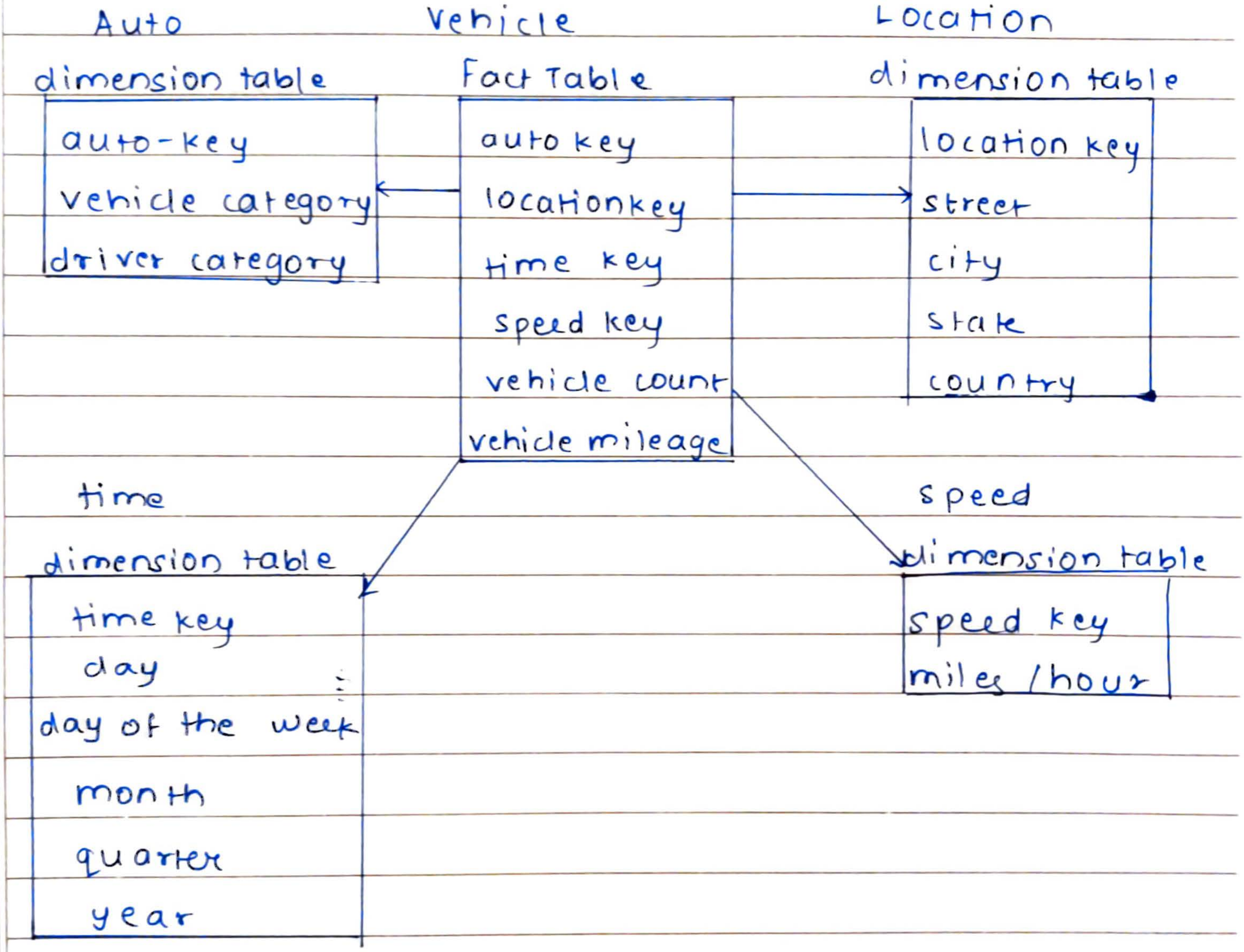
Shubham Bhate 8318 TE comps

DWM Assignment 1

Q1. a]



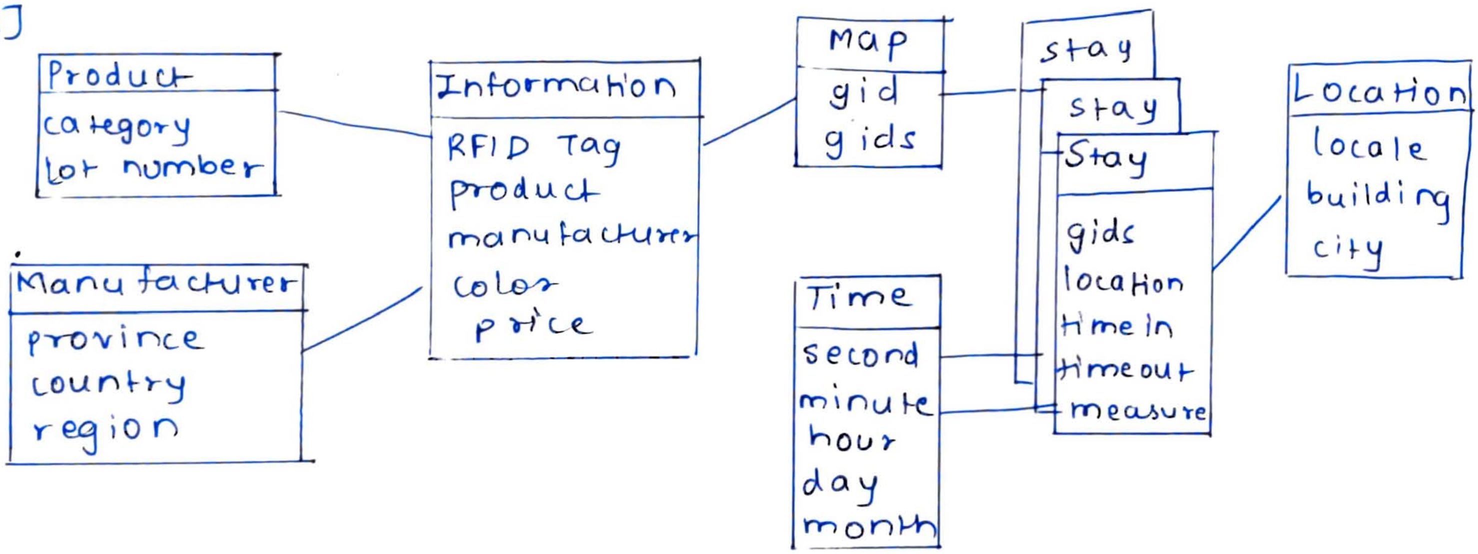
b] To handle noise, we first need to clean data. Mining values may be filled or dropped entirely. Then we can use data smoothing techniques to remove noisy data points. we can also set up rules to remove inconsistent data based on domain knowledge.

c] It is possible to get a data warehouse that is sparse. Analyzing sparse data is not reliable as single outlier may completely shift results. Hence there are few values to deal with. we have to evaluate confidence interval in such cases wherein it defines reliablity of data. confidence interval is directly proportional to accuracy of data. Hence for our vehicle database, it is computed for reducing sparsity.

d] Using this warehouse, we can look up the information for the vehicles of same vehicle and driver category. Then, using OLAP we look up the speedof a location at a specific time and will use that as a weight for the street on the city graph. Using this algorithm we dont care about direction of the street. we can also integrate the information and create a directed graph.

Q.2 RFID warehouse

a]



b) Each reader provides tuples at fixed time intervals. we can group this into a single one like (RFID, location, time in time out) eg. If a super market has reader on one line that scans every time and items stay on shelf for 1 day we get 1440 to 1 reduction in size without loss of info.

c) We can use the assumption that many RFID objects stay or move together, especially at early stages of distribution or use historically most likely paths of items to infer missing or error.

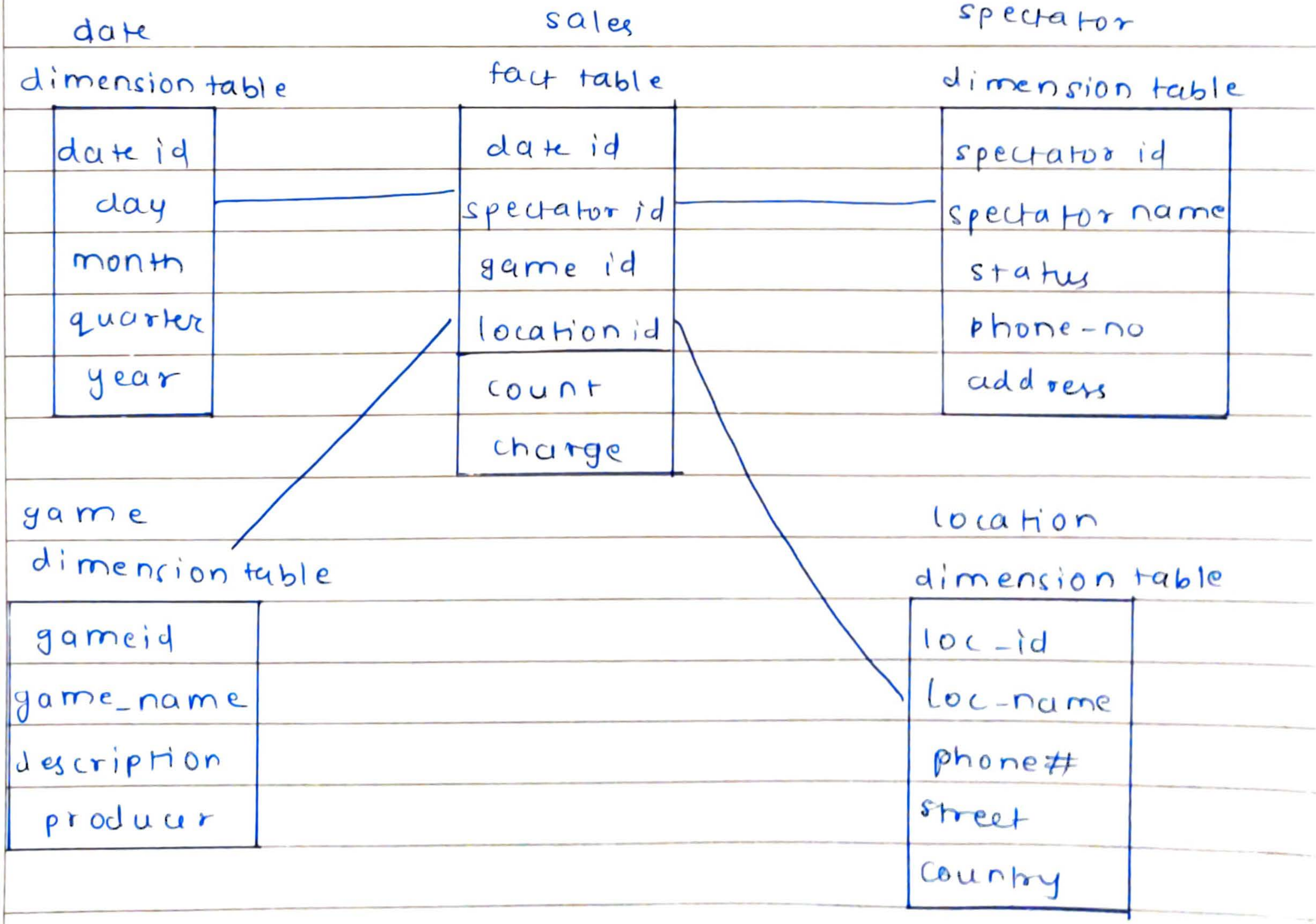
d) Compute an aggregate measure on tags that travel through set of locations and that match selection criteria on path independent dimensions.

e) For this case after RFID of milk is obtained OLAP can be directly used to get shipping and storage time efficiently

Q 3a] Use partial materialization or selected computation of cuboids. By computing only proper subset of whole sets possible cuboids, total storage space would be minimized with fast response time

b) Since this is only for 1/2 dimensions it can be done on the fly. since this feature is needed infrequently, time required for computing aggregates on those should be accepted.

Q.4 a] Game - sales data warehouse



date dimension table

date id
day
month
quarter
year

sales fact table

date id
spectator id
game id
location id
count
charge

spectator dimension table

spectator id
spectator name
status
phone-no
address

game dimension table

gameid
game_name
description
producer

location dimension table

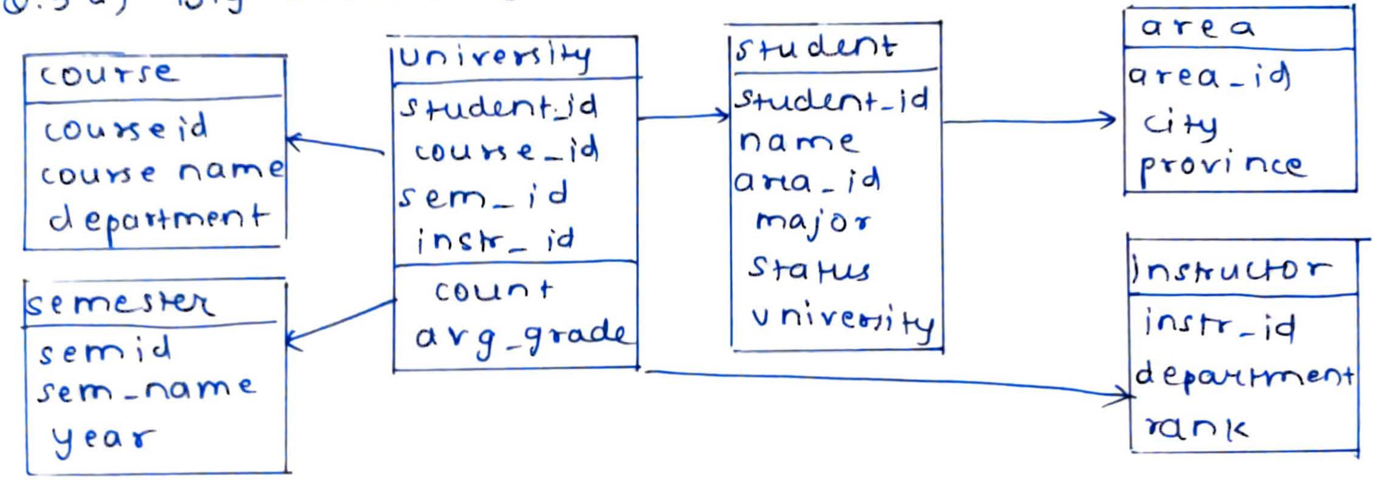
loc_id
loc-name
phone#
street
country

b] · Rollup on date from date id to year

- Roll up on game from game id to all
- Rollup on location from loc id to loc-name
- Rollup on spectator from spec id to status
- Dice with status = "student", loc_name = "GMplace" and year = "2010"

c] It is advantageous for low cardinality domains. For eg. if location is bitmap indexed comparison, join and aggregation operations over location are then reduced to bit arithmetic which substantially reduces processing time, leading to significant decrease in space and I/O time.

Q.5 a) Big University warehouse



b) · Rollup on cousre from course_id to dept

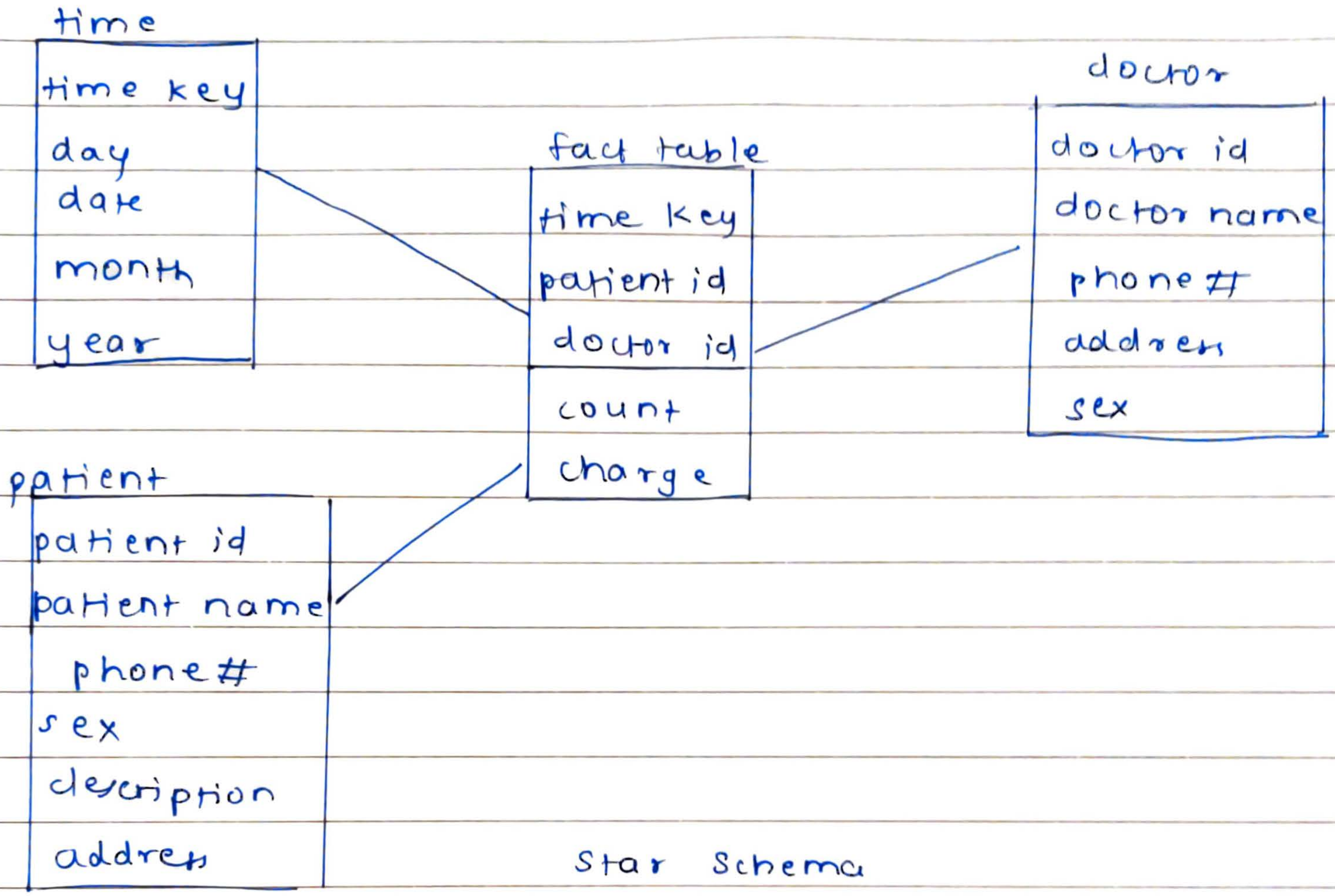
- Rollup on semester from sem id to all
- slice for course = "CS"

c) It will contain $5^4 = 625$ cuboids

Q.6 a) Three classes of schemas used to model data are

① star
② snowflake
③ fact constellation

b)



Star Schema

c)
- Roll up on time from day to year
- Roll up on patient from individual to all
- Slice for time = 2010

d) select doctor, SUM (charge) from fee
where year = 2010
group by doctor