CP5806: ASSESSMENT 2 HELP SESSION

Presented by Sisi

JAMES COOK UNIVERSITY AUSTRALIA

SUBMISSION

- Due: **Week 5 Sunday 9/08 11:59pm AEST**
- Be around **2,400** words, excluding references (word counts 10% above the required word limit will be penalised by 10% deduction of the marks available. The word count **must be accurately stated at the end of the written piece**. Every printed element between spaces is to be counted **including quotations and in-text references** (but not including reference list or appendices)
- Be **less than 16** A4 pages and in **12pt Arial** font
- APA 6th ed.style for both **in-text citations** and **reference list**:

Ponniah, P. (2010). Data Warehousing Fundamentals for IT Professionals (2nd ed.). Hoboken, NJ: Wiley.

….xxxxxx (Ponniah, 2010).

Ponniah(2010) states that xxxxx….

Ponniah(2010, p45) writes "…..xxxxxx…".

https://libguides.jcu.edu.au/ld.php?content_id=40460230

STEP 1: BUSINESS SCENARIO PHASE (5%)

Scope described in Step 3 and Step 5:

- **The number of dimension tables should be between 6-8 (including sub-dimensions)**
- **The number of metrics and measures should be between 4-5;**
- **Build a case that shows how your fact table grows to 10 billion records with your dimensions**

JCU has four campuses: Cairns, Townsville, Brisbane and Singapore. JCU offers subjects to students and lecturers teach subjects to students. PLEASE note that we make an example as simple as possible. We don't want to describe all the details JCU does but only some simple operations.

A possible scenario:

In the case of higher educator James Cook University (JCU), there are metrics related to the university performance. These are the quantitative numbers that tell the users about its performance. The set of meaningful and useful metrics is:

- Total students

And there are several business dimensions with which JCU would like to analyse these metrics. They include as follows:

- Uni organisation
- Student
- Course
- Campus.

STEP 2: INFORMATION PACKAGE DIAGRAM PHASE (5%)

Uni organisation	**Student**	**Course**	**Campus**
Division	Type	Post/under graduate	Country
College (CSE or CBLG etc)	Status	Degree	City
Academic group (such as IT or business)	Name	Code	

Measured Fact: Total students

So, each column could be rolled up to the next hierarchy to get an aggregated value. For instance, if we roll up in Uni organisation, we can get how many students for each division for each course for each campus. And if we roll up in campus, we can get how many students for each academic group in each degree for each country and so on so forth;

STEP 3: DATA DESIGN PHASE (5%)

Uni organisation
Organisation_key (PK)
Name
Fields
Office

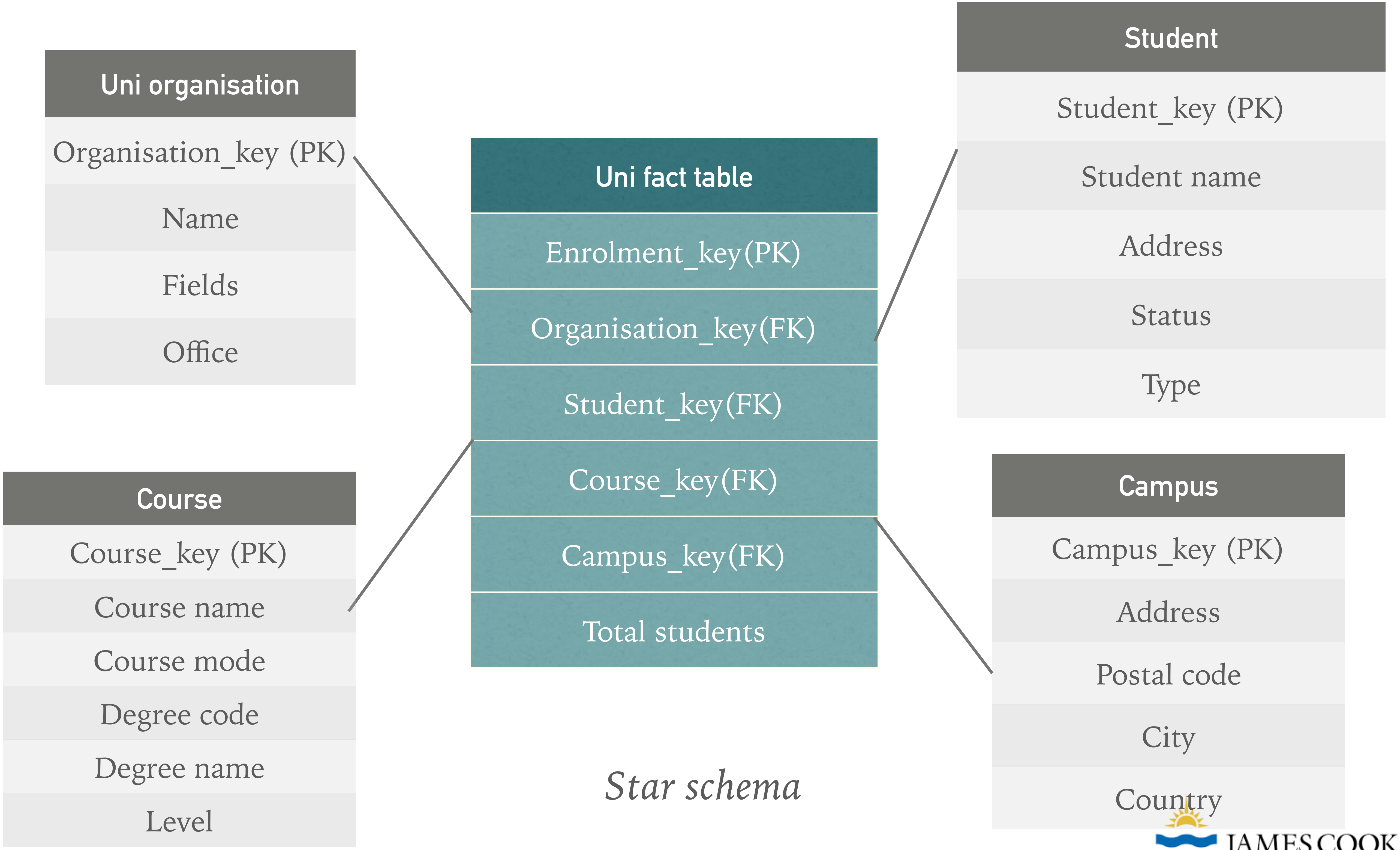
Student
Student_key (PK)
Student name
Address
Status
Type

Uni fact table
Enrolment_key (PK)
Organisation_key(FK)
Student_key(FK)
Course_key(FK)
Campus_key(FK)
Total students

Course
Course_key (PK)
Course name
Course mode
Degree code
Degree name
Level

Campus
Campus_key (PK)
Address
Postal code
City
Country

STEP 4: DIMENSIONAL MODELLING PHASE (5%)



Uni organisation
Organisation_key (PK)
Name
Fields
Office

Course
Course_key (PK)
Course name
Course mode
Degree code
Degree name
Level

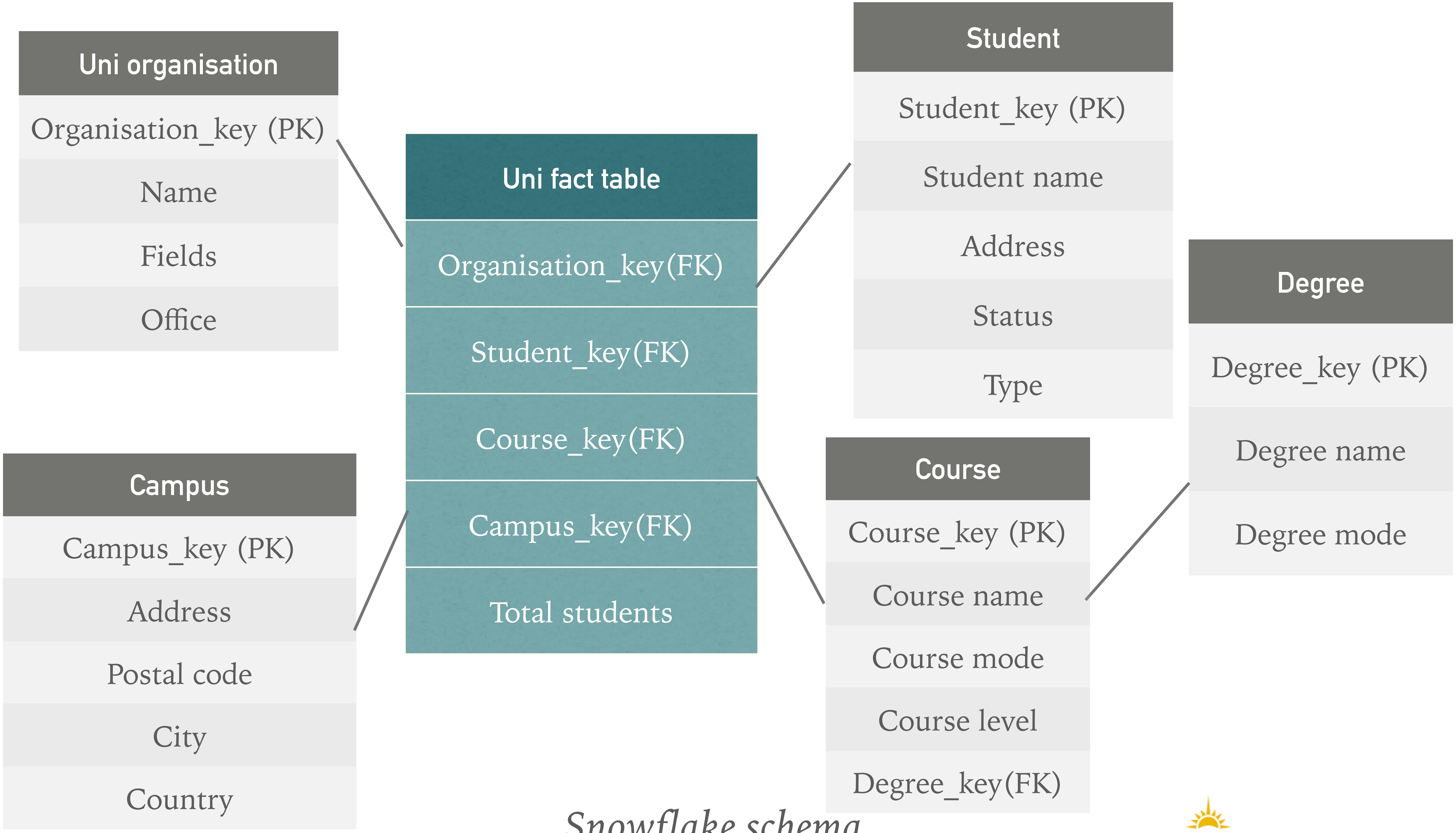
Uni fact table
Enrolment_key(PK)
Organisation_key(FK)
Student_key(FK)
Course_key(FK)
Campus_key(FK)
Total students

Student
Student_key (PK)
Student name
Address
Status
Type

Campus
Campus_key (PK)
Address
Postal code
City
Country

Star schema

STEP 4: DIMENSIONAL MODELLING PHASE (5%)



Uni organisation
Organisation_key (PK)
Name
Fields
Office

Uni fact table
Organisation_key(FK)
Student_key(FK)
Course_key(FK)
Campus_key(FK)
Total students

Student
Student_key (PK)
Student name
Address
Status
Type

Degree
Degree_key (PK)
Degree name
Degree mode

Campus
Campus_key (PK)
Address
Postal code
City
Country

Course
Course_key (PK)
Course name
Course mode
Course level
Degree_key(FK)

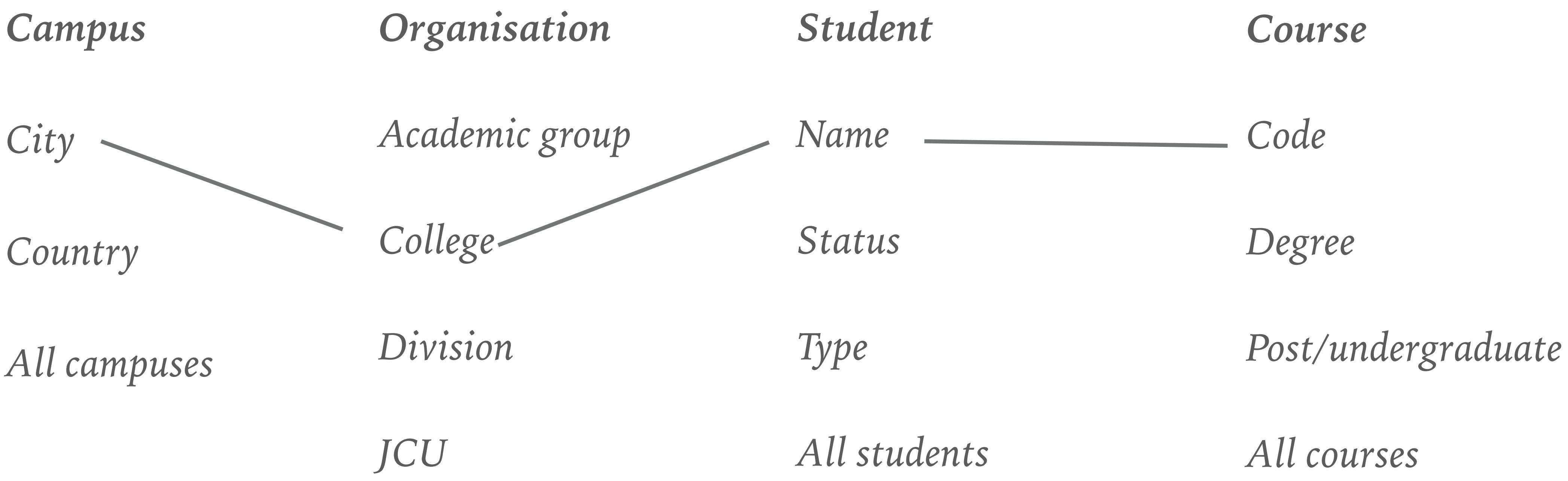
Snowflake schema

STEP 5: THE SIZE OF FACT TABLE PHASE (5%)

- Number of records for fact table at base level:
  - **Campus**: 4 campuses
  - **Course**: 30 courses for 10 master degrees + 60 courses for 20 bachelor degrees + 7 diplomas/advanced diplomas + 1 certificate = 98 courses
  - **Student**: 6,000 students
  - **Uni organisation**: 5 academic organisations x 5 colleges x 2 divisions
  - **Total: 4 x 98 x 6,000 x 50 = 117,600,000 records**
- Aggregated to students for each division for each degree in each country:
  - **Campus**: 2 countries
  - **Course**: 10 master + 20 bachelor + 2 diplomas + 1 certificate = 33 degrees
  - **Student**: 6,000 students
  - **Uni organisation**: 2 divisions
  - **Total: 2 x 33 x 6,000 x 2 = 792,000 records**

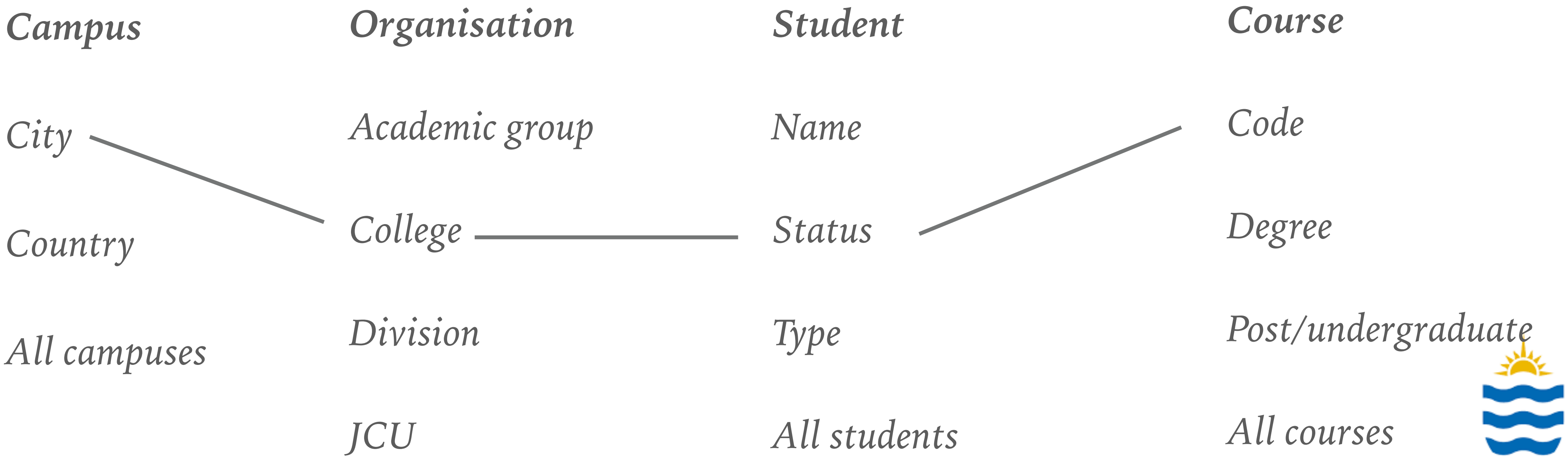
STEP 6: AGGREGATING FACT TABLE PHASE (5%)

- One-way aggregate:



Campus	Organisation	Student	Course
City	Academic group	Name	Code
Country	College	Status	Degree
All campuses	Division	Type	Post/undergraduate
	JCU	All students	All courses

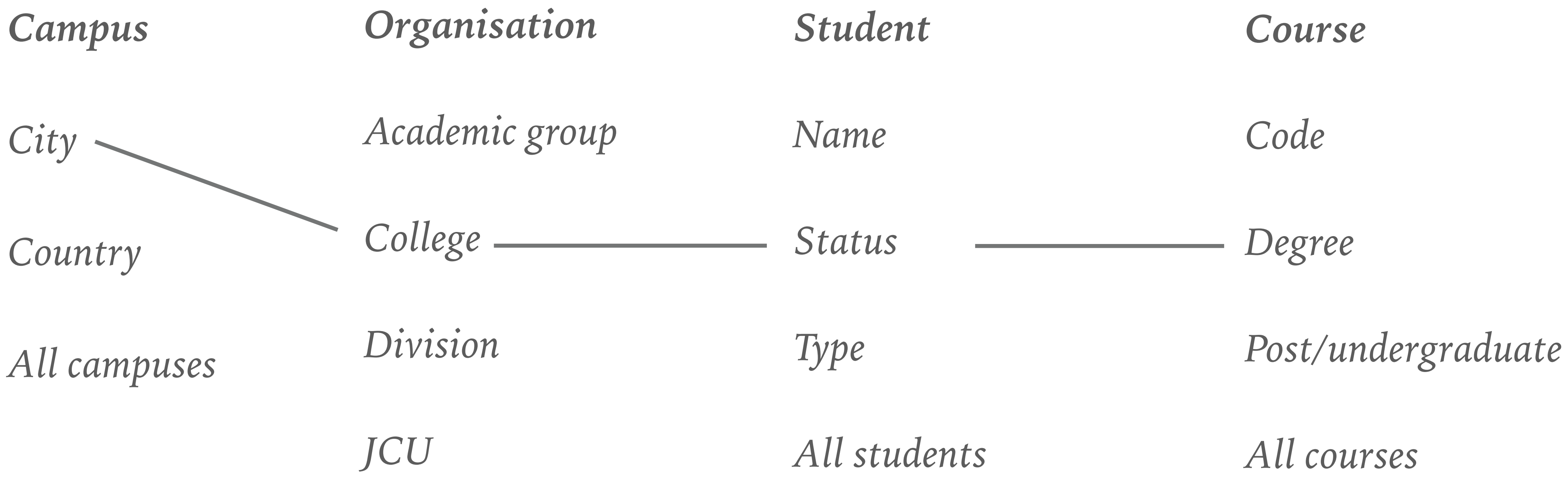
- Two-way aggregate:



Campus	Organisation	Student	Course
City	Academic group	Name	Code
Country	College	Status	Degree
All campuses	Division	Type	Post/undergraduate
	JCU	All students	All courses

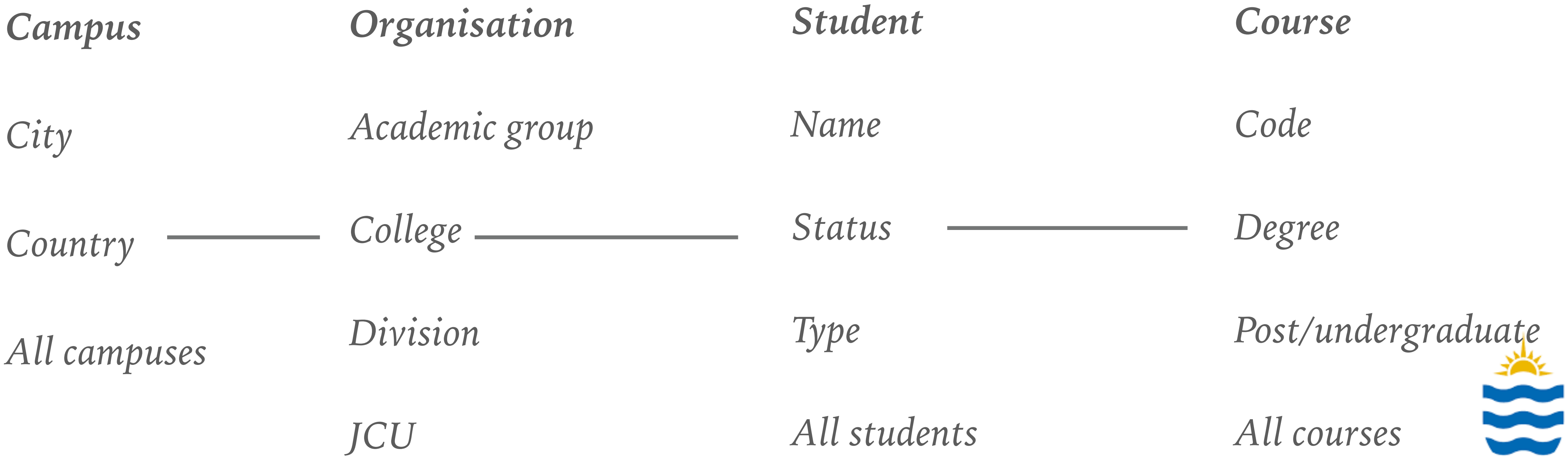
STEP 6: AGGREGATING FACT TABLE PHASE (5%)

- Three-way aggregate:



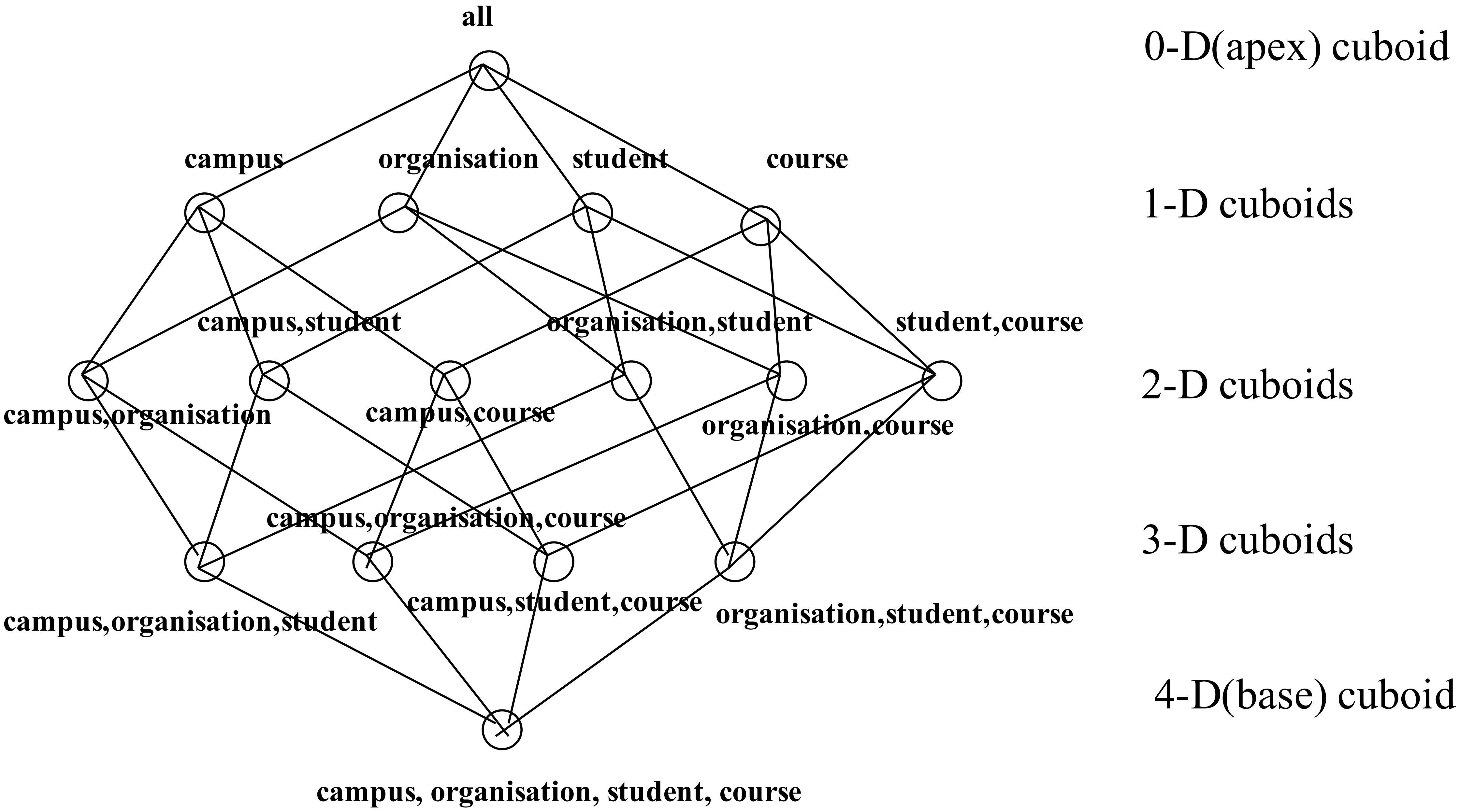
Campus	***Organisation***	***Student***	***Course***
City	*Academic group*	*Name*	*Code*
Country	*College*	*Status*	*Degree*
All campuses	*Division*	*Type*	*Post/undergraduate*
	JCU	*All students*	*All courses*

- Four-way aggregate:



Campus	***Organisation***	***Student***	***Course***
City	*Academic group*	*Name*	*Code*
Country	*College*	*Status*	*Degree*
All campuses	*Division*	*Type*	*Post/undergraduate*
	JCU	*All students*	*All courses*

STEP 7: A LATTICE OF CUBOIDS PHASE (5%)



*nCr = n! / r! * (n - r)!*

*e.g. 1-D cuboids = 4C1 = 4! / 1! * (4-1)!*

*= 4 * 3 * 2 * 1/ 1 * 3 * 2 * 1 = 4*

STEP 8: DATA CUBE COMPUTATION (5%)

- **Assume a base cuboid of 5 dimensions contains two base cells (a1, a2, a3, a4, a5) and (b1, b2, a3, a4, a5), measure is count;**

a) How many nonempty aggregate (i.e., nonbase) cells will a full cube contain?

c) How many nonempty aggregate cells will an iceberg cube contain if the condition of the iceberg cube is “count ≥ 2”?

d) A cell, c , is a closed cell if there exists no cell, d, such that d is a specialisation of cell c (i.e., d is obtained by replacing a ∗ in c by a non-∗ value) and d has the same measure value as c. A closed cube is a data cube consisting of only closed cells.

STEP 8: DATA CUBE COMPUTATION (5%)

- Assume a base cuboid of 5 dimensions contains two base cells (a1, a2, a3, a4, a5) and (b1, b2, a3, a4, a5), measure is count;

Solutions:

a) For each base cell, 2 ^5 - 1 aggregate cells:

1: (*, *, *, *, *);

5: (a1, *, *, *, *),(*, a2, *, *, *),(*, *, a3, *, *),(*, *, *, a4, *),(*, *, *, *, a5);

10: (a1, a2, *,*,*), (a1, *, a3, *, *), (a1, *, *, a4, *), (a1, *, *, *, a5),(*, a2, a3, *, *), (*, a2, *, a4, *), (*, a2, *, *, a5), (*,*,a3,a4,*),(*,*,a3, *, a5), (*,*,*,a4,a5);

10: (a1,*,*,a4,a5), (a1,a2,*,*,a5), (a1,a2,a3,*,*), (a1,*,a3,*,a5), (a1,*,a3,a4,*)(a1,a2,*,a4,*), (*, a2,*,a4,a5), (*, a2,a3,*,a5), (*,a2,a3,a4,*),(*,*,a3,a4,a5);

5: (a1,a2,a3,a4,*), (a1,*,a3,a4,a5),(a1,a2,*,a4,a5), (a1,a2,a3,*,a5) ,(*,a2,a3,a4,a5);

Current total: 2 * (2 ^ 5 - 1)

There are 2 ^ 3 cells calculated twice:

(*,*, a3, a4, a5), (*,*,a3,a4,*), (*,*,a3,*,a5),(*,*,*,a4,a5), (*,*,a3,*,*), (*,*,*,a4,*), (*,*,*,*,a5), (*,*,*,*,*)

Final aggregate cells: 2 ^6 - 2 - 8 = 2 ^ 6 - 10

STEP 8: DATA CUBE COMPUTATION (5%)

- Assume a base cuboid of 5 dimensions contains two base cells (a1, a2, a3, a4, a5) and (b1, b2, a3, a4, a5), measure is count;

Solutions:

b) There are 2 ^ 3 cells calculated twice:

(*,*, a3, a4, a5), (*,*,a3,a4,*), (*,*,a3,*,a5),(*,*,*,a4,a5), (*,*,a3,*,*), (*,*,*,a4,*), (*,*,*,*,a5), (*,*,*,*,*)

All these cells are aggregate cells, therefore 8 aggregate cells will an iceberg cube contain if the condition of the iceberg cube is "count ≥ 2"

STEP 8: DATA CUBE COMPUTATION (5%)

- Assume a base cuboid of 5 dimensions contains two base cells (a1, a2, a3, a4, a5) and (b1, b2, a3, a4, a5), measure is count;

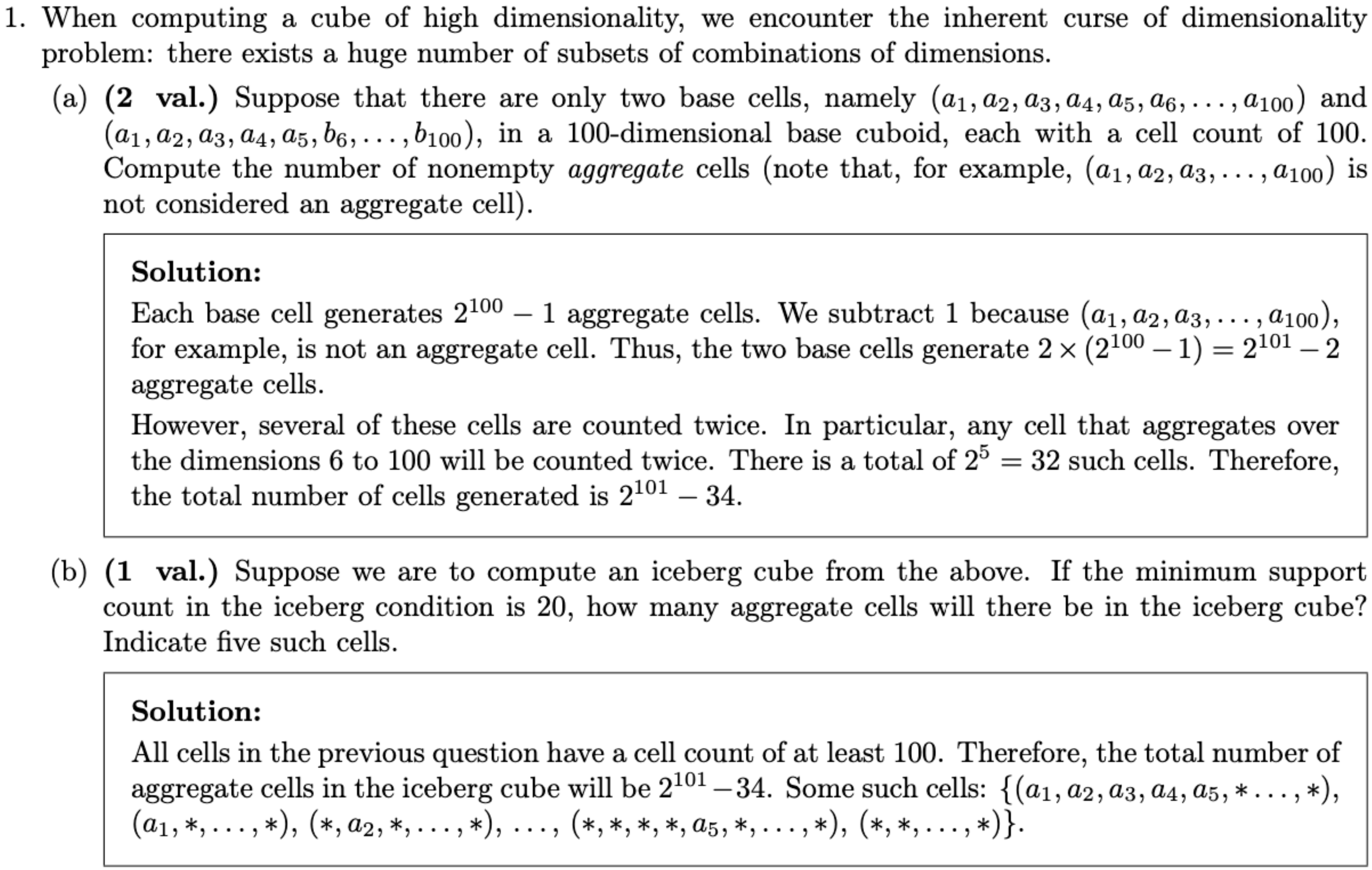
Solutions:

c) Closed cell: no descendant cell d that has the same measurement value, e.g. cell (a1, a2, a3, a4, a5) is a descendant of cell (a1, *, *, a4, a5) or cell (*,*,*, a3,a4,a5) is a descendant of cell (*, *,*,*,a5) etc.

2 base cells are closed cells as they do not have descendant cells;

(*, *, a3, a4, a5) is a closed cell as all its descendant cells, i.e. (a1, *,a3,a4,a5) or (*, b2,a3,a4,a5) only have a count of 1;

STEP 8: DATA CUBE COMPUTATION (5%)



1. When computing a cube of high dimensionality, we encounter the inherent curse of dimensionality problem: there exists a huge number of subsets of combinations of dimensions.

   (a) **(2 val.)** Suppose that there are only two base cells, namely $(a_1, a_2, a_3, a_4, a_5, a_6, \ldots, a_{100})$ and $(a_1, a_2, a_3, a_4, a_5, b_6, \ldots, b_{100})$, in a 100-dimensional base cuboid, each with a cell count of 100. Compute the number of nonempty *aggregate* cells (note that, for example, $(a_1, a_2, a_3, \ldots, a_{100})$ is not considered an aggregate cell).

   Solution:

   Each base cell generates $2^{100} - 1$ aggregate cells. We subtract 1 because $(a_1, a_2, a_3, \ldots, a_{100})$, for example, is not an aggregate cell. Thus, the two base cells generate $2 \times (2^{100} - 1) = 2^{101} - 2$ aggregate cells.

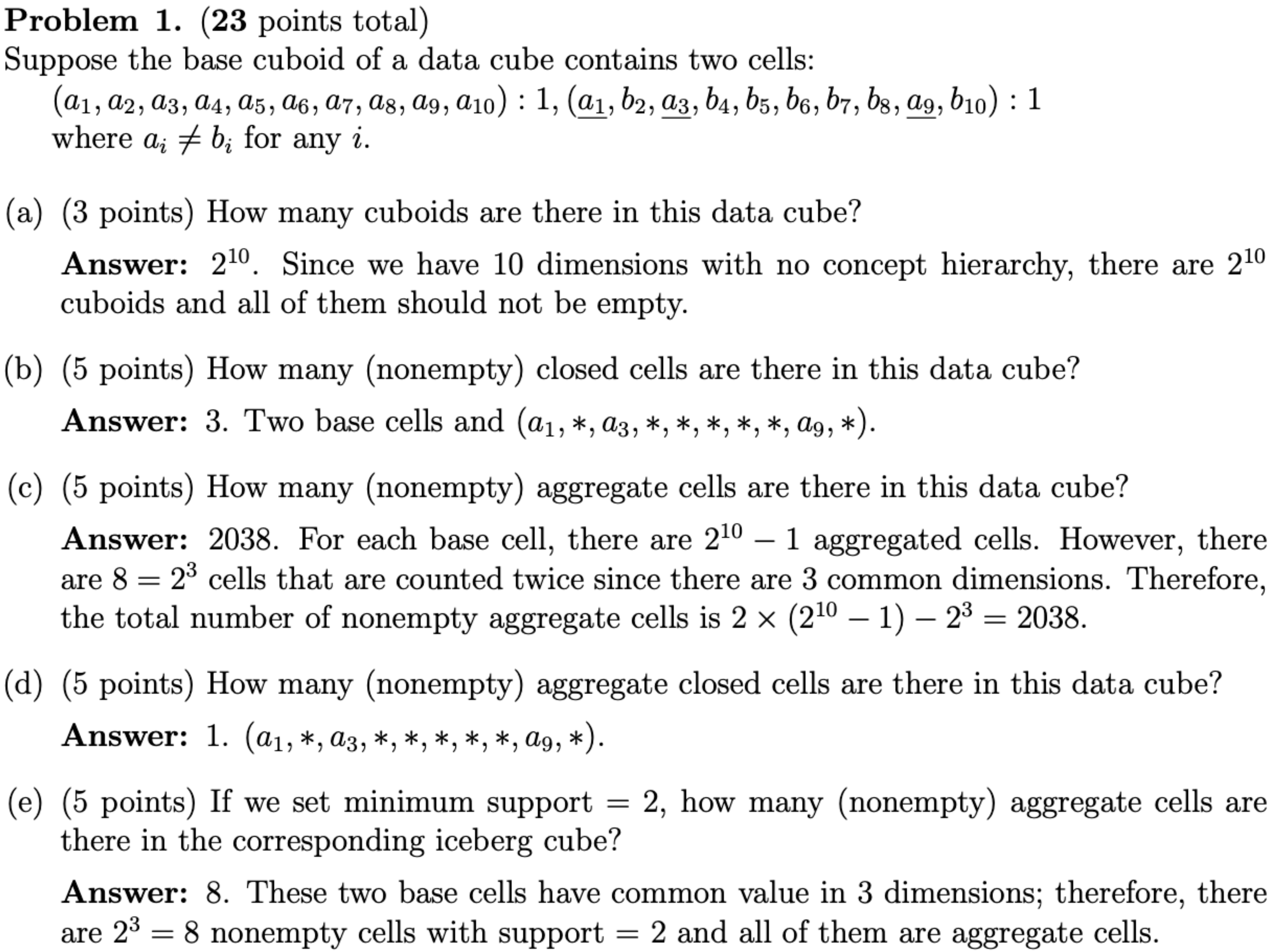
   However, several of these cells are counted twice. In particular, any cell that aggregates over the dimensions 6 to 100 will be counted twice. There is a total of $2^5 = 32$ such cells. Therefore, the total number of cells generated is $2^{101} - 34$.

   (b) **(1 val.)** Suppose we are to compute an iceberg cube from the above. If the minimum support count in the iceberg condition is 20, how many aggregate cells will there be in the iceberg cube? Indicate five such cells.

   Solution:

   All cells in the previous question have a cell count of at least 100. Therefore, the total number of aggregate cells in the iceberg cube will be $2^{101} - 34$. Some such cells: $\{(a_1, a_2, a_3, a_4, a_5, * \ldots, *)$, $(a_1, *, \ldots, *)$, $(*, a_2, *, \ldots, *)$, $\ldots$, $(*, *, *, *, a_5, *, \ldots, *)$, $(*, *, \ldots, *)\}$.

STEP 8: DATA CUBE COMPUTATION (5%)



Problem 1. (**23** points total)
Suppose the base cuboid of a data cube contains two cells:

$(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}) : 1, (\underline{a_1}, b_2, \underline{a_3}, b_4, b_5, b_6, b_7, b_8, \underline{a_9}, b_{10}) : 1$
where $a_i \neq b_i$ for any i.

(a) (3 points) How many 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.

(b) (5 points) How many (nonempty) closed cells are there in this data cube?

Answer: 3. Two base cells and $(a_1, *, a_3, *, *, *, *, *, a_9, *)$.

(c) (5 points) How many (nonempty) aggregate cells are there in this data cube?

Answer: 2038. For each base cell, there are $2^{10} - 1$ aggregated cells. However, there are $8 = 2^3$ cells that are counted twice since there are 3 common dimensions. Therefore, the total number of nonempty aggregate cells is $2 \times (2^{10} - 1) - 2^3 = 2038$.

(d) (5 points) How many (nonempty) aggregate closed cells are there in this data cube?

Answer: 1. $(a_1, *, a_3, *, *, *, *, *, a_9, *)$.

(e) (5 points) If we set minimum support $= 2$, how many (nonempty) aggregate cells are there in the corresponding iceberg cube?

Answer: 8. These two base cells have common value in 3 dimensions; therefore, there are $2^3 = 8$ nonempty cells with support $= 2$ and all of them are aggregate cells.