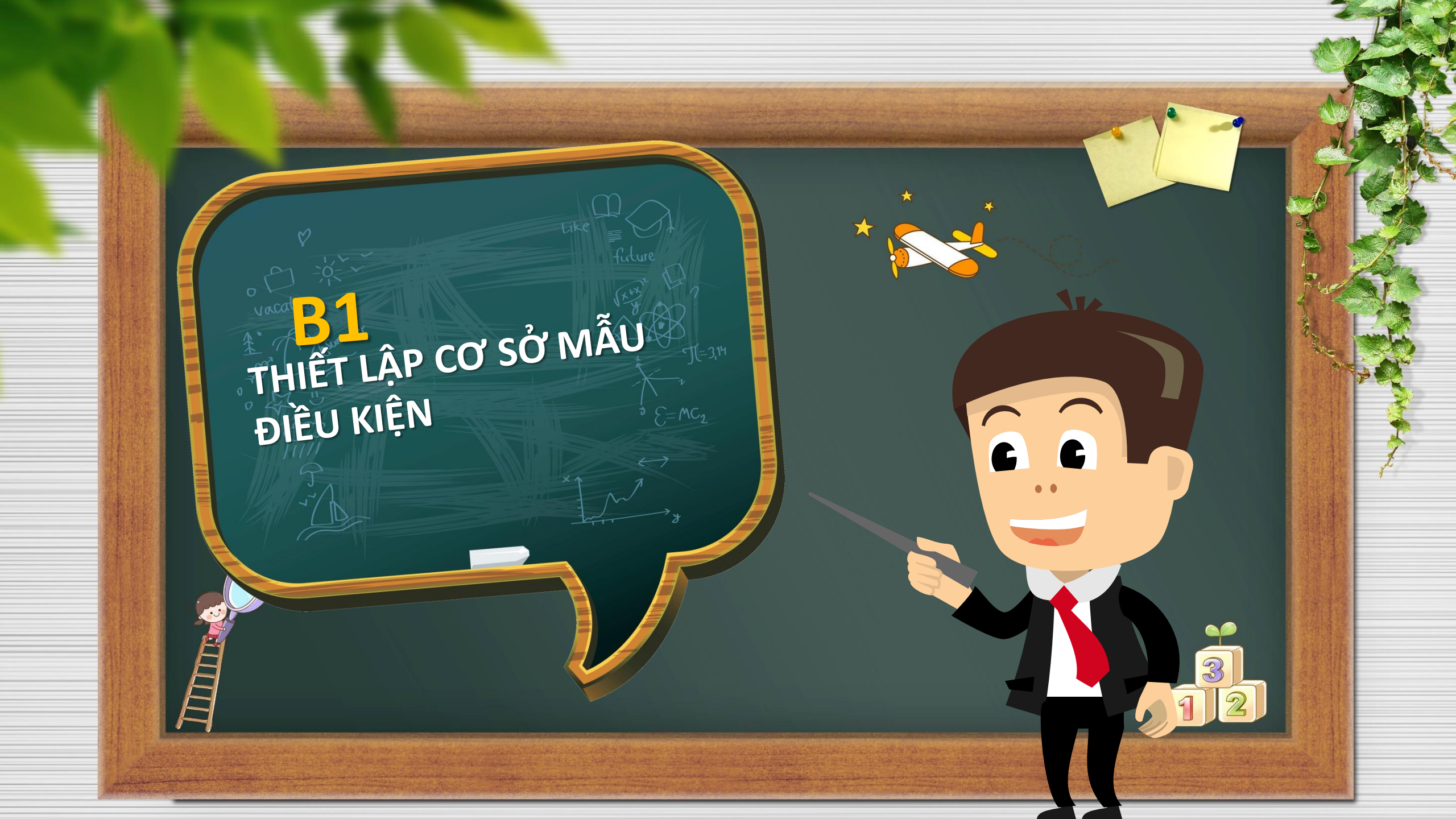
TID	Items	Items phổ biến
1	f, a, c, d, g, i, m, p	f, c, a, m, p
2	a, b, c, f, l, m, o	f, c, a, b, m
3	b, f, h, j, o	f, b
4	b, c, k, s, p	c, b, p
5	a, f, c, e, l, p, m, n	f, c, a, m, p



XÂY DỰNG CÂY



B1
THIẾT LẬP CƠ SỞ MẪU ĐIỀU KIỆN



ADD TITLE TEXT

Gom tất cả đường dẫn tiền tố
Biến đổi của hạng mục để
tạo CSMDL

STEP 03

Duyệt cây FP theo kết nối
Của mỗi hạng mục phổ biến

STEP 02

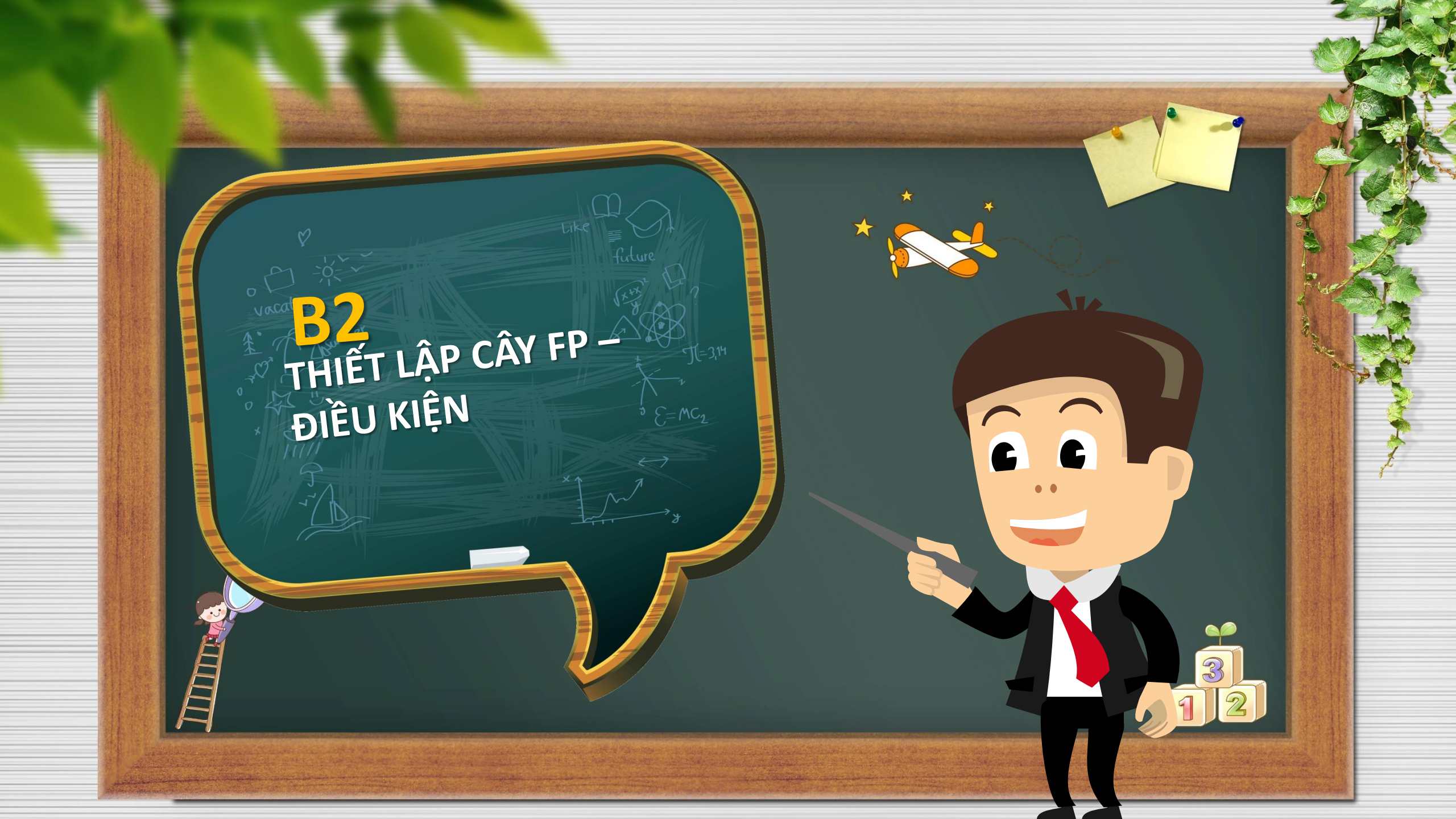
Bắt đầu từ mẫu phổ biến
cuối bảng của cây fp

STEP 01

Xây dựng cơ sở mẫu điều kiện

Conditional pattern base

B2
THIẾT LẬP CÂY FP – ĐIỀU KIỆN



Xây dựng cây FP- điều kiện

Với mỗi cơ sở mẫu

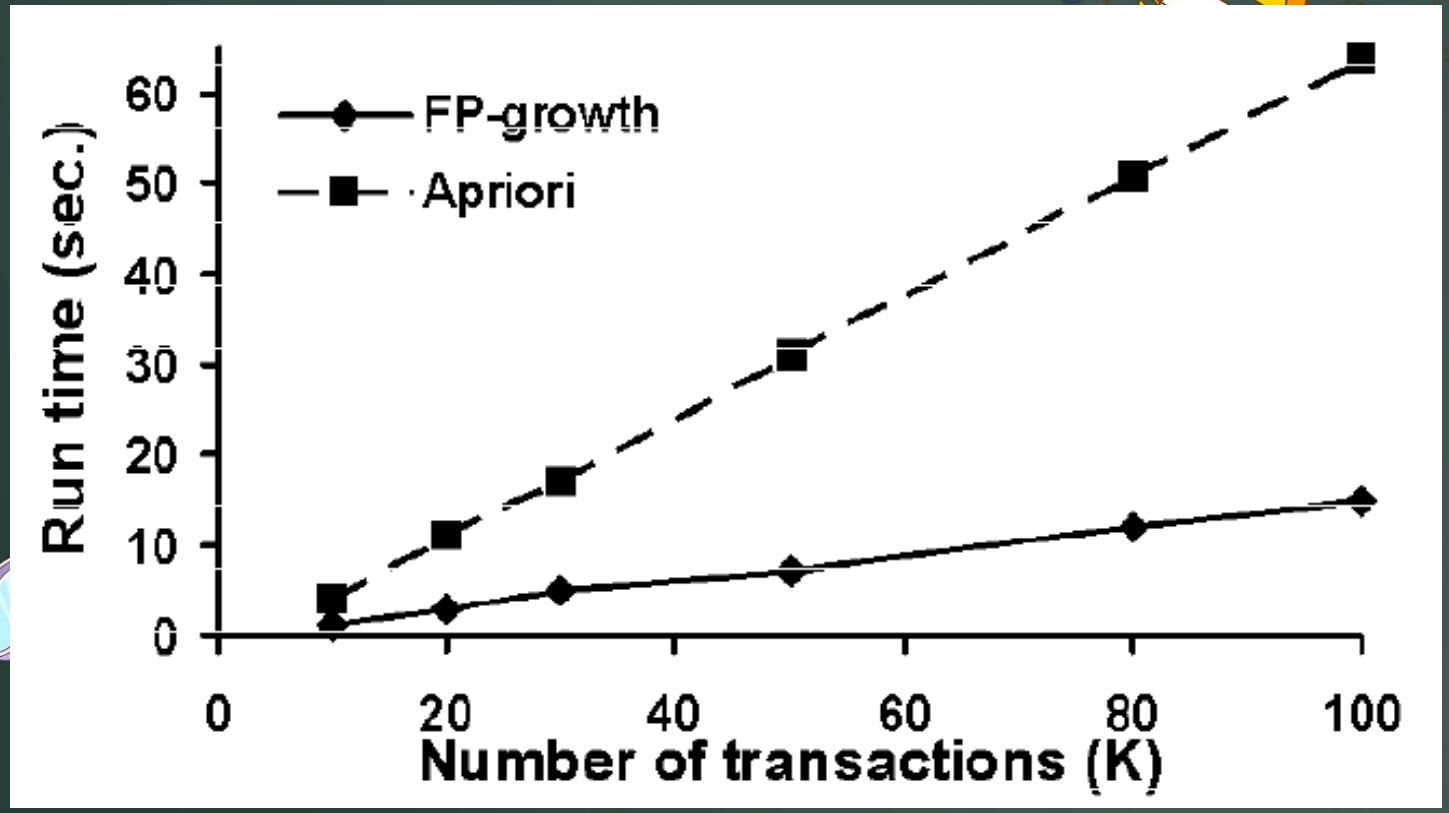
Đếm số lượng mỗi mẫu trong cơ sở mẫu.
Xác định tập phổ biến của mẫu cơ sở

Xây dựng cây FP – điều kiện cho tập phổ biến của mẫu cơ sở (tương tự như bước 0)

ƯU ĐIỂM

- **Hiệu quả hơn so với Apriori.**
- **Phân chia và kiểm soát quá trình xử lý.**
- **Sử dụng cây FP để biểu diễn các mẫu phổ biến thì dữ liệu giảm rất đáng kể so với cách biểu diễn trong CSDL.**

SO SÁNH FP VÀ Apriori



Khuyết điểm

Không thể xây dựng cây FP trong bộ nhớ chính khi CSDL là lớn.

Đầu tiên phân chia CSDL vào trong các projected database và sau đó xây dựng một cây FP và khai thác cây này trong mỗi projected database.

Tiếp theo

Giải thuật FP-Growth (bài tập)

Cho CSDL gồm 5 giao dịch tìm các mẫu phổ biến

ID	ITEMSETS
100	M,O,N,K,E,Y
200	D,O,N,K,E,Y
300	M,A,K,E
400	M,U,C,K,Y
500	C,O,O,K,I,E

MIN_SUP = 3

ITEM	SUP
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

Loại bỏ những tần suất nhỏ hơn 3

Giải thuật FP-Growth (bài tập)

ITEM	SUP
K	5
E	4
M	3
O	3
Y	3

Sau khi loại bỏ các mục có tần suất nhỏ hơn 3

ID	ITEMSETS	ITEM PHỔ BIẾN
100	M,O,N,K,E,Y	K,E,M,O,Y
200	D,O,N,K,E,Y	K,E,O,Y
300	M,A,K,E	K,E,M
400	M,U,C,K,Y	K,M,Y
500	C,O,O,K,I,E	K,E,O

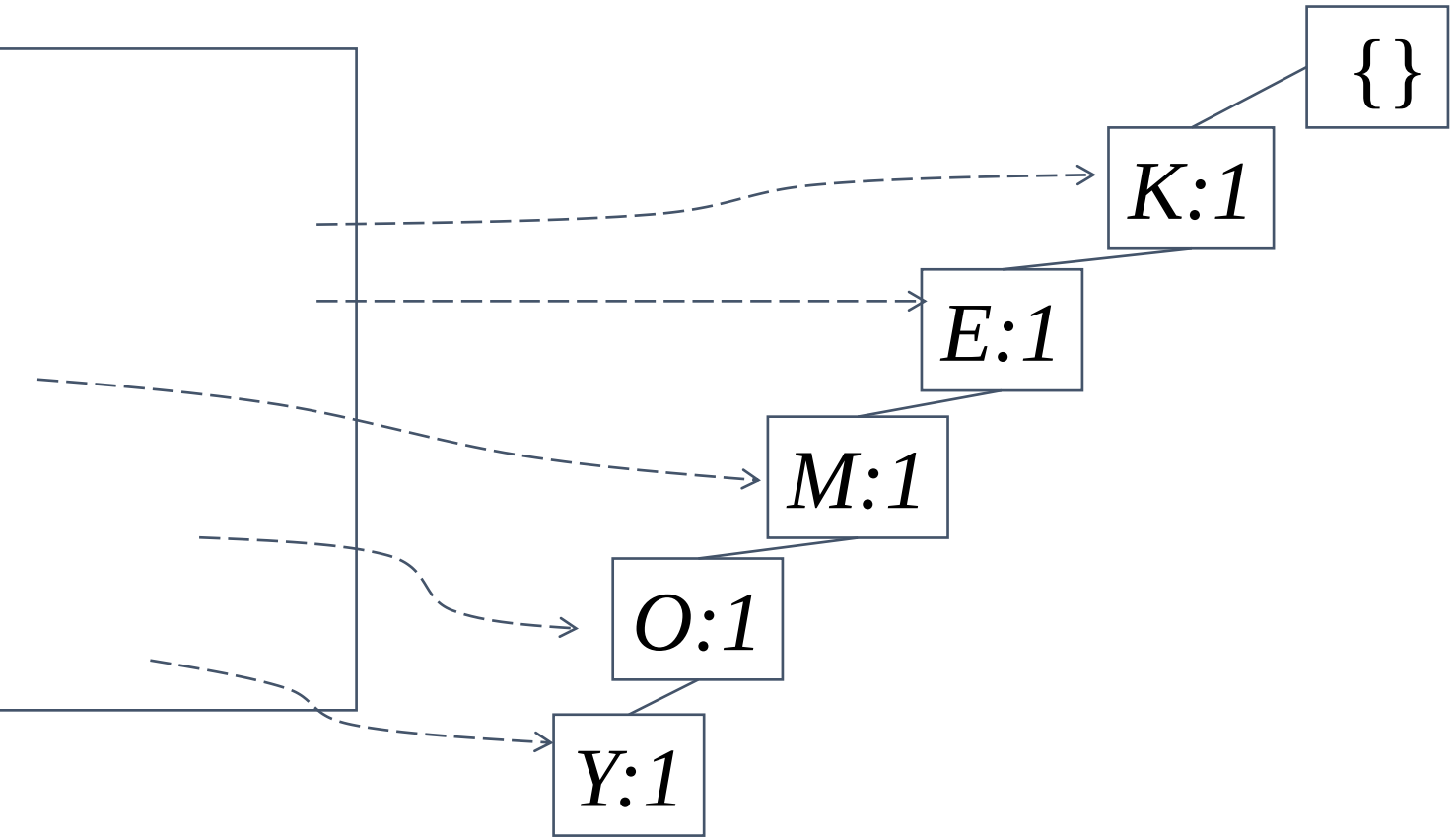
Giải thuật FP-Growth

B1: Xây dựng FP tree:

Min sup = 3

Bảng

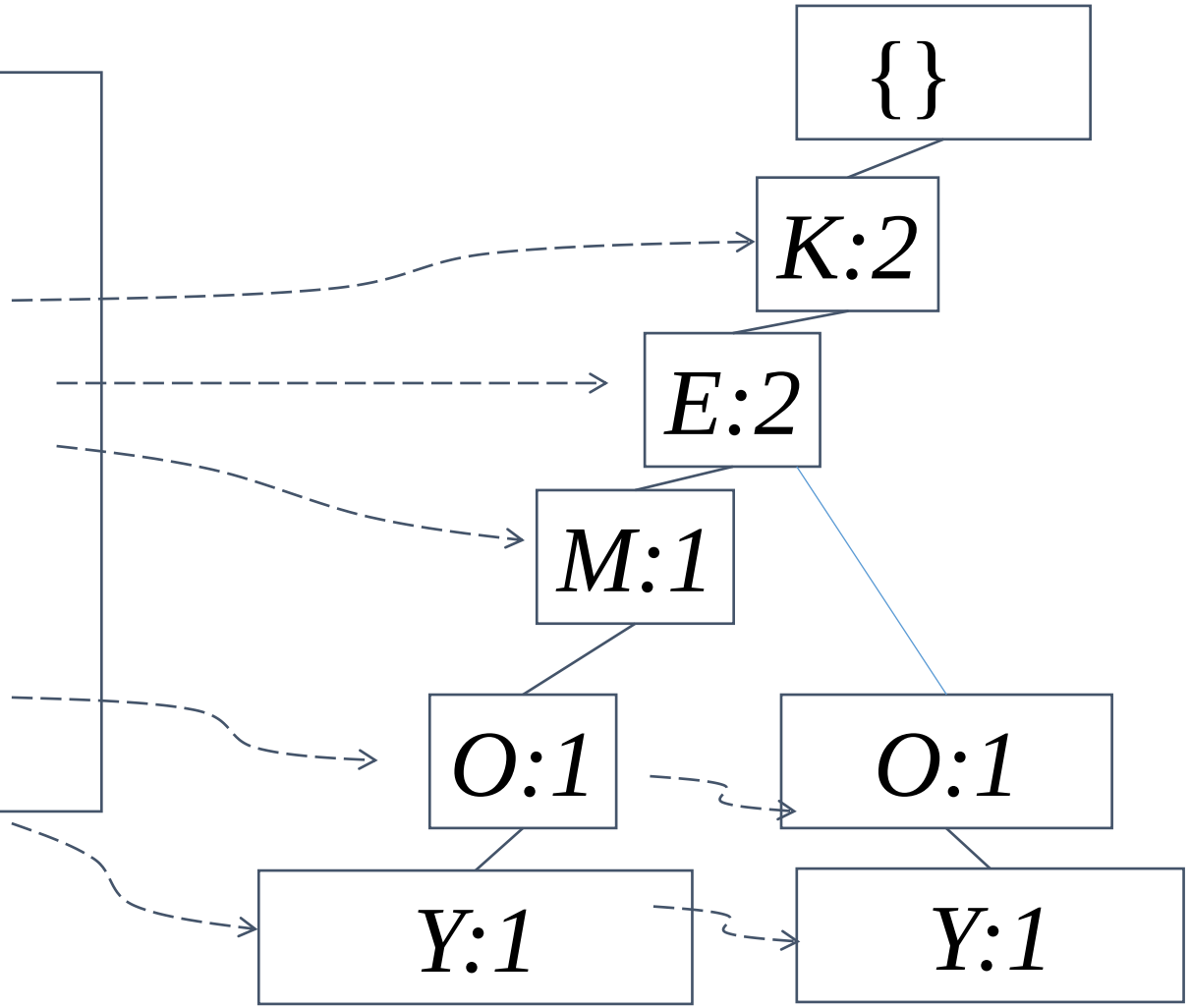
Item	*Tần số*
K	*5*
E	*4*
M	*3*
O	*3*
Y	*3*



Giải thuật FP-Growth

Bảng

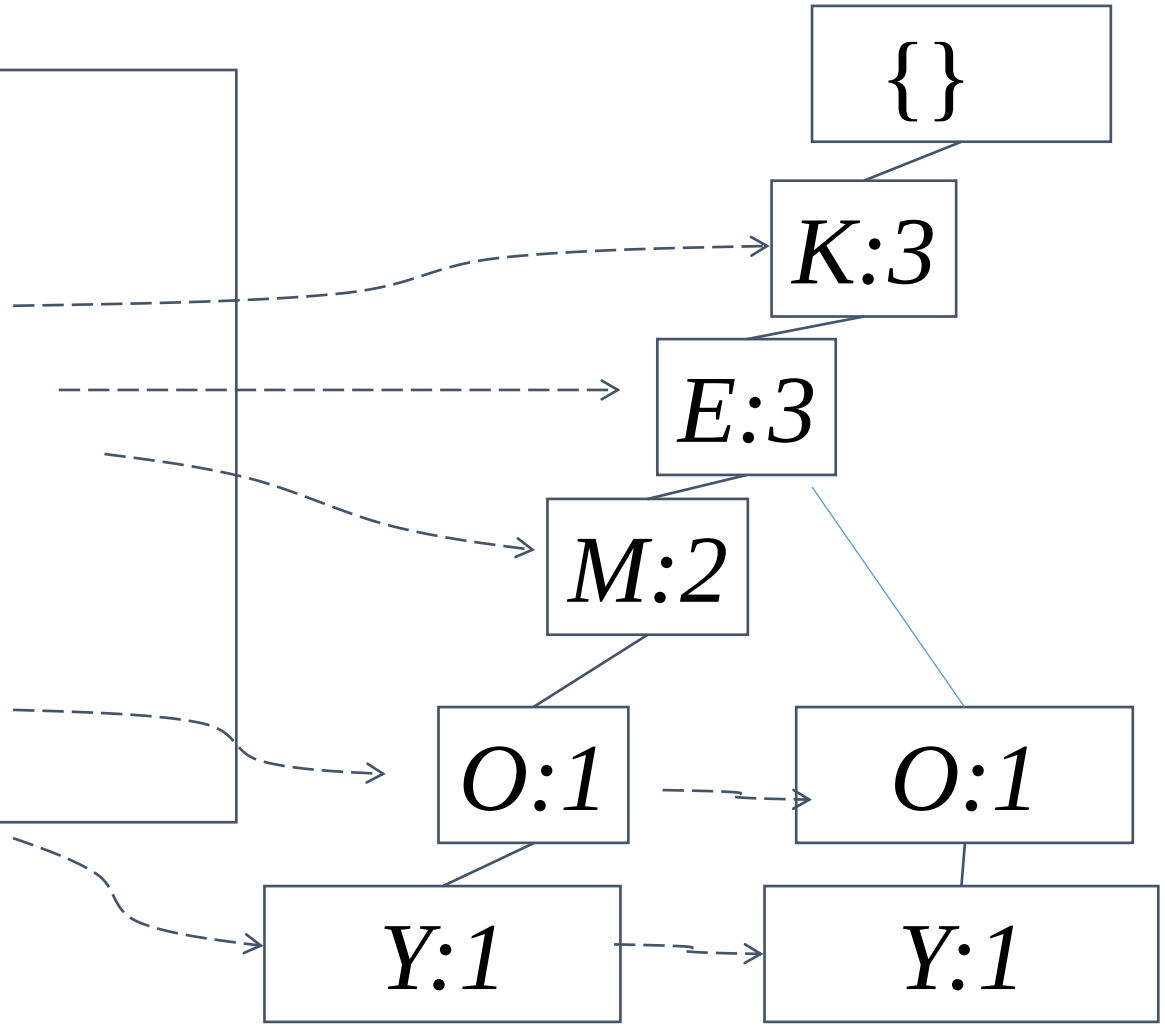
Item	*Frequency*
K	*5*
E	*4*
M	*3*
O	*3*
Y	*3*



Giải thuật FP-Growth

Bảng

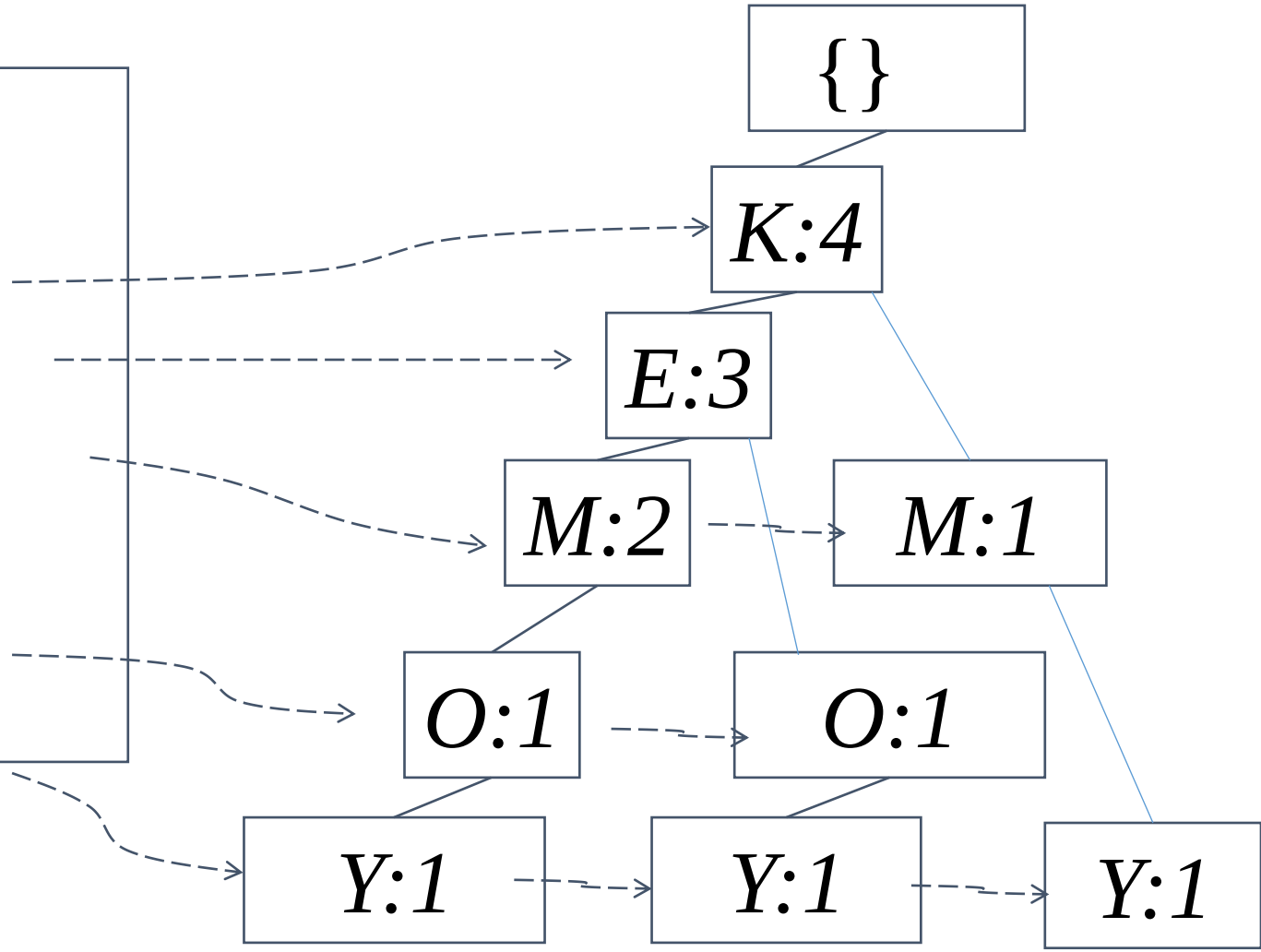
Item	***Frequency***	***Head***
K	*5*	
E	*4*	
M	*3*	
O	*3*	
Y	*3*	



Giải thuật FP-Growth

Bảng

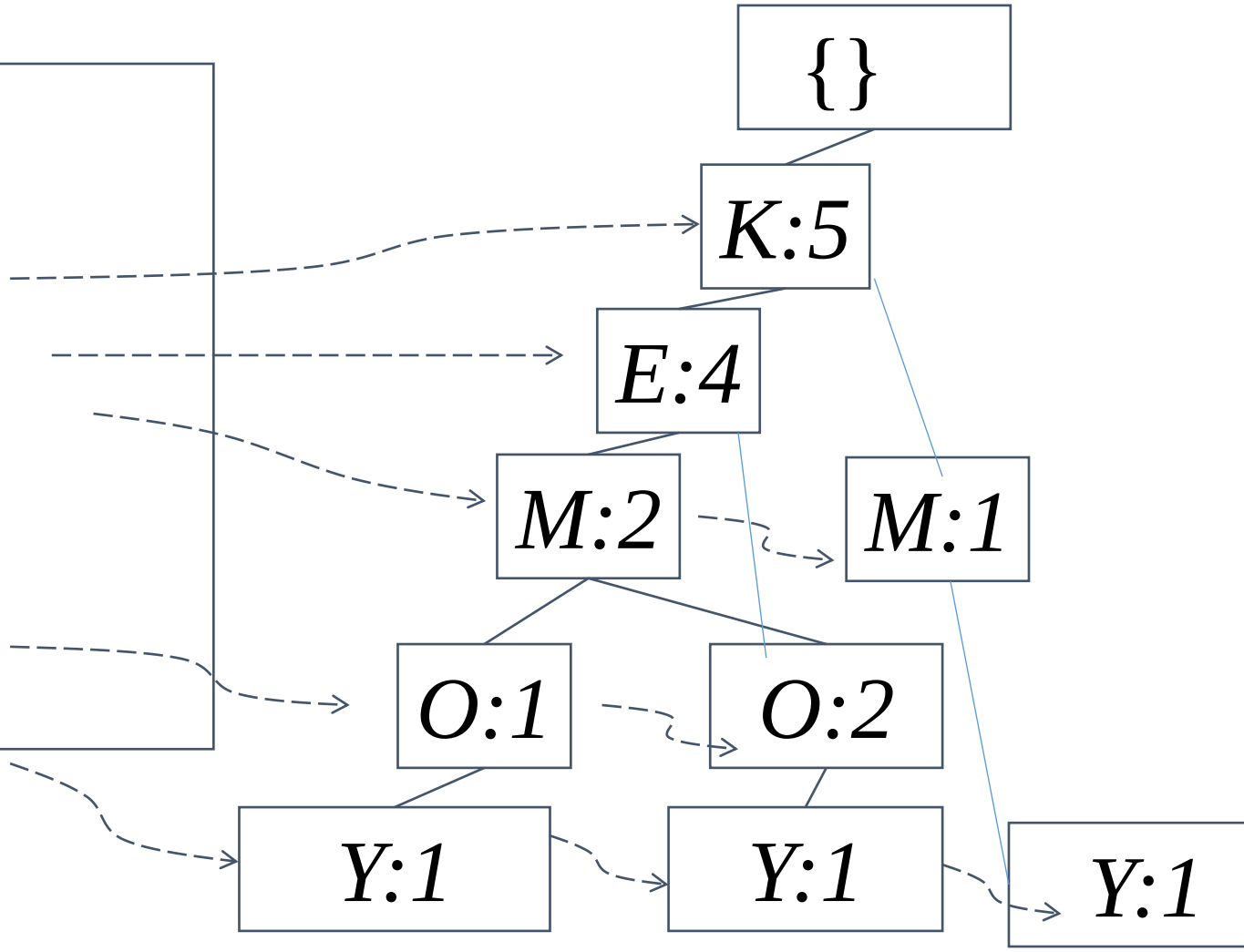
Item	***Frequency***	***Head***
K	*5*	
E	*4*	
M	*3*	
O	*3*	
Y	*3*	



Giải thuật FP-Growth

Bảng

Item	***Frequency***	***Head***
K	*5*	
E	*4*	
M	*3*	
O	*3*	
Y	*3*	



ID	Cơ sở mẫu đk	FP-Tree đk	Các mẫu phổ biến
Y	{KEMO : 1 , KEO: 1 KM: 1}	{K:3}-Y	Y,KY
O	{KEM:1, KE:2}	{K:3,E:3}-O	O,KO,EO,KEO
M	{KE:2, K:1}	{K:3}-M	M,KM
E	{K:4}	{K:4}-E	E,KE
K	Ø	Ø	K