

## Tutorial 5: Data Warehousing

**INFS3200/7907**  
**Advanced Database Systems**

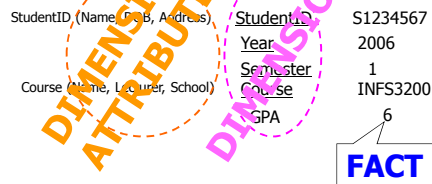
## Data Warehouse

- A data warehouse is a collection of computerised data that is organised to most optimally support **reporting** and **analysis** activity.
- This tutorial is of 2008 Olympic Games Attendance Report.

## Overview

- Question 1
  - Fact
  - Dimension and its Attributes
- Question 2
  - Fact and Dimension Table
  - Star Schema
  - Snowflake Schema
  - Drill Down
  - Roll Up
  - Pivoting Dimension

## Fact, Dimension and its Attributes



## Question 1 – Fact

Explain the use of facts, dimension, and attributes in the star schema. **Facts are numbers attending, value of ticket sales.**

- **Facts are numeric measurements (values)** that represent a specific business aspect or activity, e.g. number of attendance (**10,000 persons**), values of ticket sales (**\$10,000**).
- **Facts are normally stored in a fact table** which is the center of the **star schemas**.

## Q1 – Fact Continues

Explain the use of facts, dimension, and attributes in the star schema.

Olympiad 2008				
Olympic City	Beijing			
	Beijing Olympic Committee Email: beijing@olympics.org			
Region	Venue	Sport	Men	Women
Central-Enclosed	Pool	Swimming	10345	11011
		...	...	...
Remote	Rowing Area	Rowing	2156	2203

## Q1 – Fact Continues

Explain the use of facts, dimension, and attributes in the star schema.

Olympiad 2008				
Olympic City	Beijing			
	Beijing Olympic Committee Email: beijing@olympics.org			
Region	Venue	Sport	Men	Women
Central-Enclosed	Pool	Swimming	10345	11011
		...	...	...
Remote	Rowing Area	Rowing	2156	2203

Table Name: Attendance

**Olympiad**  
**Venue**  
**Sport**  
**Gender**  
**Attendance**

FACT TABLE

## Q1 – Dimensions

Explain the use of facts, dimensions, and attributes in the star schema. Dimensions include Olympiad, Venue, Sport, Type (heat, semifinal, final), men/women. Venues are classified by location and type of building into central-enclosed, central-open, remote. Sport subdivided into events.

- The fact table contain facts that are linked through their dimensions.
- Dimensions are qualifying characteristics that provide additional perspectives to a given fact.
- For examples
  - Olympiad Fact versus Olympiad Dimension
- Dimensions are normally stored dimension tables.

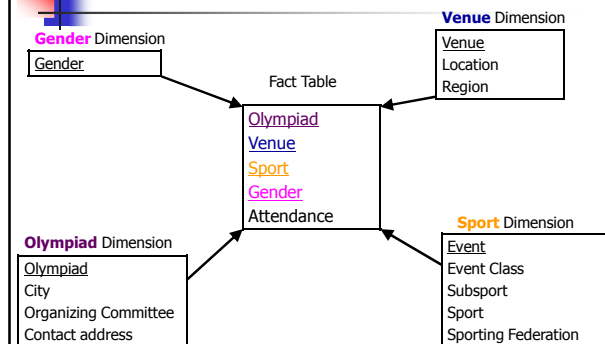
## Q1 – Attributes

Explain the use of facts, dimensions, and attributes in the star schema.

- Each dimension table contains attributes.
- The attributes are often used to search, filter, or classify facts.
- Dimensions provide descriptive characteristics about the facts through their attributes.
- Examples
  - Olympiad is a fact. The attributes of its dimension are the Olympiad, City, Organizing Committee and Contact address.

## Q1 –Continues

Explain the use of facts, dimensions, and attributes in the star schema.



## Question 2 – background

Olympiad 2008				
Olympic City	Beijing			
	Beijing Olympic Committee Email: beijing@olympics.org			
Region	Venue	Sport	Men	Women
Central-Enclosed	Pool	Swimming	10345	11011
		...	...	...
Remote	Rowing Area	Rowing	2156	2203

## Q2 (a)

Identify some important pieces of information available from this record. "Dimensions include Olympiad, Venue, Sport, Type (heat, semifinal, final), men/women. Venues are classified by location and type of building into central-enclosed, central-open, remote. Sport subdivided into events" from Question 1

- Olympiad
  - Olympiad 2008, the city, committee name and contact
- Region
  - Central Enclosed
    - Venue
      - Pool
        - Sport: Swimming, Diving, and so on
    - Central – Open ...
    - Remote ...
  - Gender
    - Men and Women

## Q2 (b) - Fact Table

Using the information found in (a) construct a Fact Table.  
What is its key?

Olympiad 2008				
Olympic City	Beijing			
	Beijing Olympic Committee Email: beijing@olympics.org			
Region	Venue	Sport	Men	Women
Central-Enclosed	Pool	Swimming	10345	11011
Remote	Rowing Area	Rowing	2156	2203

FACT TABLE

Olympiad  
Venue  
Sport  
Gender  
Attendance

The Key: Olympiad X Venue X Sport X Gender

## Q2 (c) – Star Schema

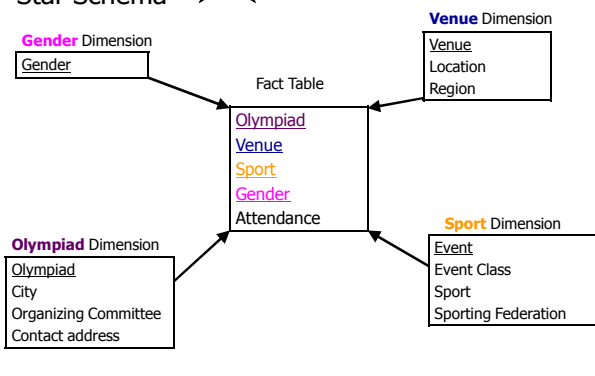
A star schema with the fact table designed in (b) above, and appropriate dimension tables.

The star schema consists of a fact table with a single table for each dimension.



## Q2 (c)

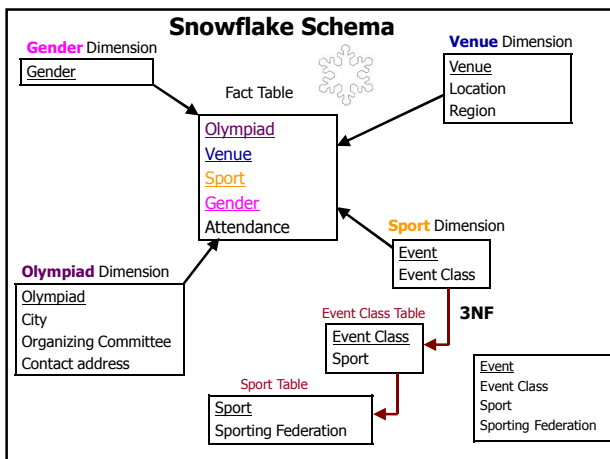
### Star Schema



## Q2 (d) – Snowflake Schema

A snowflake schema from (c) above, with the dimension tables normalized to 3NF (Third Normalization Form).

The snowflake schema is a variation of the star schema in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.



## Brief Review on Normalizations

- Normalization
  - Normalization is the restructuring of database fields and tables.
  - It aims at eliminating redundancy, organizing data efficiently, and reducing the potential for anomalies during data operations and improve data consistency.
- We will discuss three normal forms
  - 1<sup>st</sup> Normal Form
  - 2<sup>nd</sup> Normal Form
  - 3<sup>rd</sup> Normal Form

## Normalisation – 1<sup>st</sup> Normal Form

- The only attribute values permitted by 1NF are single atomic (or individual value), e.g. a sales person could have more than one phone number.

Employee ID	Phone number
S394849	0404040412, 34990000, 33749094
...	...

Employee ID	Phone number
S394849	0404040412
S394849	34990000
S394849	33749094

## Normalisation – 2<sup>nd</sup> Normal Form

- Is based on the concept of full functional dependency.

EMP_ID	Project_Number	Working_Hours	Employee_Name	Project_Name	Project_Location
--------	----------------	---------------	---------------	--------------	------------------

EMP_ID	Project_Number	Working_Hours
--------	----------------	---------------

EMP_ID	Employee_Name
--------	---------------

Project_Number	Project_Name	Project_Location
----------------	--------------	------------------

## Normalisation – 3<sup>rd</sup> Normal Form

- Is based on the concept of transitive dependency.

Employee_Name	EMP_ID	Date_of_Birth	Address	Dept_Number	Dept_Name	Dept_Manager_ID
---------------	--------	---------------	---------	-------------	-----------	-----------------

Employee_Name	EMP_ID	Date_of_Birth	Address	Dept_Number
---------------	--------	---------------	---------	-------------

Dept_Number	Dept_Name	Dept_Manager_ID
-------------	-----------	-----------------

## Q2 (e)- Drill Down

Perform a Drill down operation on the given data, and determine attendance by individual events in Swimming (you may assume attendance at a few swimming events you know about).

### Before Drill Down

Region	Venue	Sport	Men	Women
Central-Enclosed	Pool	Swimming	10345	11011

### After Drill Down

Region	Venue	Sport	Men	Women
Central-Enclosed	Pool	Swimming	10345	11011
		400 Metres freestyle	1200	1204
		200 Metres freestyle	1195	1195
		...	...	...

## Q2 (f) – Roll Up

Perform a Roll up operation on the given data and determine total attendance at the Stadium.

### Before Roll up

Region	Venue	Sport	Men	Women
Central-Open	Stadium	Track	22457	22783
		Field	9624	8333
	Total Stadium		32081	31116

### After Roll up

Region	Venue	Attendance
Central-Open	Stadium	63197

## Q2 (g) – Pivoting Dimension

How would the report change if you were to pivot on three dimensions: Sport, men/women and Event Type (heat/semifinal/final)

- What is current pivot dimension?
  - Venue, Sport, Men/Women
- Solution
  - Replacing *Venue* with *Event Type*