Average in the Fact?

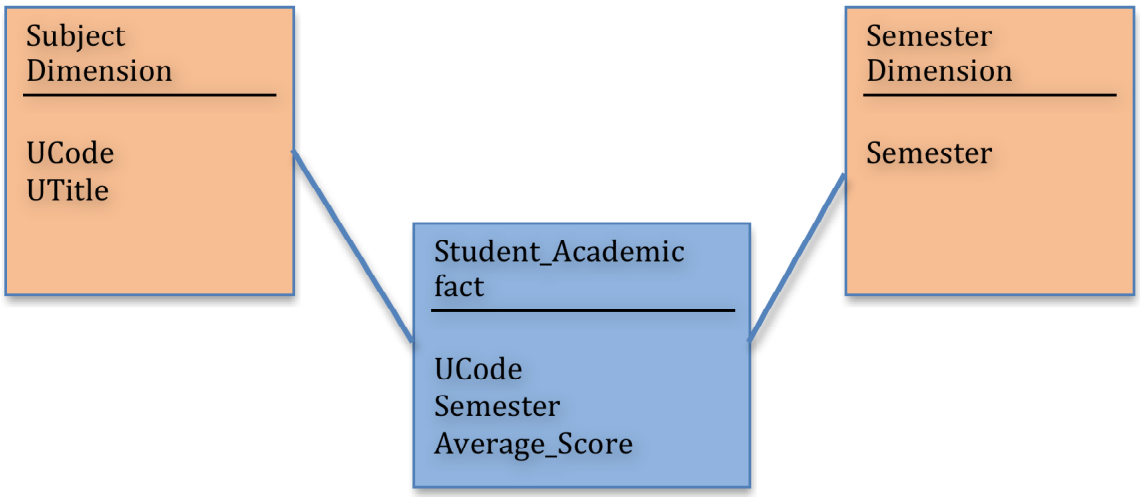
Should we store an "average" fact measure in the fact table?
No!!!!!!!

Consider the following example. Suppose we have the following 16 records as our data source in the operational database. Note that there are 9 Semester One and 7 Semester Two records, respectively. Out of the 8 Database unit recodes, 6 of them are Semester One, and 2 of them are Semester Two.

An Operational Database

Ucode	Utitle	Semester	Sfname	Score
IT001	Database	1	Mirriam	81
IT001	Database	1	Allan	41
IT001	Database	1	Ben	74
IT001	Database	1	Kate	85
IT001	Database	1	Larry	87
IT001	Database	1	Leonard	75
IT001	Database	2	Juan	64
IT001	Database	2	Andy	32
IT002	Java	1	Ally	65
IT002	Java	1	Menson	47
IT002	Java	2	Mirriam	78
IT002	Java	2	Ben	73
IT002	Java	2	Larry	64
IT003	SAP	1	Ally	63
IT004	Network	2	Juan	53
IT004	Network	2	Menson	52

The star schema of the above operational database contains one fact and two dimensions. The dimensions are: Subject and Semester (One or Two); and the fact measure is Average Score.



The fact table aggregates these score records based on their dimensions, which are subject and semester. If we store Average Score in the fact table, this is how the fact table will look like:

Fact Table

UCode	Semester	Average_Score
IT001	1	73.833
IT001	2	48
IT002	1	56
IT002	2	71.667
IT003	1	63
IT004	2	52.5

The dimension tables look like as follows:

SubjectDIM Table

UCode	UTitle
IT001	Database
IT002	Java
IT003	SAP
IT004	Network

SemesterDIM Table

Semester
1
2

Looking at the Fact Table, the average score for the unit Database in Semester One is 73.833 (average of the first 6 score records); the average score for the unit Database in Semester Two is 48 ((64 + 32)/2).

Is this fact table correct?
It looks correct. But actually it is incorrect.

For example, if we want to query the fact table to find out what is the average score of the Database unit, by looking at the above fact table, the answer would be (73.833+48)/2= **60.9165**.

The SQL to query the Fact Table is as follows:

```
Select Avg(Average_Score)
From FactTable
Where UCode  = 'IT001';
```

Is this correct? No.

In the Operational Database, there are 8 records for Database unit in Semester One and Two (see the first eight records in the operational database). If we sum

all the score of these eight records and divided by eight records, the result will be 539/8=**67.375;** not 60.9165.

Let's do further comparisons:
The average score for Java unit in Semester One and Two using the above fact is (56+71.667)/2=**63.833**. The actual average score for Java unit in Semester One and Two is not 63.833, but **65.4** (see the next 5 records in the above score list, and sum these scores and then divide by 5, 327/5). So again, the above fact table, which stores the average score, will not produce correct results.

Ok, now let's calculate further. The average score for Semester One using the above fact is (73.833+56+63)/3=**64.278**. In the above score list records, there are nine Semester One records, and the average is in fact **68.667**.

For Semester Two, using the above fact the average scare for Semester Two is (48+71.667+52.5)/3=**57.389**; whereas the actual average score for the seven Semester Two records is **59.4286**.

So, storing average as a fact measurement is not a good idea.
How do we solve the above problems?

In the fact table, we should store the "total score" and "number of students" in each aggregate group. Hence, the fact table should look like this.

Fact Table 2

UCode	Semester	Total_Score	NumberofStudents
IT001	1	443	6
IT001	2	96	2
IT002	1	112	2
IT002	2	215	3
IT003	1	63	1
IT004	2	105	2

Note that the dimension tables remain unchanged:

SubjectDIM Table

UCode	UTitle
IT001	Database
IT002	Java
IT003	SAP
IT004	Network

SemesterDIM Table

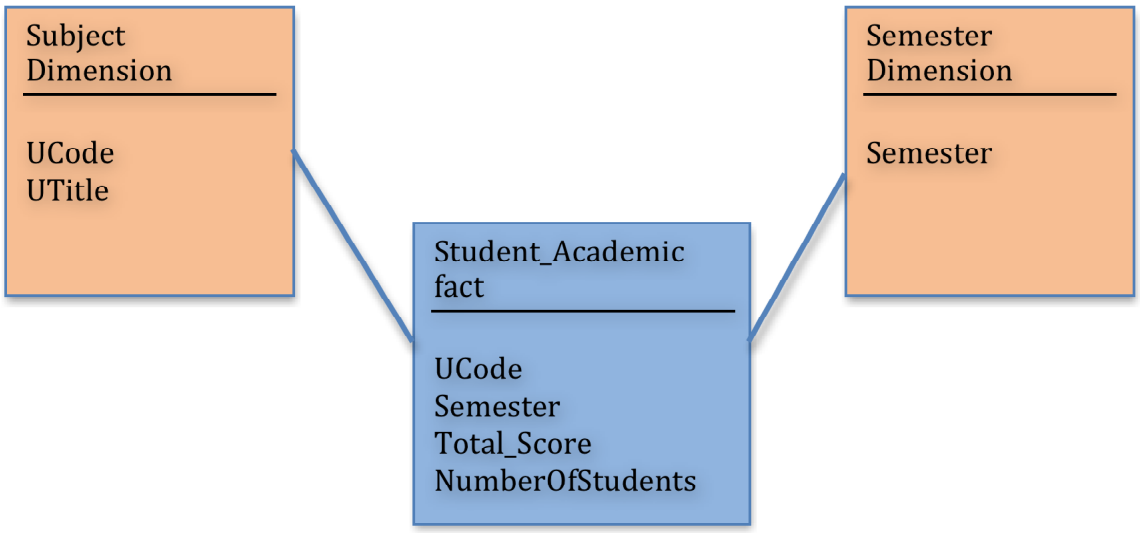
Semester
1
2

Using the correct fact table above, it is easy to calculate the average score of Database unit, which is (443+96)/(6+2)=67.375.

The SQL to query the Correct Fact Table is as follows:

```
Select Sum(Total_Score)/Sum(NumberofStudents)
From FactTable2
Where UCode  = 'IT001';
```

The correct star schema is then as follows:



Conclusion

The problem of AVG in the Fact is known as the "Average of an Average" problem. This problem is well known in Mathematics. Average of an average will simply produce an incorrect average result (almost all the time). Hence, it is not desirable to have an average measure in the fact – unless the analysis **ALWAYS** uses all the dimensions.

How about Min or Max in the Fact? Can we do it? – Yes we can.
Because Max of Max is always a global max, and Min of Min is always a global min. For example, using the above sample data, assume we have Max_Score and Min_Score in the Fact, as follows:

Fact Table 3

Ucode	Semester	Min_Score	Max_Score
IT001	1	41	87
IT001	2	32	64
IT002	1	47	65
IT002	2	64	78
IT003	1	63	63
IT004	2	52	53

SubjectDIM Table

UCode	UTitle
IT001	Database
IT002	Java
IT003	SAP
IT004	Network

SemesterDIM Table

Semester
1
2

(*Note: the dimension tables are unchanged)

Assuming we want to get the Max_Score of IT001, then the max of {87, 64} will produce 87, and **87** is the maximum score of IT001, because 87 is the max in semester 1, which is greater than any max of IT001 (e.g. in semester 2). In other words, "Max of Max" is correct.

The SQL to retrieve the maximum score of IT001 is as follows:

```
Select Max(Max_Score)
From FactTable3
Where UCode  = 'IT001';
```

The same applies to "Min of Min". If we want to get the minimum score of IT001, the result will be **32**, which is the minimum between 41 and 32.

```
Select Min(Min_Score)
From FactTable3
Where UCode  = 'IT001';
```

We certainly don't want to mix between min and max. For example, retrieving the minimum of Max_Score would be meaningless; the same as retrieving the maximum of Min_Score.

As a final conclusion:

- **Average** in the fact is not desirable, although technically it satisfies the two criteria of the fact (e.g. must be a numerical and aggregate value)
- **Min** and **Max** in the fact can still be used, since min_score and max_score are valid fact measures (e.g. they are numerical and aggregated values)
- In general, count and sum are more common. **Count** is "number of", and **Sum** is "total of".