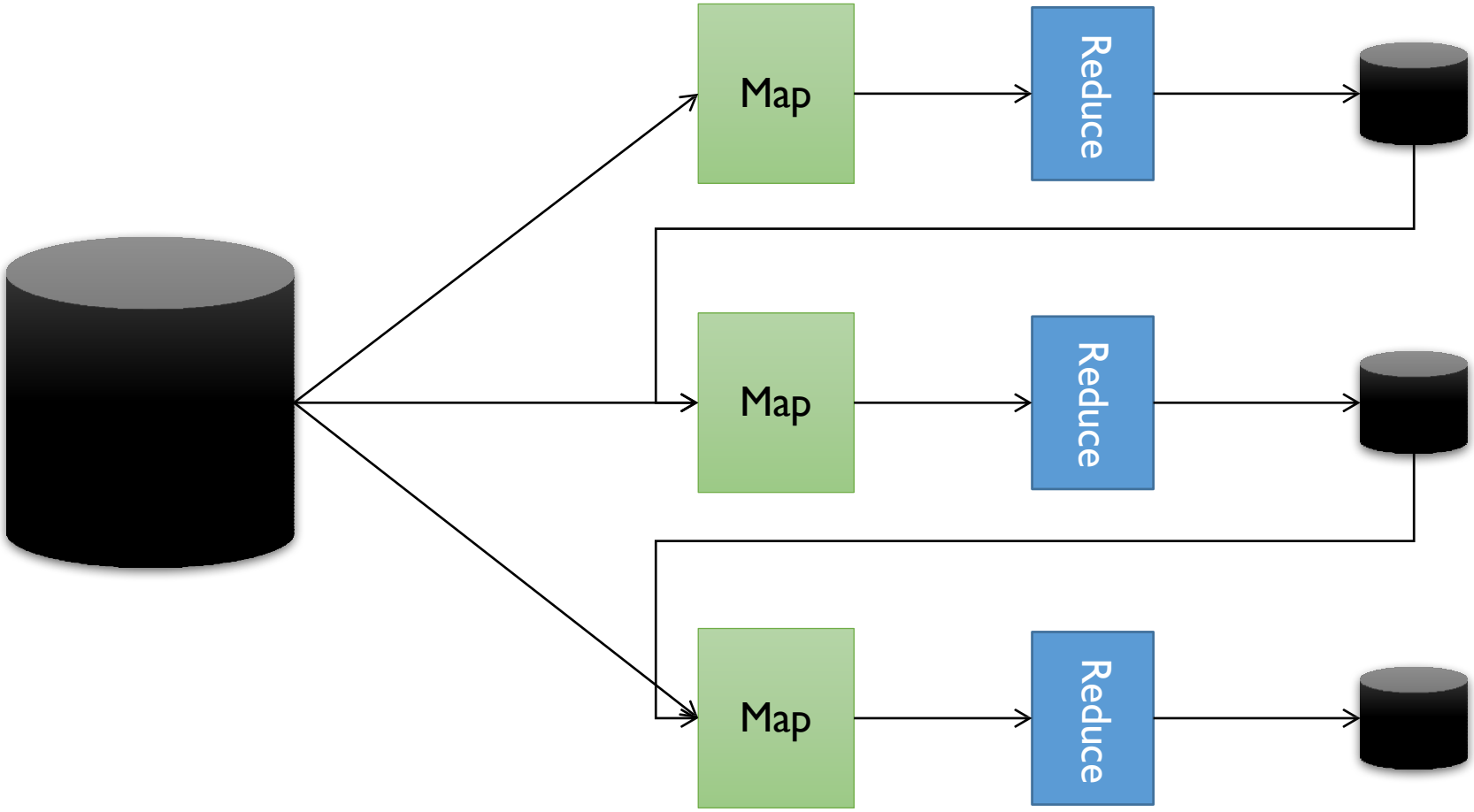
Data Engineering

MG-GY 8441

Distributed File Systems



Market Basket Analysis

Agenda

- Frequent Items
- Association Rules
- Metrics

References

- Han, Kamber, Pei, *Data Mining: Concepts and Techniques* (Chapter 6.1, 6.2)
- Optional
  - Hand, Mannila, Smyth *Principles of Data Mining* (Chapter 13)

Example

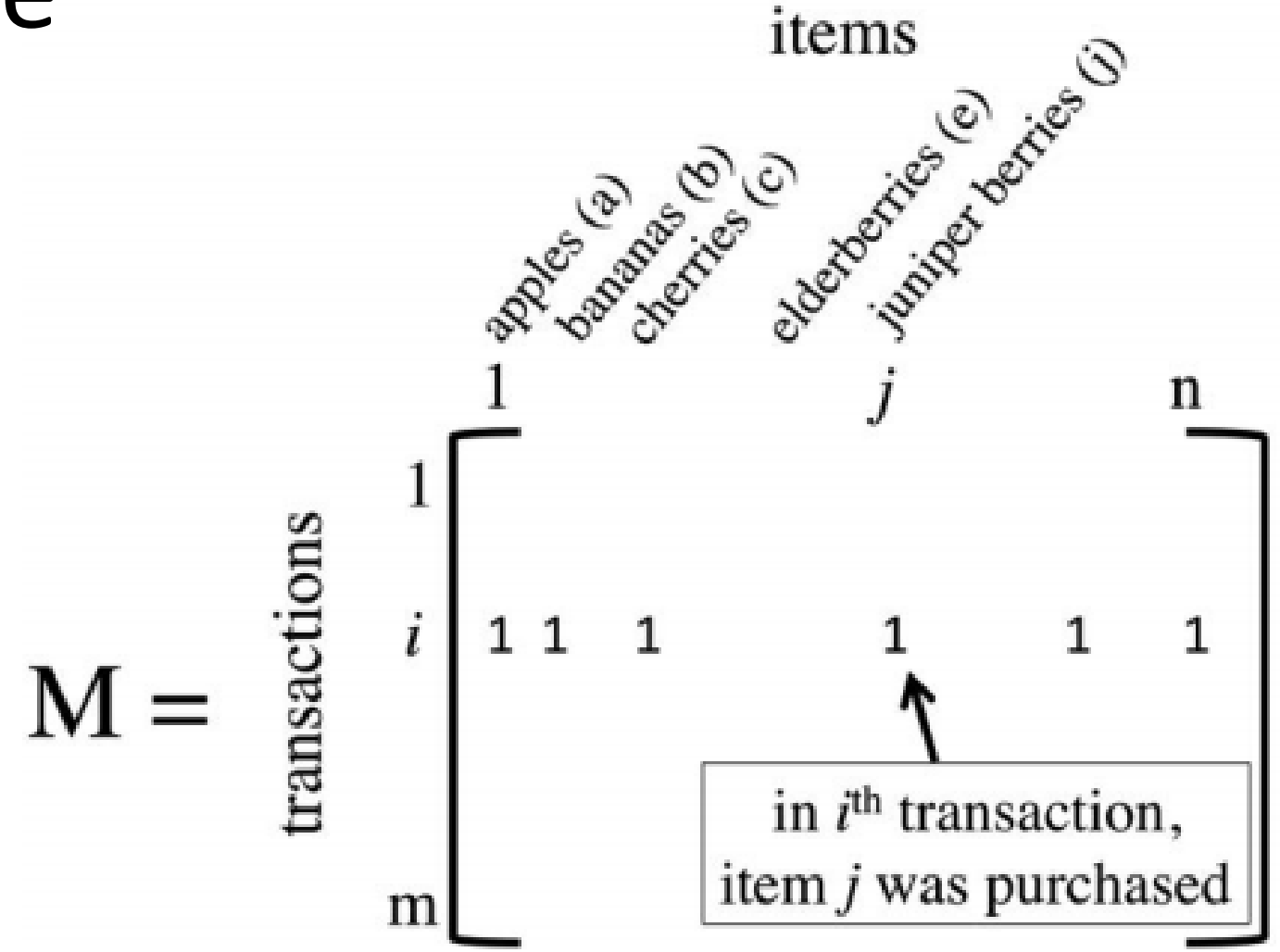
Example from Marketing

Suppose you are a business analyst within the marketing division of your company.

Your group has been studying customer records to determine patterns in transactions.

You have access to a data warehouse indicating purchases of items from the company inventory over time.

Example


items
apples (a)
bananas (b)
cherries (c)
elderberries (e)
juniper berries (j)
1
j
n
M =
transactions
1
i
m
1 1 1
1
1 1
in i^{th} transaction,
item j was purchased

Example

Example from Marketing

You want to determine frequently occurring collections of items

{apples, cherries, bananas}

From the frequently occurring collections of items, you want to find association rules

{bananas, cherries} imply {apples}

These rules could inform discount programs, store layout, catalogue design,…

Terminology

- **Support** of an itemset: number of transactions containing it,

$$\text{Supp}(\text{bananas, cherries, elderberries}) = \sum_{i=1}^{m} M_{i,2} \cdot M_{i,3} \cdot M_{i,5}.$$

- **Confidence** of rule $a \to b$: the fraction of times itemset b is purchased when itemset a is purchased.

$$\begin{aligned} \text{Conf}(a \to b) &= \frac{\text{Supp}(a \cup b)}{\text{Supp}(a)} = \frac{\#\text{times } a \text{ and } b \text{ are purchased}}{\#\text{times } a \text{ is purchased}} \\ &= \hat{P}(b|a). \end{aligned}$$

- **Itemset**: a subset of items, e.g., (bananas, cherries, elderberries), indexed by $\{2, 3, 5\}$.

Calculating Support

	apples	*bananas*	*cherries*		*elderberries*		*grapes*
1-itemsets:	a	b	c	d	e	~~f~~	g
supp:	25	20	30	45	29	5	17

2-itemsets:	~~{a,b}~~	{a,c}	{a,d}	{a,e}	...	~~{e,g}~~
supp:	7	25	15	23		3

3-itemsets:	{a,c,d}	{a,c,e}	{b,d,g}	...
supp:	15	22	15	

4-itemsets:	{a,c,d,e}
supp:	12

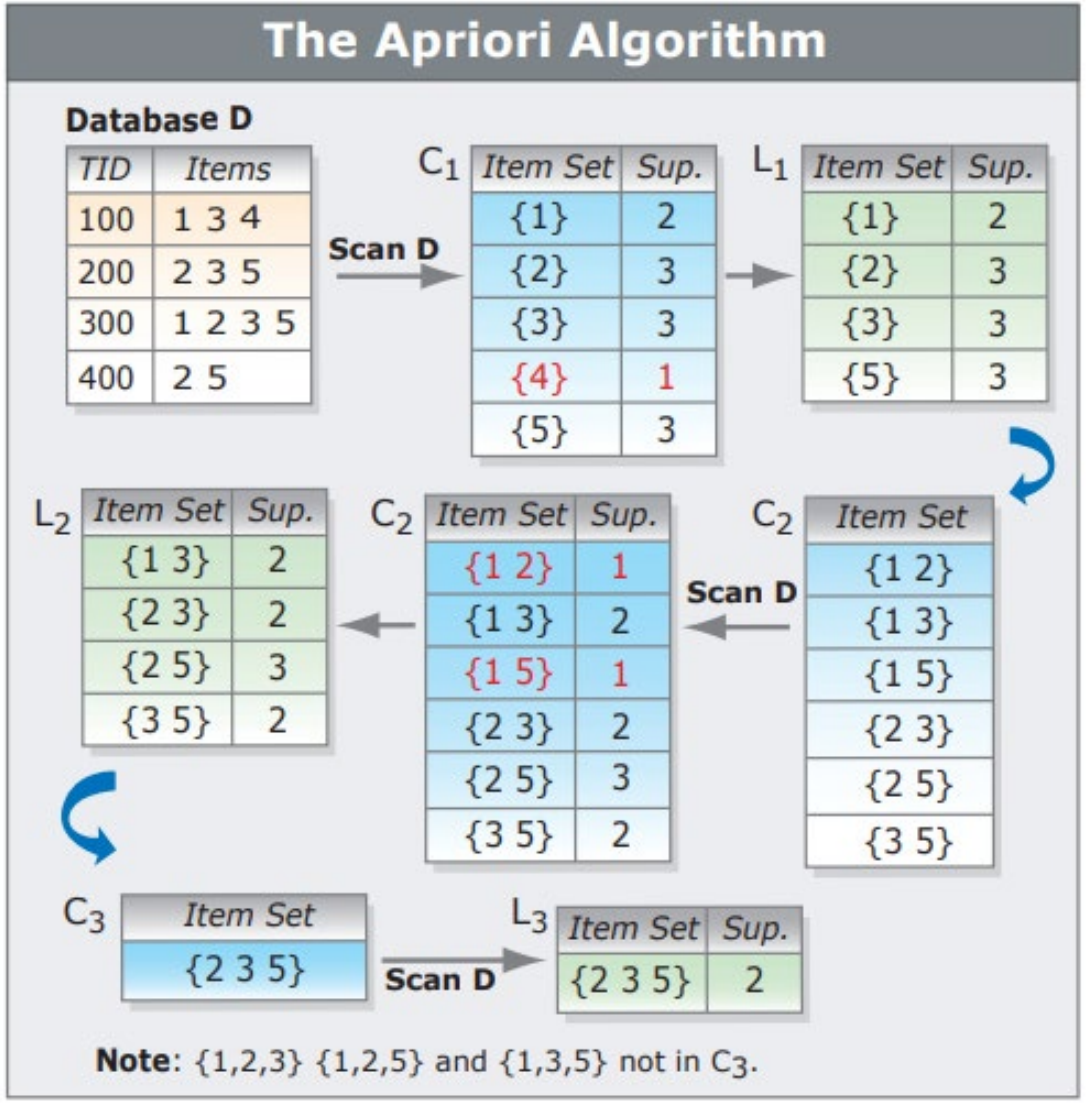
If $\text{Supp}(a \cup b) \geq \theta$ then $\text{Supp}(a) \geq \theta$ and $\text{Supp}(b) \geq \theta$.

Calculating Confidence

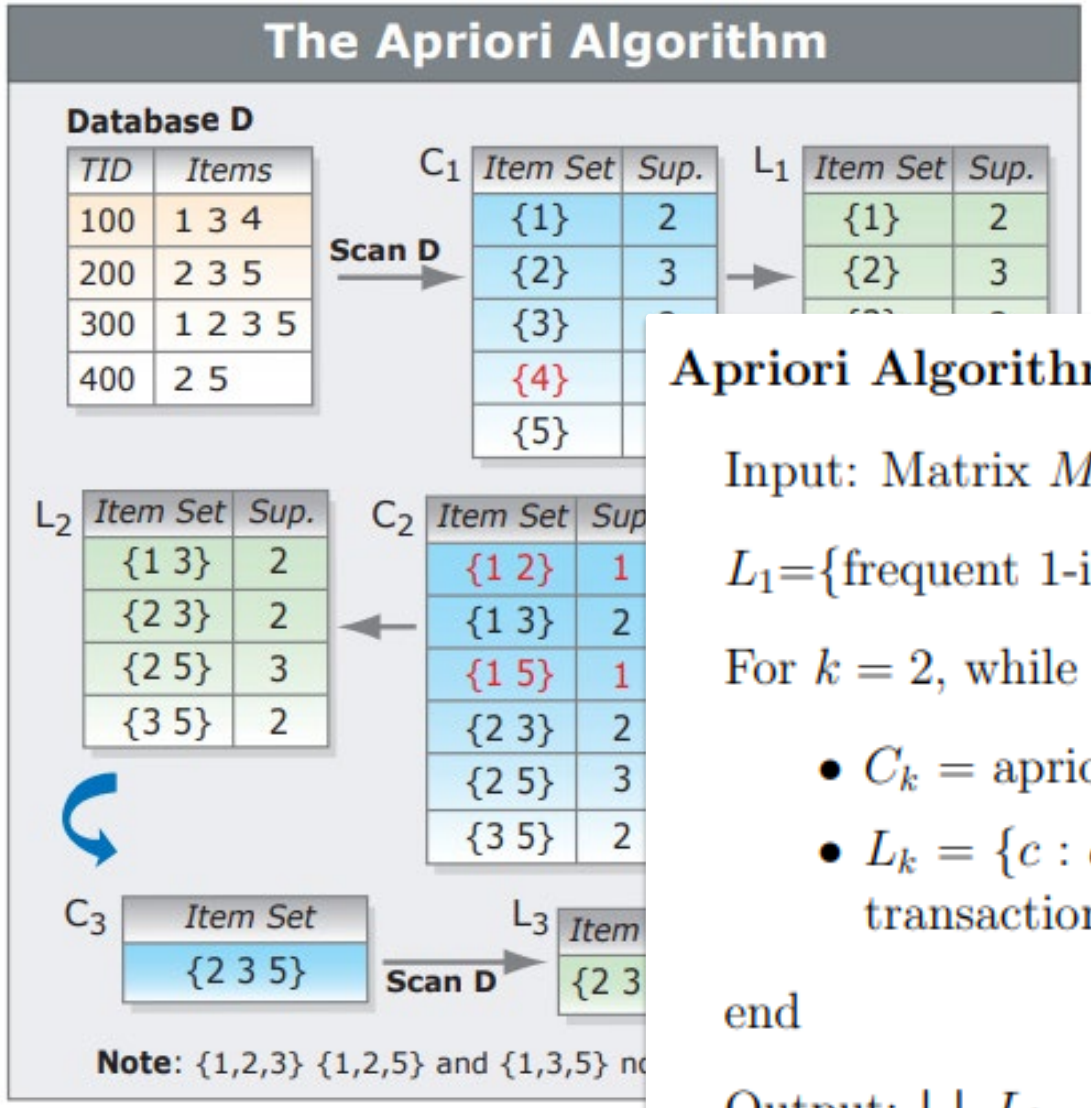
- For each frequent itemset ℓ:
  - Find all nonempty subsets of ℓ
  - For each subset a, output $a \rightarrow \{\ell \setminus a\}$ whenever

$$\frac{\text{Supp}(\ell)}{\text{Supp}(a)} \geq \varepsilon$$

Apriori Algorithm



Apriori Algorithm



Apriori Algorithm:

Input: Matrix M

L_1={frequent 1-itemsets; i such that Supp$(i) \geq \theta$}.

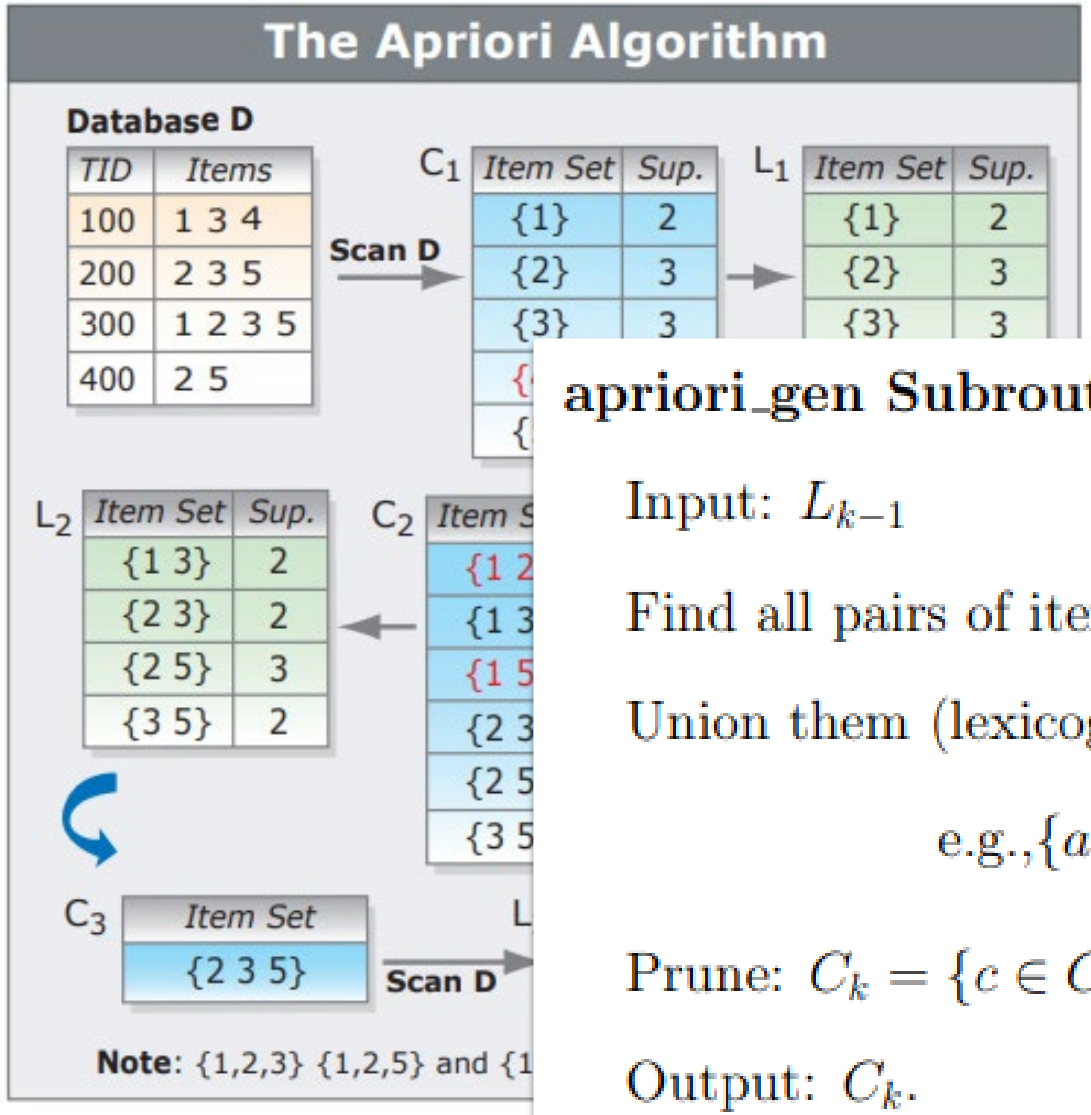
For $k = 2$, while $L_{k-1} \neq \emptyset$ (while there are large $k-1$-itemsets), $k++$

- C_k = apriori_gen(L_{k-1}) generate candidate itemsets of size k
- $L_k = \{c : c \in C_k,$ Supp$(c) \geq \theta\}$ frequent itemsets of size k (loop over transactions, scan the database)

end

Output: $\bigcup_k L_k$.

Apriori Algorithm



apriori_gen Subroutine:

Input: L_{k-1}

Find all pairs of itemsets in L_{k-1} where the first $k-2$ items are identical.

Union them (lexicographically) to get $C_k^{\text{too big}}$,

$$\text{e.g.,} \{a,b,c,d,e,f\}, \{a,b,c,d,e,g\} \rightarrow \{a,b,c,d,e,f,g\}$$

Prune: $C_k = \{c \in C_k^{\text{too big}}, \text{all } (k-1)\text{-subsets } c_s \text{ of } c \text{ obey } c_s \in L_{k-1}\}$.

Output: C_k.

Eclat Algorithm

Items	TID
1	100,300
2	200,300,400
3	100,200,300
4	100
5	200,300,400

C_1

Item Set	Sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

Item Set	Sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Item Set	TID
{1,2}	300
{1,3}	100,300
{1,5}	300
{2,3}	200,300
{2,5}	200,300,400
{3,5}	200,300

C_2

Item Set	Sup.
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

L_2

Item Set	Sup
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2

C_3

Item Set	TID
{1,2,3}	300
{1,3,5}	300
{2,3,5}	200,300

C_3

Item Set	Sup
{1,2,3}	1
{1,3,5}	1
{2,3,5}	200,300

L_3

Item Set	Sup
{2,3,5}	200,300

Metrics

$$\text{Lift}(a \rightarrow b) = \frac{\text{Supp}(a \cup b)}{\text{Supp}(a) \cdot \text{Supp}(b)} = \frac{\text{Conf}(a \rightarrow b)}{\text{Supp}(b)}$$

$$\text{Conviction}(a \rightarrow b) = \frac{1}{\text{Lift}(a \rightarrow \text{not } b)}$$