

- 1) Suppose that a data warehouse consists of the three dimensions *time*, *doctor*, and *patient*, and the two measures *count* and *charge*, where charge is the fee that a doctor charges a patient for a visit.
 - a) Draw a star schema diagram for this data warehouse.
 - b) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2020?
- 2) A database has five transactions. Let $\text{min_sup} = 60\%$ and $\text{min_conf} = 80\%$.

<i>TID</i>	<i>items_bought</i>
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

- 3) The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, and hamburgers refers to the transactions containing hamburgers.

	<i>hot dogs</i>	$\overline{\text{hot dogs}}$	Σ_{row}
<i>hamburgers</i>	2000	500	2500
$\overline{\text{hamburgers}}$	1000	1500	2500
Σ_{col}	3000	2000	5000

- a) Suppose that the association rule “hot dogs \Rightarrow hamburgers” is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?
- b) Calculate lift and chi-square on the given data.
- c) What kind of correlation relationship exists between the purchase of hot dogs and the purchase of hamburgers?