6	"PipeSort combines the optimizations share-sorts and smallest parent to get the minimum total cost." Elaborate. PipeSort algo. ✓ Reduces the min total cost when operations like grouping of attributes is done on vast & huge datasets. It uses methods like: 1) Optimizing share sorts: It waits for the Input streams & then do the sorting & Grouping together. It dont go through dataset again & again. for ex → if we have to group all A with all B and func" happening on those 2 all. so we will join them first & then cost 2) Smallest Parent: If we have to group attributes like A,B, and use already have grouped A,B so we will use it (from cache memory) instead of grouping from Scratch again.
7	"When we consider computing the cube on data stored in arrays, one can once again use the ROLAP trick of computing one aggregate from another. However, none of the other techniques that have been developed for ROLAP cube computations apply to the data stored in arrays." Explain the reasons. ROLAP: Relational Online Analytical Processing. It works on Relational database systems. For arrays, we typically use MOLAP, but just aggregation on arrays can be done by ROLAP also. But other techniques like Grouping diff. attributes is not basable on array, storing parent info & using it not poss. as we always traverse the array again & again. ROLAP techniques are faster than MOLAP as continuous traversing of arrays is not these. And it has functionalities applicable on Relational Databases as seen above, and hence using them on arrays is not a good choice.
8	Discuss how the ordering of dimensions with respect to cardinality of attribute and skew of the attribute will impact the performance of BUC approach. Bottom to Up approach: Make Data Marts ↓ Create Data Warehouse Ordering of Dimensions will affect a lot as it will decide in which order data pre-processing will happen. If there's any col with a lot of null values and we process it in start by dropping all those rows, it will lead us to loose on a lot of useful information. 1) Cardinality of attribute: How many distinct values are there in the attribute. If there are less distinct values, pruning will be easier. Hence, less Cardinality should be done first as less loss of info will happen. 2) Skew of attribute: More skewed attributes should be done first because skewed attributes have outliers and its good to remove them in start itself, otherwise they can skew other results as well.

3	Explain the purpose of normalization of data. Compare min-max and z-score normalization methods. Take example of ordinal data. There's a ranking order but it can go to 1,2,3 or 1,2,3,4,5 or 1,2,3, ... 1000, etc. Hence, to standardize the results and make them uniform across all the attributes of schema, normalization is used. With normalization, all possible values are in range (typically [0,1]) which helps in comparison, aggregation, generalization, etc. It also helps in standardization while analysis. Two types of Normalization: 1) Min-max: $x_i = \frac{x - Min(A)}{Max(A) - Min(A)}$ (x is the value of attribute. Min(A) = min val. in att. A, Max(A) = max) It is used when the data is not skewed a lot. 2) Z-score: $x_i = \frac{x - \mu}{\sigma}$ (μ → mean of values in att, σ → S.d) It is used when values are tend to follow Normal distribution. Outliers?
4	Major features of data warehouse are: subject-oriented, integrated, time-variant and nonvolatile. What do you understand with key words: (i) "subject-oriented" and (ii) "nonvolatile" Data warehouse is the entity where all the data is stored after pre-processing stage. → Subject-Oriented: It is divided into data marts which are actually the subsets of Data Warehouse. They keep information about diff. sectors eg. customers, sales, returns, management, etc. This helps in better querying of results. → Non-volatile: Data is stored and added in warehouse continuously but typically never deleted. It's main purpose only is for analysts to analyze historical trends, check present & predict future applications & trends for betterment of the Company.
5	Can we extend "Learning from Examples" paradigm to learn concepts like "attribute oriented induction method" from relational databases? Discuss. "Learning from examples" means using datasets & knowledge available to train a ML model for ex. But "attribute oriented Method" works on relational databases not by training on typically 1000s of tuples but by grouping tuples together & then learning. For ex → Age categorised as <18: Child, <35: Young, >35: old So, in AOI method Model is trained on this Grouped information and hence, it's not feasible to extend Learning from examples to this. Grouping, aggregating, comparing, etc. will take a lot of time & cost. Negative EX?? (1.5)

Note: Answer all questions. Make appropriate assumptions. Give a brief answer. There are 8 questions. Each question is for 2 marks.

1	Compare the two methods for filling the missing values: one is using any of the central tendency and the other is using most probable value. Filling missing values is one of the most imp. step in data cleaning. Otherwise, data type errors, arithmetic errors, etc comes. 1. Fill by Central Tendency (Mean/Median/Median): We impute the null value with the Mean/Median/Mode of that attribute. Mean (symmetrical data), median (asymmetrical so that outliers aren't that big problem). 0.25 2. Fill by most probable value (Medoid): It's not median but in any dimension, a point from the dataset which is equi-distant from all other points. Filling by most probable value will be better as 0.25 outliers won't affect the result a lot than Mean won't be good for skewed datasets & median only applicable in 1D attribute.
2	Explain the issue of discrepancy in the data. Discuss the role of metadata in discrepancy detection. Give two examples. Discrepancy in data occurs when there is some mis-match in the data in datamarts, or there is some null values, improper grouping of columns, etc. 0.75 Metadata keeps the store of schema of warehouse, data types of attributes, all the actions being done, etc. Hence, in case of discrepancies it will be useful for verification about where our data preprocessing went wrong or some mistake occured. eg. • Attribute Age having String data type instead of column • Attribute is blue having null values instead of binary 0/1. • An attribute missing from the schema.

1.75