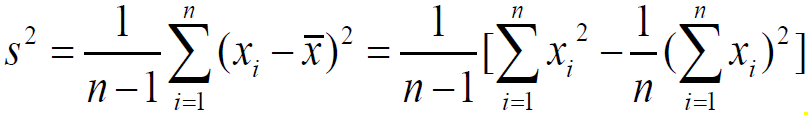
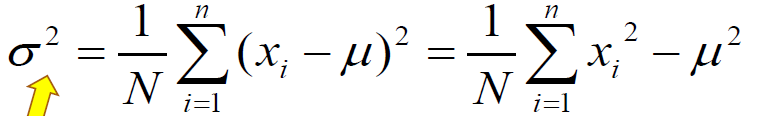
**Know your data(Chapter 2)**

**1.Type: Nominal：**定类 **Binary： Ordinal:** 定序 \*Numeric（**Interval-scaled：**定距，0不代表没有 **Ratio-scaled：**定比，0代表没有）

**2.Unimodal 单峰：**mean-mode=3\*(mean-median)

**3.Symmetric:** mode=median=mean

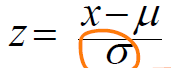
**positively skewed:** < < **negatively skewed:** > >

**4. Variance: sample(n): population(N):**

**5.Q1:**25% **Q3:**75% **Inter-quartile range:**IQR=Q3-Q1

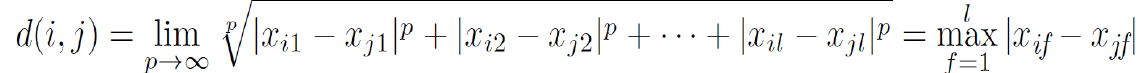
**6.Graphic Displays of Basic Statistical Descriptions: Boxplot:** min Q1 Q2/median Q3 max max-Q3/A1-min=whisker **Histogram:** x values, y frequencies, show distributions of variables & binned quantitative data. But bar charts: compare variables & categorical data **Quantile plot:** Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences). **Quantile-q plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another **Scatter plot:** Provides a first look at bivariate data to see clusters of points, outliers

**7.Data Visualization: \*Pixel oriented:** m dimensions with m windows. To save space and show the connections among multiple dimensions, circle segment **\*Geometric Projection: (Direct Data Visualization: Scatterplot Matrices(2D):** k\*(k-1)/2 **Landscape: Parallel Coordinates:** axes [minimum, maximum]**) \*Icon-based:** features of icons. Shape/color/tile bars(**Chernoff Faces/Stick Figures)** \***Hierarchical: (Dimensional Stacking**: Partitioning of the attribute value ranges into classes. **Words within worlds:** Assign the function and two most important parameters to innermost world **Tree-Map:** **InfoCube:)** **Tag cloud: Social Networks**

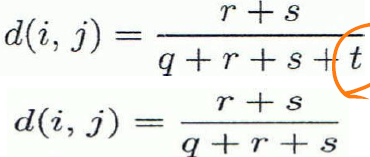
**8.Similarity: \*Z-score:**

**Minkowski Distance:**

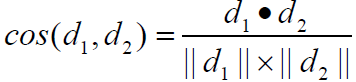
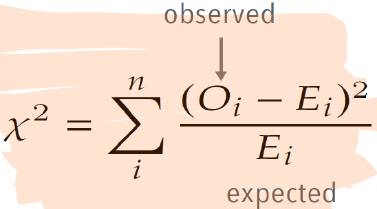
**Manhattan/City Block Distance:** p=1 **Euclidean Distance:** p=2

**Supremum D:**

**9.JaccardCoefficient/Coherence(i,j):**  j: 1 0

symmetric i:1 q r

asymmetric 0 s t

**Cosine Similarity:**

**Chi-square(X2):**

Null hypothesis: two distributions are independent 自由度=(行数-1）\*(列数-1)

查表得到不相关的概率，卡方值越大说明关联性越强。

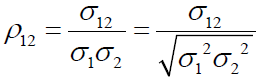
**10.Variance:**

Sample var: Est avg == real avg, N; Est avg != real avg, N-1

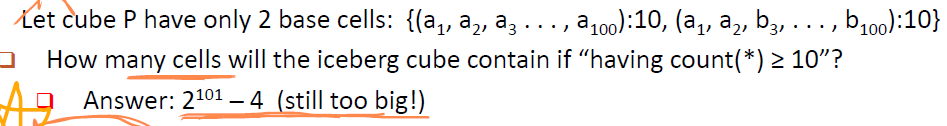
**Covariance:** 

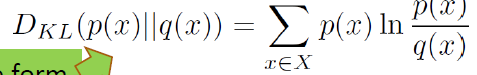
Sample cov: (Total Cov/n). >0, pos;独立则==0(反不可); <0, neg

**11. Pearson Correlation**: normalize covariance with standard deviation [-1, 1]

>0, pos correlated; ==0, independent;

<0, neg

**12.KL Divergence:** measure the difference between two probability distributions over the same x. D(p(X)||q(X)), P true/observation, 1 model/approximation of p, from q(prior( to p(posterior). D>=0 and only p==q,D == 0. P:0,D:0;q:0,D:正无穷

smoothing: not counting the possibility of unseen

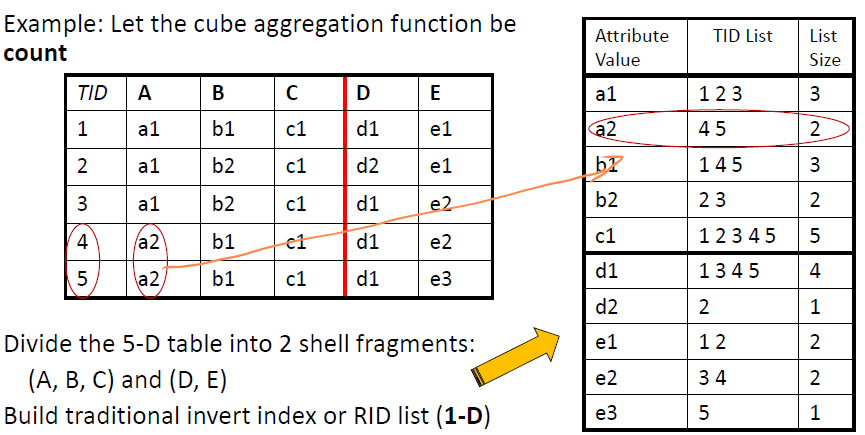
**\*\*Value Range\*\*** jacc\_coef,=[0,1]; cov=(-inf, inf); KL-div=[0, inf]; z=(-inf,inf); pearson cor coef=[-1, 1]; L\_inf=[0, inf]=[min(s(ij)), max(s(ij))] s(ij)=supdist(I,j)

**Data Processing(Chapter3)**

**1.Quality:** accuracy, completeness, consistency, timeliness, believability, interpretability

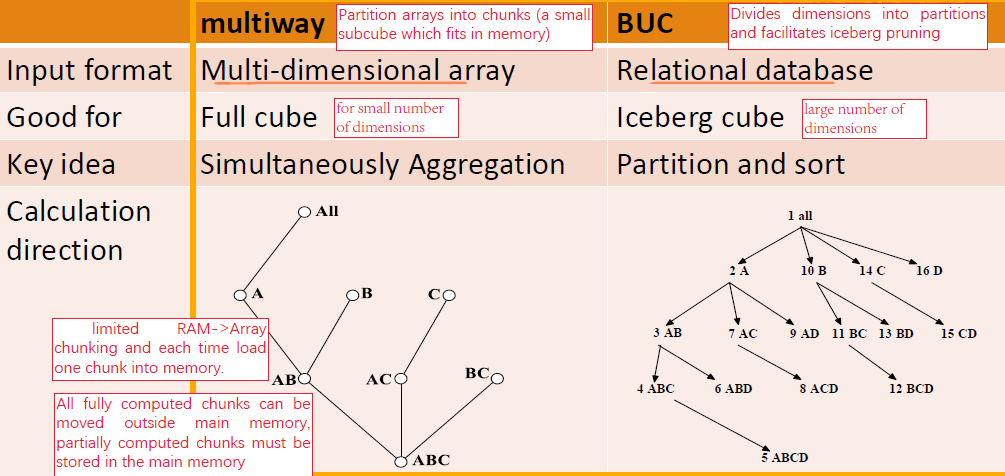
**Cleaning:** Incomplete-ignore, Noisy(Binning-Equal-frequency), Regression, Clustering-outlier), Inconsistent, Intentional;

**2.Integration:** schema integration( feature一样但不同名)

**3.Data Reduction: Parametric(可以用mode表示)/Non-parametric Methods; dependent/response variable, independent/explanatory variable; Regression/ Log-Linear/ Histogram Analysis/ Clustering Analysis/ Sampling(**random sampling-equal probability, without replacement-selected & removed, stratified-partition/cluster first then sampling)/ **Data Cube aggregation/ Data Compression****4. Data Transformation: Normalization(min-max/z-score/decimal scaling)**

**4.Discretization: Binning-**unsupervised (Equal-width/distance partition Equal-depth (frequency) partition, Top-down **Histogram**-un,td **Clustering**-un,td or bottom-up **Classification(Decision-Tree**-su,td) **Correlation(x^2)**-su, bu **Concept Hierarchy Generation**

**5.Dimensionality Reduction: Feature Selection:** subset; **Feature extraction: 降维**（**PCA:** linearly uncorrelated variables called principal components. Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space. **Attribute Subset Selection:** Redundant attributes. Irrelevant attributes. **Attribute Creation**: (**Attribute extraction-**domain-specific**, Mapping data to new space-Fourier/wavelet transformation, Attribute Construction**-Combining features**)**

**Data Warehousing and OLAP for Data Mining (Chapter4) 1.Basic Concepts:** Subject-oriented, Integrated, Time Variant, Nonvolatile( initial loading of data & access of data).

**2.Type:Enterprise warehouse:** collect all of the information about subject spanning the entire organization **Data Mart:** A subset of corporate-wide data that is of value to a specific groups of users **Virtual warehouse:** a set of views over operational databases

**3.Operations:Data extraction/Data cleaning/ Load/ Data transformation**(host format to warehouse format)/ **Refresh**

**4.Dimension Tables**-item; **Fact Table-**measures; **Data Cube**-(nD: base cuboid, 0D: apex cuboid)

**5.Modeling**:**Star Schema**: A fact table in the middle connected to a set of dimension tables **Snowflake schema:** dimensional hierarchy **Fact constellations:** Multiple fact tables

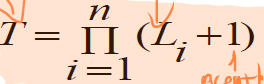
**6.OLAP Operations:** **Roll up/Drill-up:** summarize, dimension reduction; **Drill down/roll down:** new dimensions **Slice:** dimension’s subset **Dice:** data’s subset **Pivot/rotate: Drill Across:** involving across more than one fact table **Drill through:**

Through the bottom level of cube to its back-end relational tabl

**7.Measures:** **Distributive:** if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning E.g., count(), sum(), min(), max() **Algebraic:** if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate functionavg(x) = sum(x) / count(x), average, std, maxN, minN, CenterOfMass **Holistic:** if there is no constant bound on the storage size needed to describe a subaggregate .E.g., median(), mode(), rank(),Q1, Q3

**8.Materialization: Full/No/Partial Materialization**

**9.How many cuboids in an n-dimensional cube with L levels:**

3 dimensions: 3 levels + 4 levels + 5 levels

T = (3+1) (4+1) (5+1)

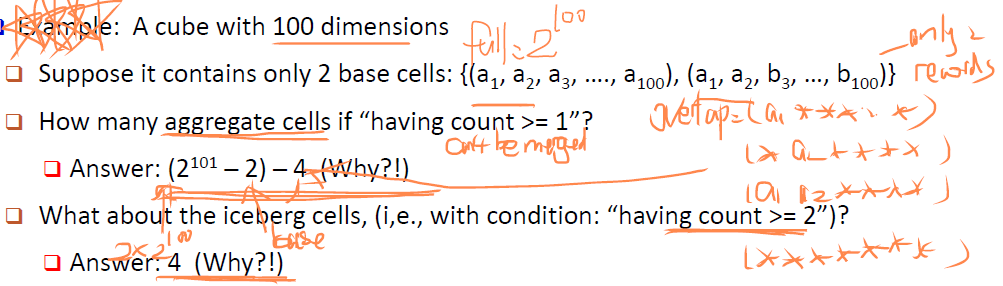
**10.Bitmap Index:** speed up + reduce storage

**11.Architectures: Relational(ROLAP):**use relational or extended-relational DBMS **Multidimensional(MOLAP):** sparse array-based multidimensional storage engine **Hybrid(HOLAP):** low level: relational, high-level: array

**\* P:** 和顺序有关 P(n,m) = n! / (n-m)!

**\* C:** 和顺序无关 C(n,m)=n! / [(n-m)! \* m!]

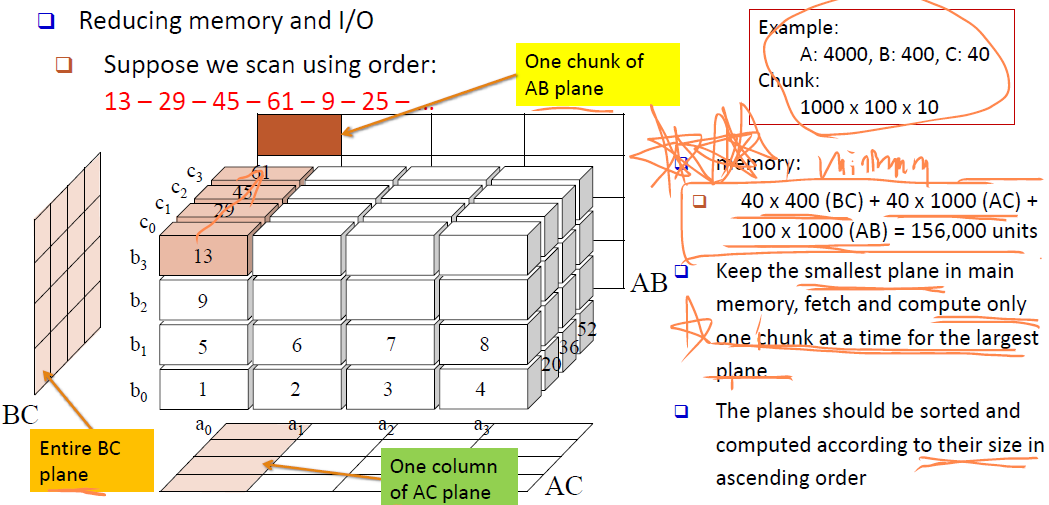
**Data Cube Technology (Chapter 5)**

1. **Base Cell / Aggregate Cell; Full Cube/Iceberg Cube**

**2.Close Cube:** a cell c is closed if there exists no cell d, such that d is a descendant of c, and d has the same measure value as c

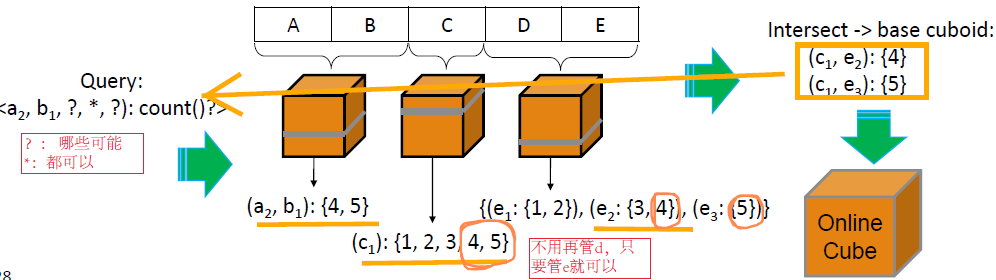
**3. Computation Methods:**

**General Heuristics (**share-sorts/share-partitions **Aggregate** may be computed from previously computed aggregates: **Smallest-child/Cache-results/Amortize-scans)**

**Multi-Way Array Aggregation** (Bottom-up): A>B>C for ABC

**BUC**(Top-down): **A Shell-Fragment Approach** (High-dimensional OLAP) **Online Query Computation with Shell-Fragments**: Frag-Shells -> Offline & Online

**总结：Multiway:** dense full cube. no iceberg, no high-D; **BUC:** ice, no high-D; **shell:** high-D, no dense full cube, no large cube



**1)** Suppose a data relation has 100 attributes and 106 tuples. Each attribute has 50 distinct values. Suppose each cell takes 16 bytes of space, and the shell-fragments are all 4 dimensions. (i) [5] What is the size (in bytes) of one pre-computed shell-fragment of size 4? Answer: 16 \* (1 + 4 \*50 + 6\*50^2 + 4\* 50^3) Bytes. For the 0-D (apex) cuboid, there is 1 cell. For each 1-D cuboid, there are 50 \*1\*1\*1 = 50 cells. There are 4 such cuboids. For each 2-D cuboid, there are 50\*50\*1\*1 = 502 cells. There are 6 such cuboids. For each 3-D cuboid, there are 50\*50\*50\*1 = 503 cells. There are 4 such cuboids. The total size of these cells is 16\* (1 + 4\*50 + 6\*50^2 + 4\* 50^3), or equivalently, 16\* ((50 + 1)4- 50\*4). Note that if we also count the base cells, since the number of tuples is smaller than 504, we can store the data in 106 cells. We will add 16\*106 to the size. (ii) [4] If an OLAP query contains 2 instantiated variables and 6 inquired variables, what is the number of shell fragments this query must access in the best case and in the worst case, respectively? Answer: Best case: 2, worst case: 8. There are 8 dimensions we care about. In the best case, these 8 dimensions happen to be in 2 shell fragments since there are 4 dimensions each. In the worst case, all these 8 dimensions are in different shell fragments and we need to access 8 shell fragments.

**2)**Cannot support the following operations efficiently? i. Computing an iceberg cube [[Multiway cannot support, aggregation bottom-up and Apriori principle/pruning cannot. ii. Processing an OLAP query on 30 dimensions [Multiway and BUC – 30 dimensions is too many]

**3)**list one best method and another worst (a) computing a dense full cube of low dimensionality (e.g., less than 6 dimensions), B: multiway. W: shell-fragment (since most part of cube is not precomputed) (b)performing OLAP operations in a high-dimensional database (e.g., over 50 dimen-sions) B: shell-fragment，W: the other (c)computing a large iceberg cube of around 10 dimensions. B: BUC W: multiway

**4)**CubeSparse Sparse array compression: Use chuck to partition the data, and use (chunk id, offset) to store only those cells contain (nonempty) values.

**5)** \*\* Cube & Cuboid\*\* Suppose the base cuboid of a data cube contains two cells(a1; a2; a3; a4; : : : ; a10) : 1, (a1; b2; a3; b4; : : : ; b10) : 1 where ai 6= bi for any I i. [3] How many nonempty cuboids are there in this data cube? Answer: 2^10. Since we have 10 dimensions with no concept hierarchy, there are 2^10

cuboids and all of them should not be empty. ii. [3] How many (nonempty) aggregate closed cells are there in this data cube? Answer: 1. There are 3 closed cells, including the two base cells and (a1; \_; a3; \_; a5; \_; a7; \_; a9; \_). But only the latter one is a aggregated closed cell. iii. [3] How many (nonempty) aggregate cells are there in this data cube? Answer: 2014. For each base cell, there are 2^10􀀀1 aggregated cells. However, there are 25 cells that are counted twice since there are 5 common dimensions. Therefore, the total number of nonempty aggregate cells is 2\*(2^10-1)-2^5 = 2014. iv. [3] If we set minimum support = 2, how many (nonempty) aggregate cells are there in the corresponding iceberg cube? Answer: 2^5. These two base cells have common value in 5 dimensions; therefore, there are 2^5 nonempty cells with support = 2 and all of them are aggregate cells.

**6)** A cell c is closed if any descendent cell d have a smaller measure than c. measure(closed cell) <= #(base cell)

(a1,a2,c3,...,ck),(a1,b2,c3,...,ck),(b1,a2,c3,...,ck), (b1,b2,c3,...,ck)

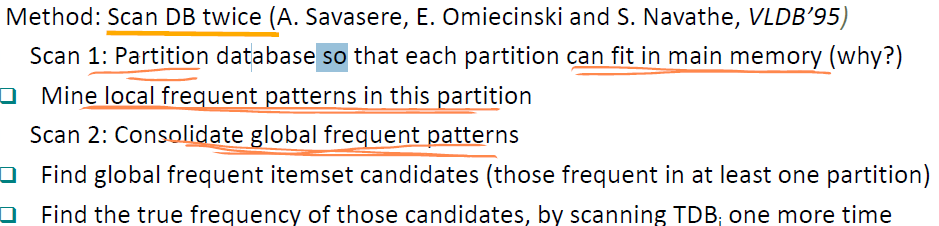
\* Cuboids = 2^k, given k dimensions

\*aggregate Closed cells: (a1,∗,c3,...,ck) : 2, (b1,∗,c3,...,ck) : 2, (∗,a2,c3,...,ck) : 2, (∗,b2,c3,...,ck) : 2, (∗,∗,c3,...,ck) : 4

\*aggregate cells: 4×(2^k-1)−4×2^(k−2)−3×2^(k−2)

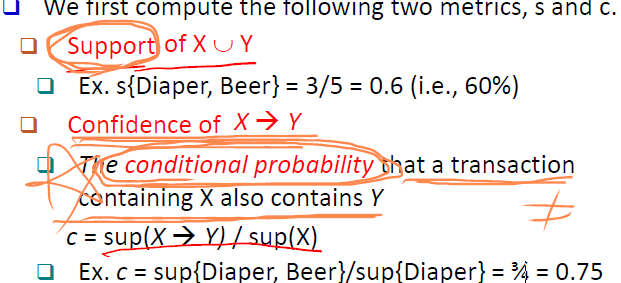
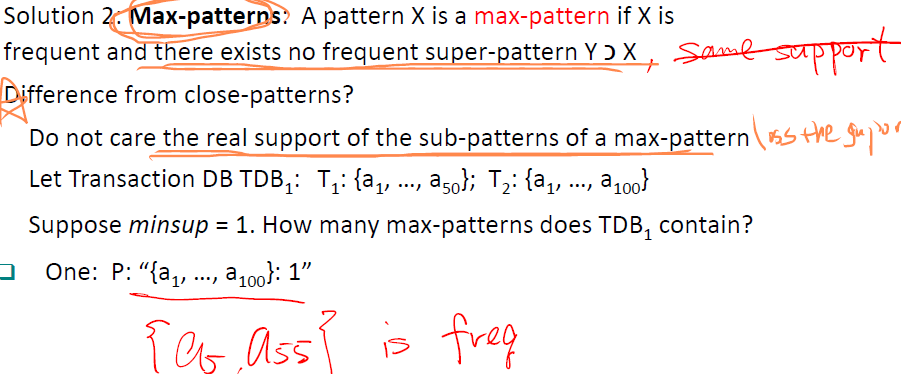
Hint: “(a1,∗,c3,...,ck) : 2” merges 2 × 2k−2 cells into 1 × 2k−2 cells and thus net loss is 2k−2 cells. There are 4 such cells. Thus their total loss is 4 × 2k−2 cells. “(∗,∗,c3,...,ck) : 4” merges 4 × 2k−2 cells into one 2k−2 cells and thus net loss is 3 × 2k−2.

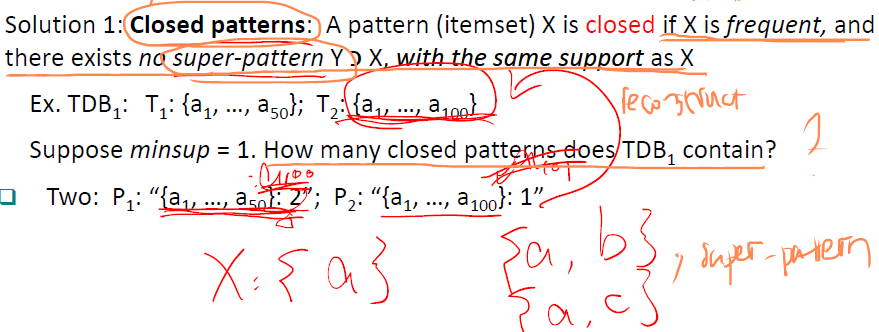
\*If we set min\_support = 2, how many aggregate cells are there in the corresponding iceberg cube? 5 × 2^(k−2) since there are 5 aggregate closed cells, each will cover 2k−2 aggregate cells.

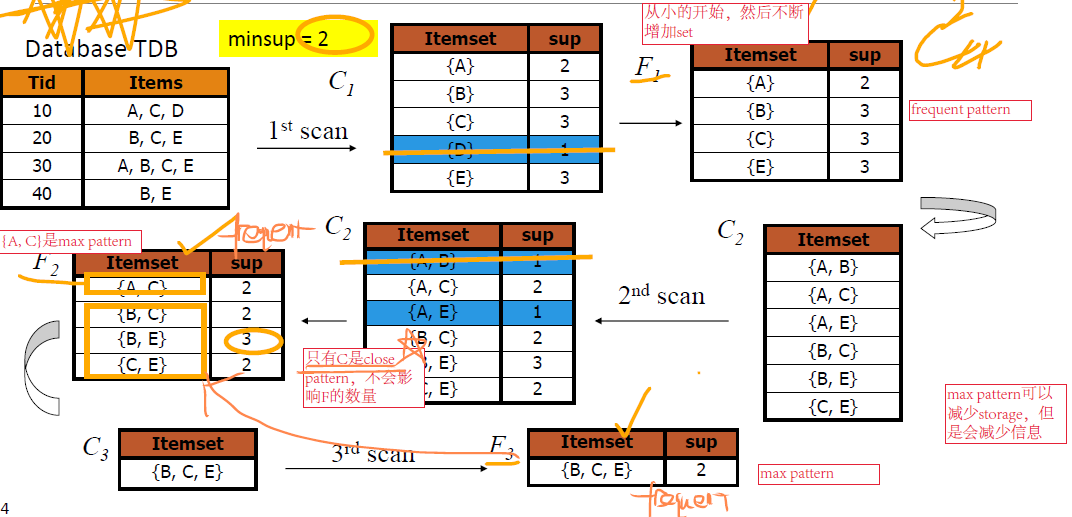
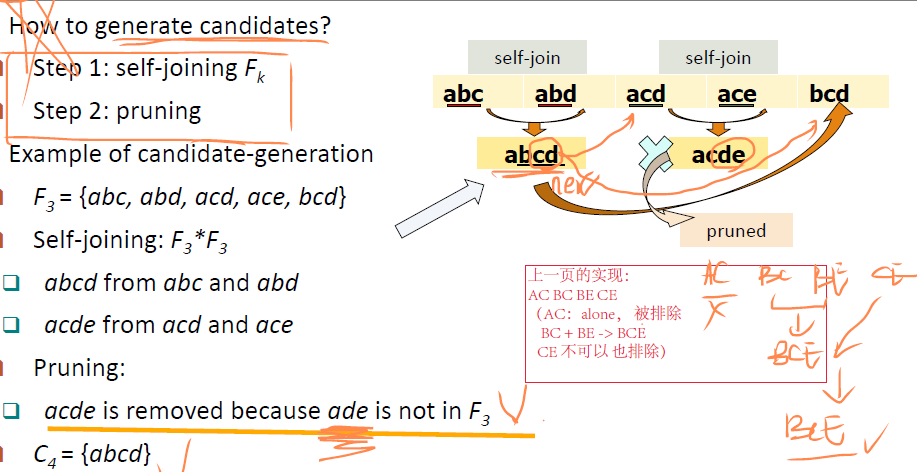
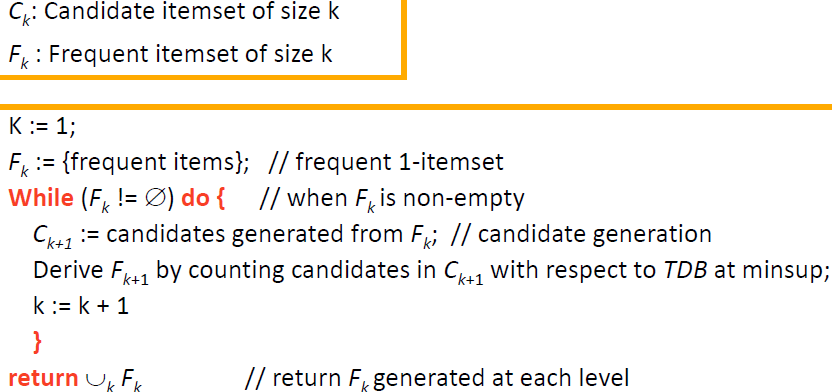
**7)** Explain why good intra-cuboid expansion may enhance the quality of drill-down in sampling cube, but a bad intra-cuboid expansion may reduce its quality. [Enhance quality: if we expand on a dimension that is not correlated with the target measure, we have more data available (so we can have tighter confidence bounds) Reduce quality: if we expand on a dimension correlated with the target measure, the new data will skew the target measure. We effectively end up answering a different question than the one posed]

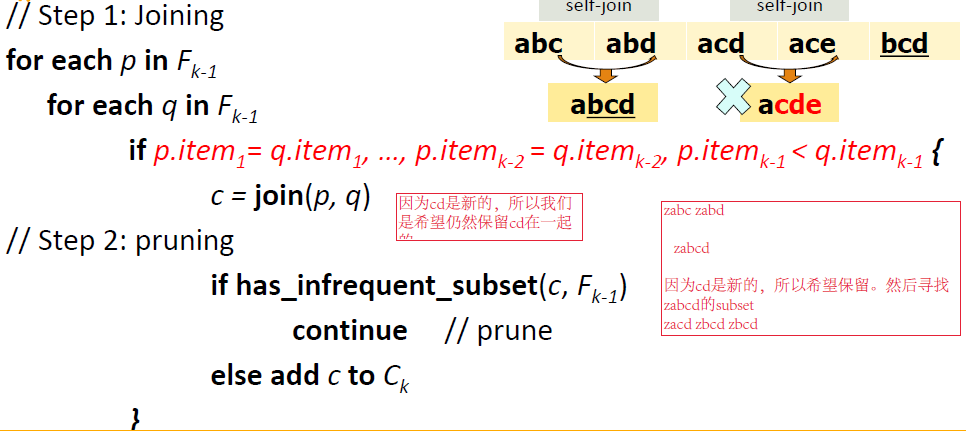
**8)** What are the major challenges to support OLAP on sampling data? [Outline a method that may support such operations effectively. Major challenges: Some drilling down cells may contain few or no data. Method: Intra or inter-cuboid expansion to combine with other cells whose attributes has low correlations with the dimensions interested.]

**Mining Frequent Patters, Association and Correlations(char6)**

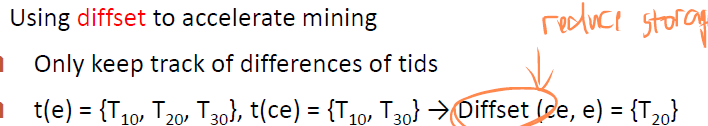
**1.Transactional Database(TDB): itemset, k-itemset, absolute/relative support. Frequent:** if the support of X is no less than a minsup threshold. **Association Rule:**

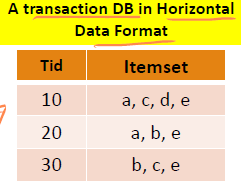
closed pattern： lossless compression/ Max-pattern: lossy

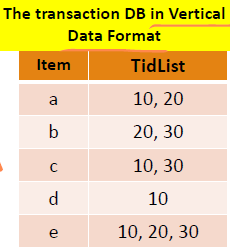
**2.Downward Closure Property Apriori:** any subset of a frequent itemset must be frequent

**Partitioning:** any itemset that is potentially frequent in TDB must be frequent in at least one of the partitions of TDB

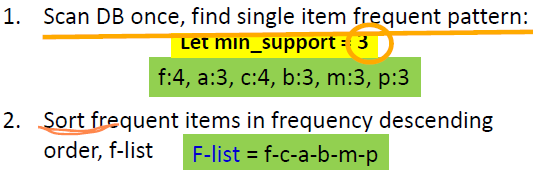
**Direct Hashing and Pruning:** different itemsets may have the same hash value, can speed up while the size of buckets重要

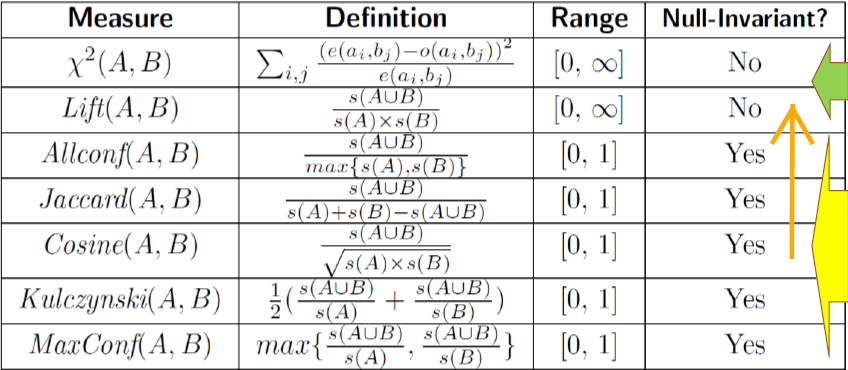
**3.Vertical Data Format:**



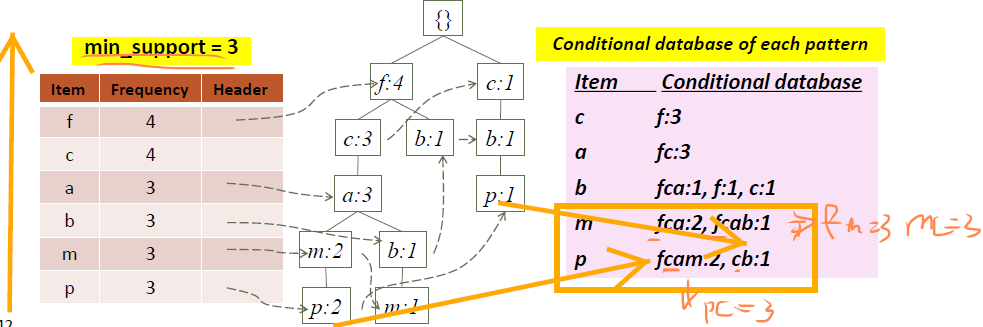
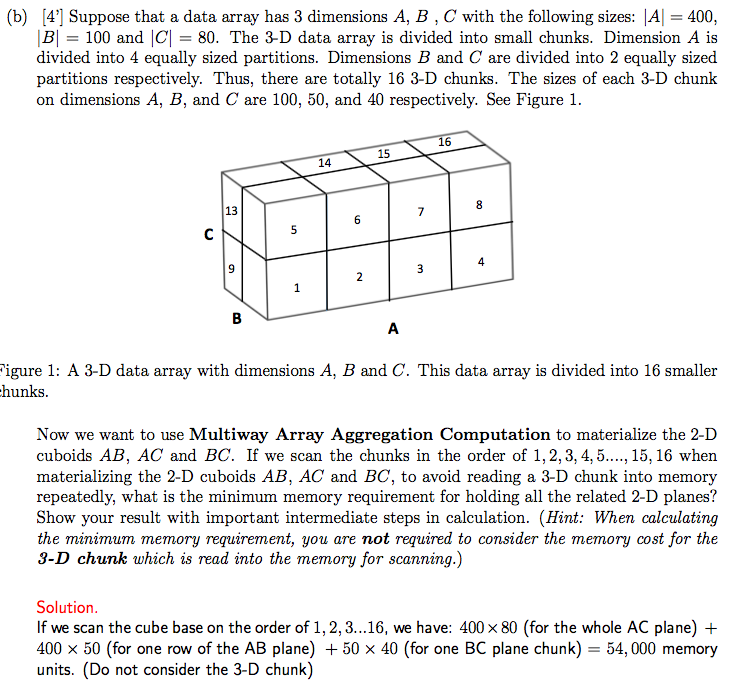


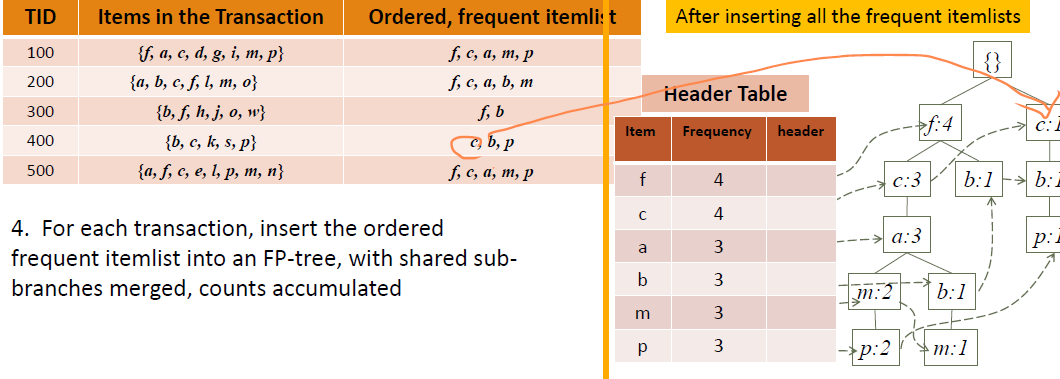
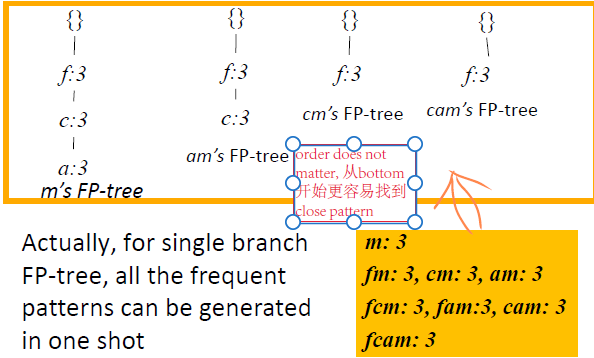
**4.FP-Tree**

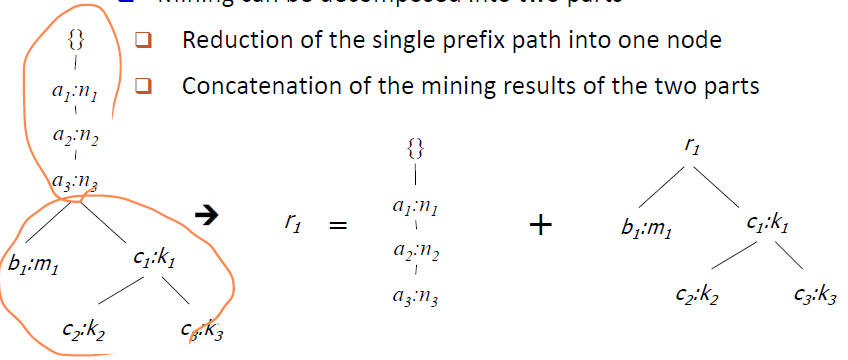
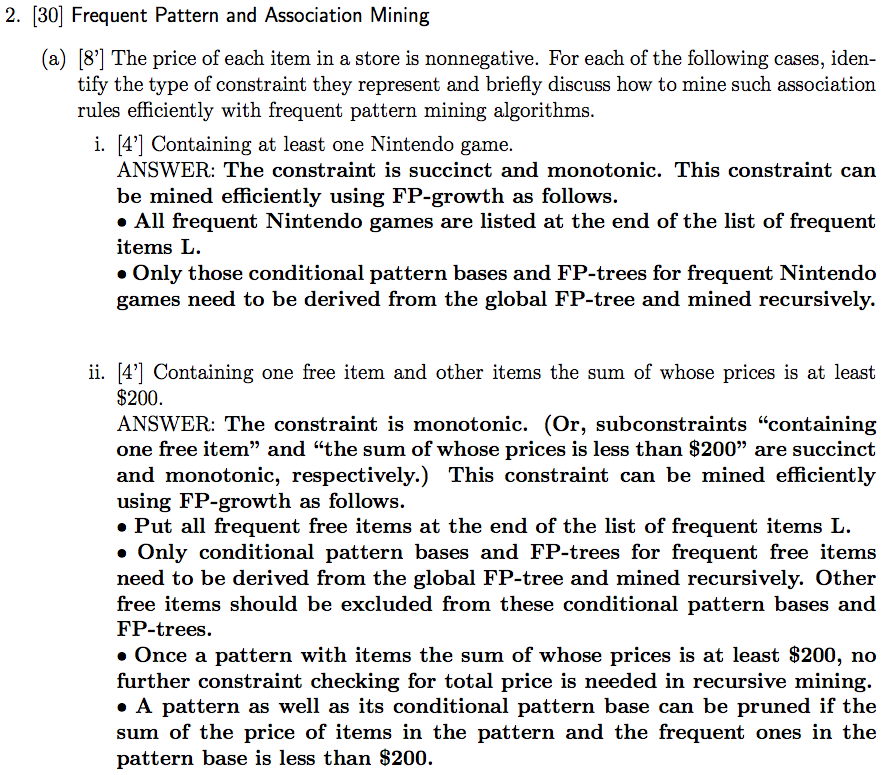


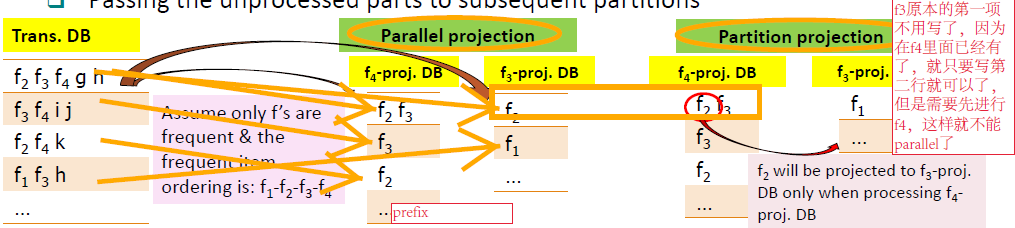






**Mine Each Conditional Database recursively:**

**Single Prefix Path in FP-tree;**

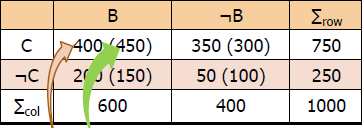
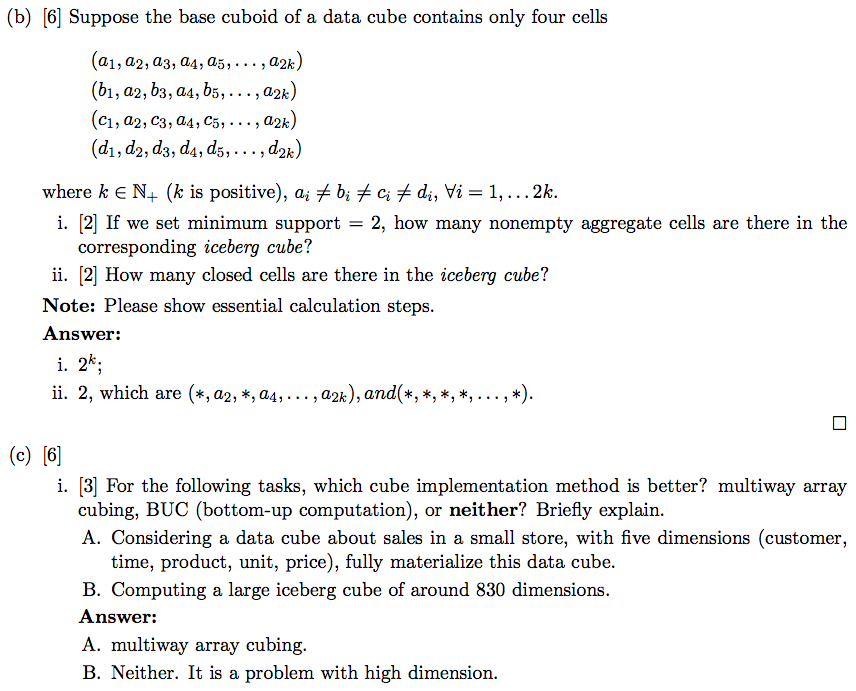
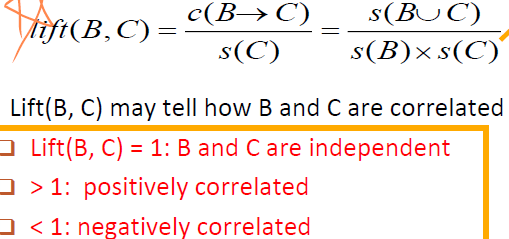
**Scaling FP-growth by Item-Based Data Projection**

**Closet” mining closed itemsets by pattern-growth**

**5.Pattern Evaluation: Strong Association Rule**

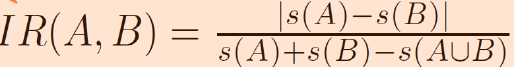
**6.Interstingess Measure:**

**Lift(**dependent/correlated)

**X2**

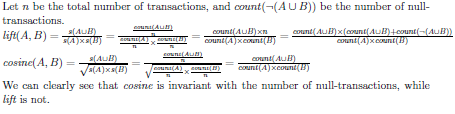
**7.Null-Invariance:** value does not change with the # of null-transactions。改变null-transactions时lift和x2会变，所以不是null-invariant. Not good to evaluate data.

**8. Imbalance Ratio with Kulczynski Measure:**



**\*\*Null invariant\*\*** Suppose a transaction database contains N transactions, ct transactions contain both coffee and tea, ct transactions contain coffee but not tea, ct transactions contain tea but not coffee, and ct transactions contain neither tea nor coffee. At what condition that Lift may not be a good measure to indicate correlations between coffee and tea? and why? 【】Condition: If the number of ct transactions is substantially larger than the remaining # of transactions, LIFT is not a good measure. / Reasoning: Lift is not null-invariant. That is, although LIFT is affected by the number of ct transactions, the correlation should not be (because transactions that do not contain c or t tell us nothing about the correlation - so they shouldn’t affect the measure).

\*\* **Null invariant** \*\*Explain why null-invariance property is important in the study which authors are“correlated?【】 Since most authors are not coauthoring papers, null value is very high. Null-invariance is critical otherwise the ”correlation value could be greatly influenced by null values.

(1) mine FPs from the new DB; (2) get counts of those patterns frequent in new DB but not in old DB by scanning the old DB once, (3) get counts of those that are frequent in old DB by not in new DB. The merge counts will be used to just which patterns are globally frequent