DATA WAREHOUSE ASSIGNMENT

QUESTION 1

1   BUSINESS PROCESS

Design a Data Warehouse for IPL Cricket Tournament

- Where is the tournament taking place?
- How many countries are participating in the tournament?
- When is the tournament starting and ending?
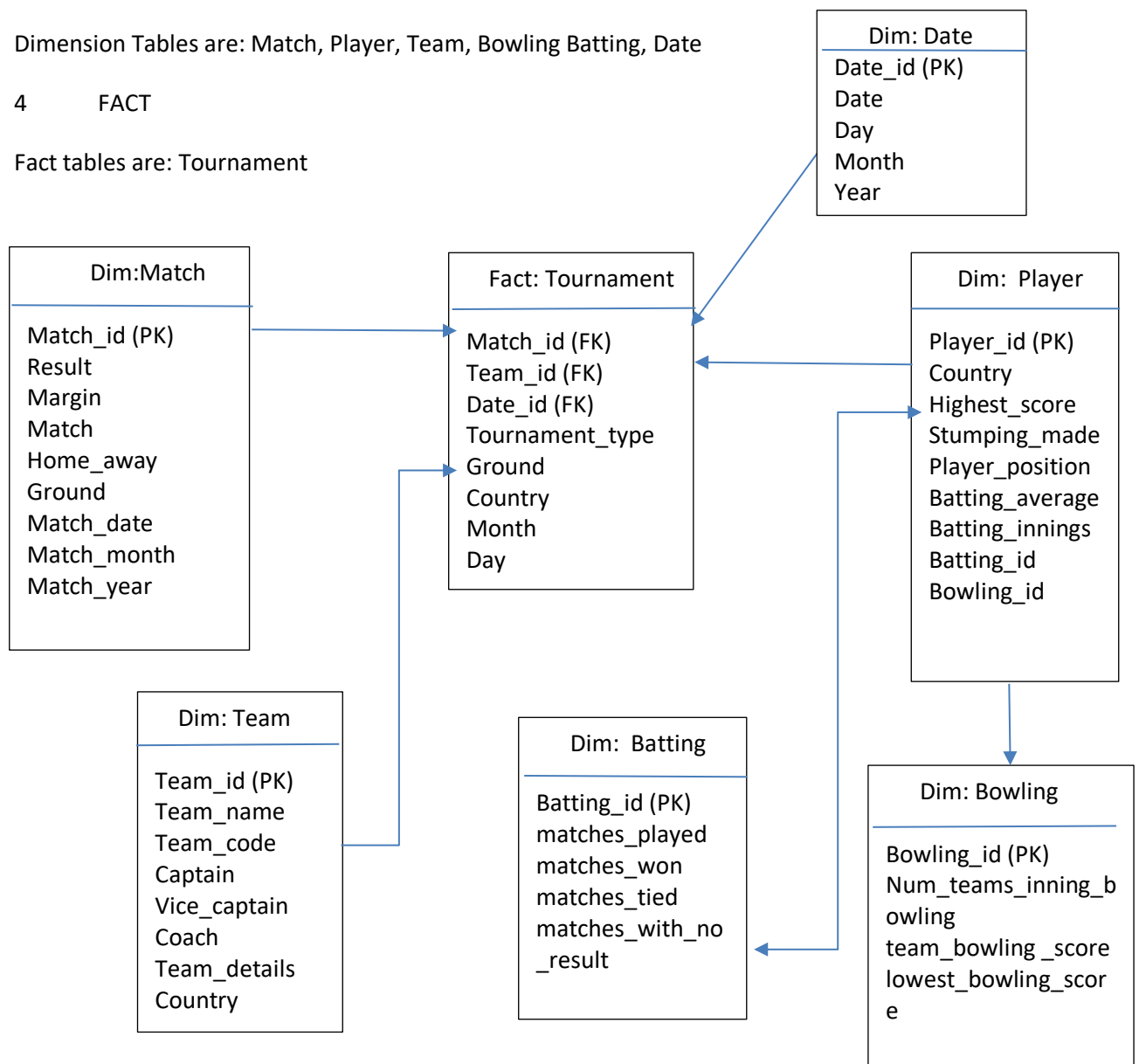- How many games are played per day?

2   THE GRAIN

Each team involved in the tournament. A single entry in the fact table will describe the entire tournament activities.

3   DIMENSION

Dimension Tables are: Match, Player, Team, Bowling Batting, Date

4   FACT

Fact tables are: Tournament

Dim: Date
Date_id (PK)
Date
Day
Month
Year

Dim:Match
Match_id (PK)
Result
Margin
Match
Home_away
Ground
Match_date
Match_month
Match_year

Fact: Tournament
Match_id (FK)
Team_id (FK)
Date_id (FK)
Tournament_type
Ground
Country
Month
Day

Dim: Player
Player_id (PK)
Country
Highest_score
Stumping_made
Player_position
Batting_average
Batting_innings
Batting_id
Bowling_id

Dim: Team
Team_id (PK)
Team_name
Team_code
Captain
Vice_captain
Coach
Team_details
Country

Dim: Batting
Batting_id (PK)
matches_played
matches_won
matches_tied
matches_with_no _result

Dim: Bowling
Bowling_id (PK)
Num_teams_inning_bowling
team_bowling _score
lowest_bowling_score

QUESTION 2

1 BUSINESS PROCESS

Design a Data Warehouse for Food Delivery App e.g Swiggi

- Where is the tournament taking place?
- How many countries are participating in the tournament?
- When is the tournament starting and ending?
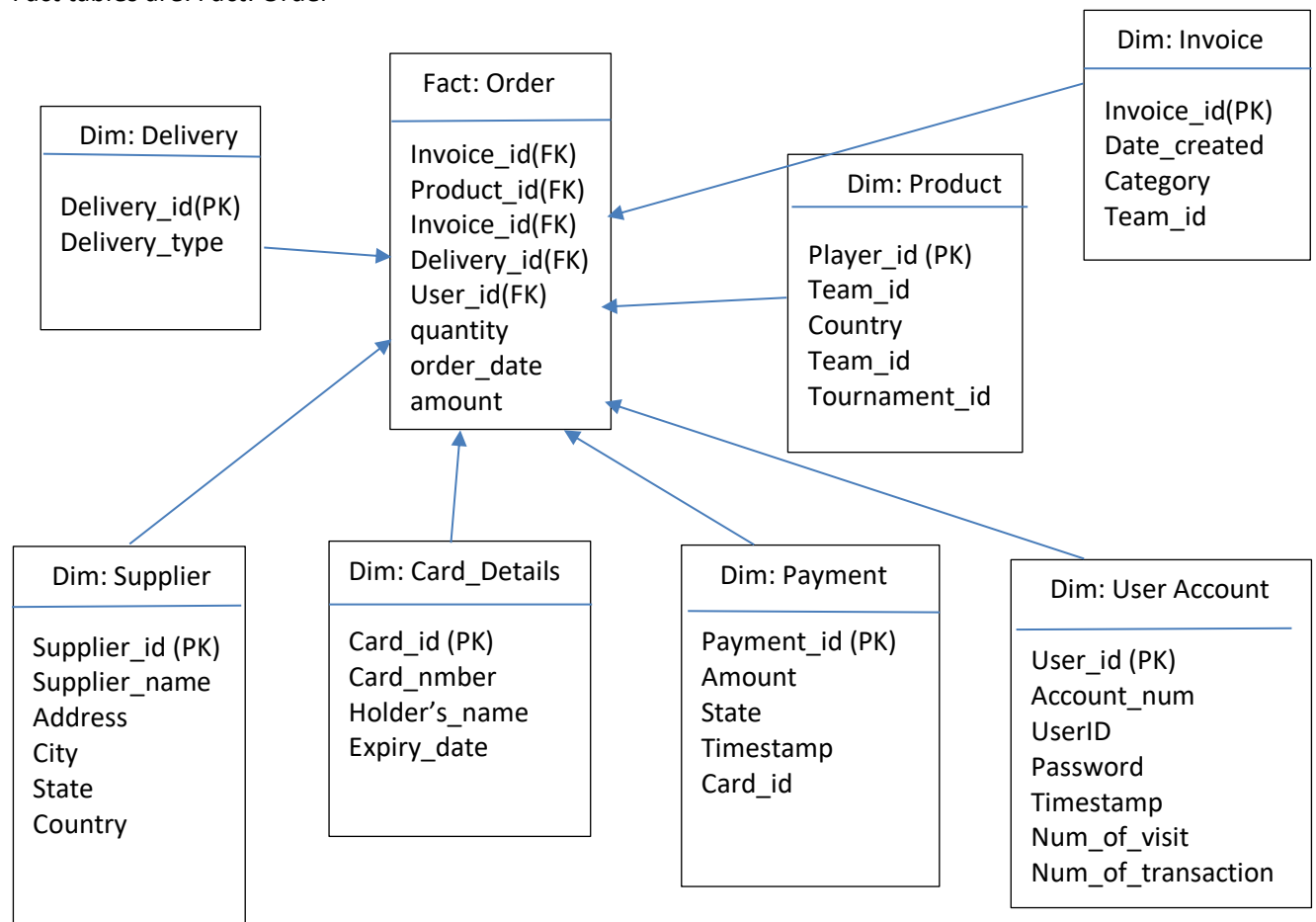- How many games are played per day?

2 THE GRAIN

The granularity of FactDeliveres is each line on the invoice. The Product dimension includes supplier information.

3 DIMENSION

Dimension Tables are: Invoice, Delivery, Product, Supplier, Payment, User_Account, Card_details

4 FACT

Fact tables are: Fact: Order

Dim: Invoice

Invoice_id(PK)
Date_created
Category
Team_id

Fact: Order

Invoice_id(FK)
Product_id(FK)
Invoice_id(FK)
Delivery_id(FK)
User_id(FK)
quantity
order_date
amount

Dim: Delivery

Delivery_id(PK)
Delivery_type

Dim: Product

Player_id (PK)
Team_id
Country
Team_id
Tournament_id

Dim: Supplier

Supplier_id (PK)
Supplier_name
Address
City
State
Country

Dim: Card_Details

Card_id (PK)
Card_nmber
Holder's_name
Expiry_date

Dim: Payment

Payment_id (PK)
Amount
State
Timestamp
Card_id

Dim: User Account

User_id (PK)
Account_num
UserID
Password
Timestamp
Num_of_visit
Num_of_transaction

QUESTION 3

1 BUSINESS PROCESS

Design a Data Warehouse for Cab Ride Services e.g Uber

- Track rides done by a driver and their performance?
- How many rides are taken to known destinations each day?
- How many trips were cancelled per day?
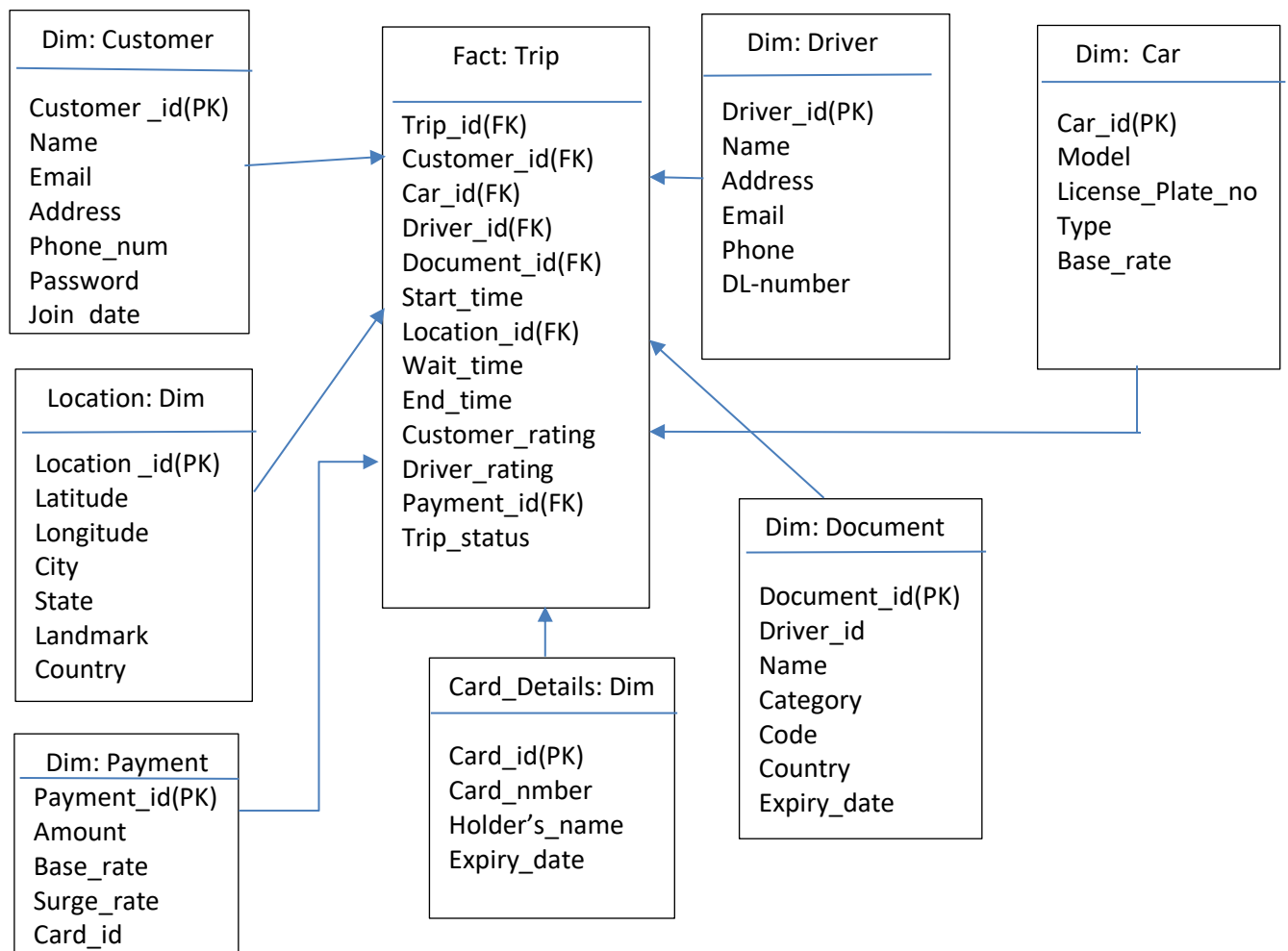- How many rides and the average price during peak hours per day?

2 THE GRAIN

An individual trip on each transaction level. One single entry in the fact table that will describe the entire process of the business trip.

3 DIMENSION

Dimension Tables are: Customer, Driver, Car, Location, Payment, Card_details and Document

4 FACT

Fact tables are: Trip

Dim: Customer

Customer _id(PK)
Name
Email
Address
Phone_num
Password
Join date

Fact: Trip

Trip_id(FK)
Customer_id(FK)
Car_id(FK)
Driver_id(FK)
Document_id(FK)
Start_time
Location_id(FK)
Wait_time
End_time
Customer_rating
Driver_rating
Payment_id(FK)
Trip_status

Dim: Driver

Driver_id(PK)
Name
Address
Email
Phone
DL-number

Dim: Car

Car_id(PK)
Model
License_Plate_no
Type
Base_rate

Location: Dim

Location _id(PK)
Latitude
Longitude
City
State
Landmark
Country

Dim: Document

Document_id(PK)
Driver_id
Name
Category
Code
Country
Expiry_date

Card_Details: Dim

Card_id(PK)
Card_nmber
Holder's_name
Expiry_date

Dim: Payment

Payment_id(PK)
Amount
Base_rate
Surge_rate
Card_id

QUESTION 4

1   BUSINESS PROCESS

Design a Data Warehouse for Restaurant Table Booking

- Where is the Restaurant located?
- How many people are the reservations for?
- When is the reservation booking?
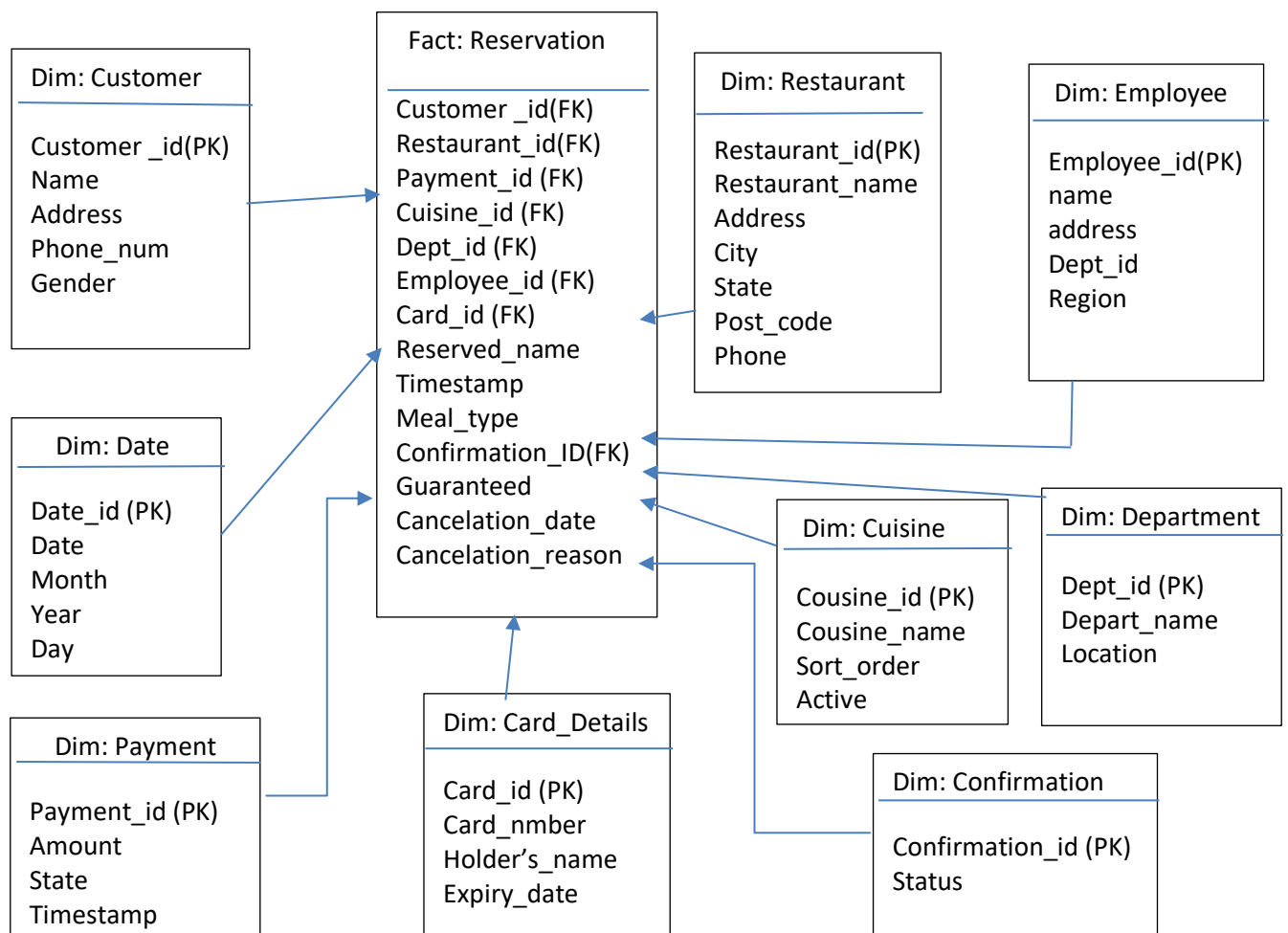- What time is the reservation?

2   THE GRAIN

The granularity of FactReservation is one row per reservation per customer.

3   DIMENSION

Dimension Tables are: Customer, Restaurant, Employee, Date, Payment, Card_details, Cuisine, Department and Confirmation

4   FACT

Fact tables are: Reservation

Fact: Reservation
Customer _id(FK)
Restaurant_id(FK)
Payment_id (FK)
Cuisine_id (FK)
Dept_id (FK)
Employee_id (FK)
Card_id (FK)
Reserved_name
Timestamp
Meal_type
Confirmation_ID(FK)
Guaranteed
Cancelation_date
Cancelation_reason

Dim: Customer
Customer _id(PK)
Name
Address
Phone_num
Gender

Dim: Restaurant
Restaurant_id(PK)
Restaurant_name
Address
City
State
Post_code
Phone

Dim: Employee
Employee_id(PK)
name
address
Dept_id
Region

Dim: Date
Date_id (PK)
Date
Month
Year
Day

Dim: Cuisine
Cousine_id (PK)
Cousine_name
Sort_order
Active

Dim: Department
Dept_id (PK)
Depart_name
Location

Dim: Payment
Payment_id (PK)
Amount
State
Timestamp

Dim: Card_Details
Card_id (PK)
Card_nmber
Holder's_name
Expiry_date

Dim: Confirmation
Confirmation_id (PK)
Status

1 BUSINESS PROCESS

Design a Data Warehouse for Covid Vaccination Application

- Where is the Hospital located?
- How many recovered cases are reported?
- How many people test positive?
- How many death cases are reported?
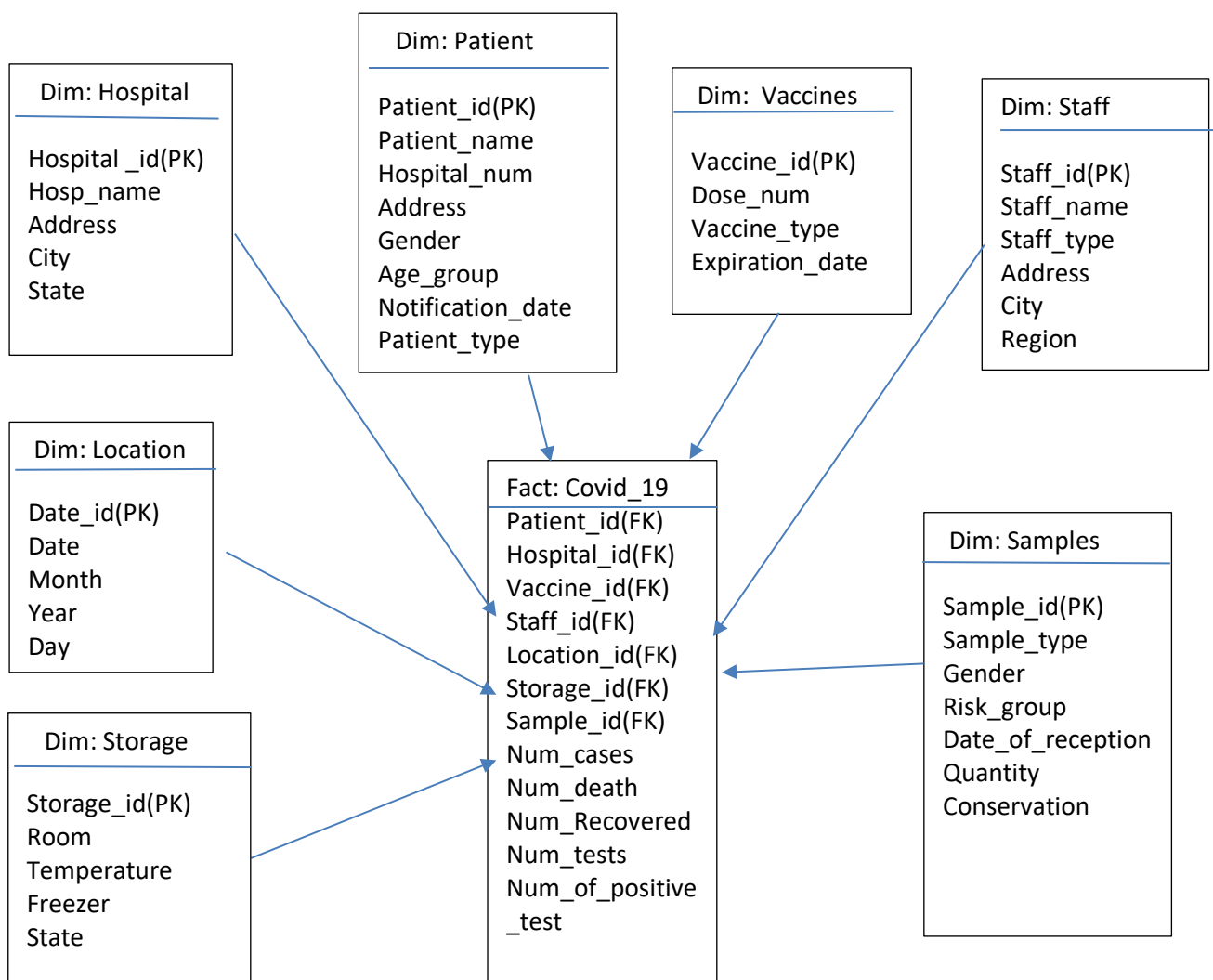- Number of reported cases?

2 THE GRAIN

The granularity of FactCovid_19 is one row per patient per vaccination.

3 DIMENSION

Dimension Tables are: Hospital, Patient, Vaccine, Staff, Location, Sample, Storage

4 FACT

Fact tables are: Covid_19

Dim: Patient

Patient_id(PK)
Patient_name
Hospital_num
Address
Gender
Age_group
Notification_date
Patient_type

Dim: Hospital

Hospital _id(PK)
Hosp_name
Address
City
State

Dim: Vaccines

Vaccine_id(PK)
Dose_num
Vaccine_type
Expiration_date

Dim: Staff

Staff_id(PK)
Staff_name
Staff_type
Address
City
Region

Dim: Location

Date_id(PK)
Date
Month
Year
Day

Fact: Covid_19

Patient_id(FK)
Hospital_id(FK)
Vaccine_id(FK)
Staff_id(FK)
Location_id(FK)
Storage_id(FK)
Sample_id(FK)
Num_cases
Num_death
Num_Recovered
Num_tests
Num_of_positive
_test

Dim: Samples

Sample_id(PK)
Sample_type
Gender
Risk_group
Date_of_reception
Quantity
Conservation

Dim: Storage

Storage_id(PK)
Room
Temperature
Freezer
State

SCENARIO 1

1   BUSINESS PROCESS

Design a Data Warehouse for Student enrolment in Stanford Courses

- Which courses are most popular?
- Which Instructors are most popular?
- Which courses are most popular among graduate students?
- Which courses are most popular among undergraduate students?
- Are there courses for which the assigned classrooms are too large or too small?

2   THE GRAIN

The granularity of Factcourse_enrolment, is one row per student per course enrolment.

3   DIMENSIONS

Dimension Tables are Courses, Department, Student, Term, Classroom, and Instructor

4   FACT

Fact tables are: Course_enrolment (course_id, depart_id, student_id, term_id, classroom_id, instructor_id, EnrollmentCount)

Question 1

Option A	
Strengths	**Weaknesses**
It allows for Course Instructor choices to be made	Data access and data processing time is increased
Instructor dimension is added to the fact table	Added row increases the size of the instructor entity
It includes all entities and the relationship among them	
Option B	
Database records are easily enhanced	Too many entity rows created
Create a normalised instructor entity	Improper aggregation of an enrollment count value
Easy data transformation process	
Ease of writing data to the entity instructor	
Option C	
Create a normalized Instructor entity i.e normalization occurs here	Create confusion for warehouse users
Trusted database stored in data warehouse	Double entry in the Fact table
	A bit more effort is required to enhance the data

Question 2

Option a will be most appropriate due to the following reasons;

1. It's easy to extract data from the source application
2. Data transformation and filtration are easy to complete
3. Less complication
4. Limited dimensions entity
5. Easy to query the data warehouse

Question 3

My response to the question would change and option C would have been chosen based on the circumstances described because option C would guarantee data integrity in the data warehouse. This is because normalized data is created for easy data transformation before being loaded into the data warehouse.

SCENARIO 2

1   BUSINESS PROCESS

Design a Data Warehouse for an Online Brokerage Company

- How often do they make transactions?
- What is the average transaction size?
- How recent is the transaction?
- How much profit is earned by each customer per trade?

2   THE GRAIN

The granularity of the Trades fact table is one row per stock trade.

3   DIMENSIONS

Dimension Tables are Date, Customer, Account, Security, Trade_Type, CustomerScore

4   FACT

Fact tables is: Trades (date_id, customer_id, account_id, trade_type_id)

Query

Write a query to find how many trades were placed in July by customers in each activity segment.

Question 1

Option A	
Advantages	**Disadvantages**
Easy to update data	Not ideal for historical data storage or retrieval
Changes happen on the current or latest date	Not recommended for analytical data
Option B	
Encourages data consistency and integrity when new rows are created	The size of the entity increases as the update happens
Score update is straightforward	Cannot access inactive records.
Option C	
Easy to use for a transactional process	Score updates for the customers will require row by row record filter of an entire database table
Proper Fact value aggregation	The update process takes longer
Option D	
Normalization happens at this level	Too many dimension tables are created
Record updates are quick and easy	The outcome of a customer score cannot be enhanced In the Fact table

Question 6

Option A would be preferable. This is because this option conforms with the Star Schema architecture. Also, the fact table has a relationship with all the dimension tables as minimal dimension tables are created and because it's a Type 1 Slowly Changing Dimension, data filtration and update are easy to complete. The task also required the customer's score to be updated in the database to ensure data integrity and option B provided an easy platform to do this.

Question 7

Based on the given information, Option D would be chosen. This is because option D allows Normalization to occur, which will ensure easy data transformation. Data segmentation in the data warehouse will allow easy access to information and data integrity is maintained. With option D, data warehouse update happens in the outrigger table, which allows for a faster update process, rather than going through the entire table to do a row-by-row filter.

Question 8

Another reasonable alternative would be a Type 3 Slowly Changing Dimensions (Partial History).

Here the scores are the attributes of the Customer dimension. When scores change, the new score is updated in the destination table and the updated date will be the system date at the time. There are no in-between transactions, just the current and updated records in the database.

Advantages	Disadvantages
Access to recent and updated record	Increases memory consumption
Easy data transformation process	Cannot retrieve historical data