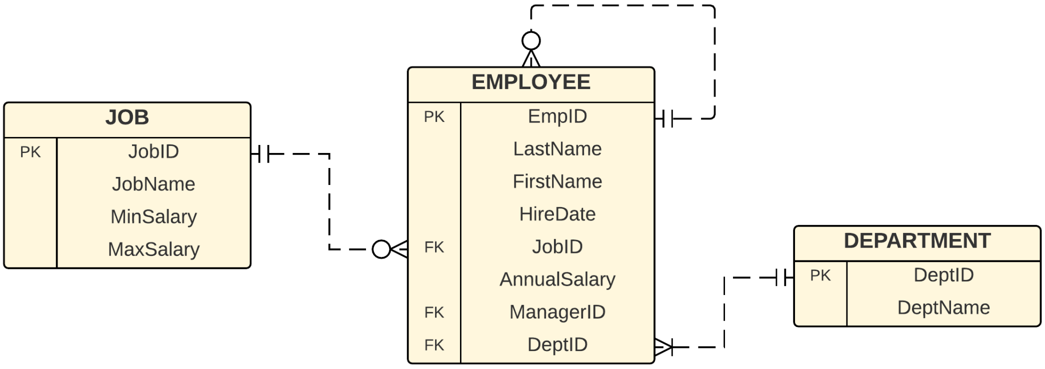
Level of Aggregation and Determinant Dimensions

In this lesson, we are going to learn the impact of determinant dimension on level of aggregation. Considering the following E/R diagram that maintains a list of employees, together with their departments and jobs. Each employee has a department and a Job ID.



The sample records are as follows. For simplicity, the details of some attributes are not displayed. However, pay a particular attention to the Hire Date attribute. This indicates when each employee started the job.

Employee Table

EmpID	LastName	FirstName	HireDate	JobID	AnnualSalary	ManagerID	DeptID
101	Koh	Katie	1-Nov-2016	SRep			D01
102	Li	Liam	1-Feb-2017	SRep			D01
103	Mao	Mary	1-May-2017	SRep			D01
104	Qi	Queeny	1-May-2017	Acc			D02
...	...	...	...				

Department Table

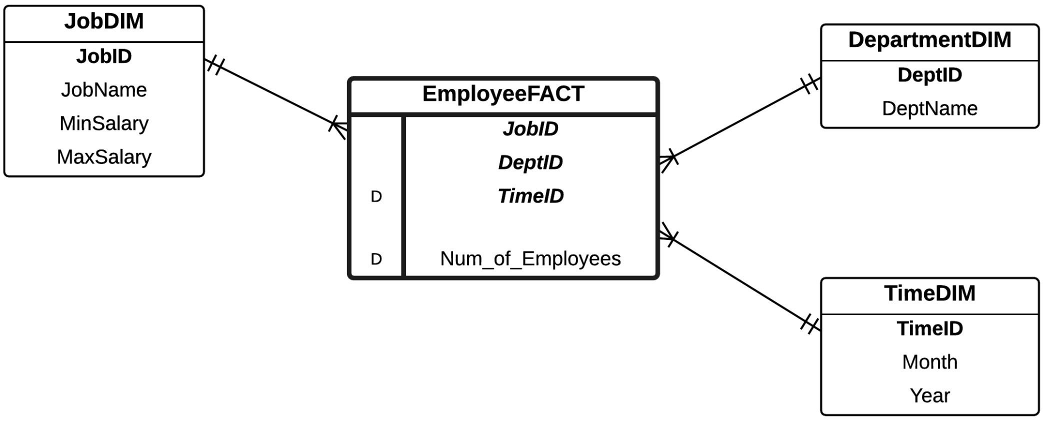
DeptID	DeptName
D1	Cosmetic
D2	Finance
...	...

Job Table

JobID	JobName	MinSalary	MaxSalary
SRep	Sales Representative		
Acc	Accountant		
...	...	...	...

The following is a star schema that captures the number of employees for each job department, and time (month). The calculation of the fact measure, namely Number of

Employees, is not that straightforward. Employee 101 started her job in November 2016, and she was the only employee on that month, until January 2017. In February 2017, Employee 102 started the job. This means that in February 2017, there were two employees (employees 101 and 102). Therefore, the TimeID is not only the Hire Date, but also the months after the hire date.



Let's assume that the data warehouse captures only for the duration from November 2016 to June 2017 (see the Time Dimension table below). The Fact table is shown as follows:

Time Dimension

TimeID	Month	Year
201611	Nov	2016
201612	Dec	2016
201701	Jan	2017
201702	Feb	2017
201703	Mar	2017
201704	Apr	2017
201705	May	2017
201706	June	2017

Employee Fact

JobID	**DeptID**	**TimeID**	Num of Employees
SRep	D01	201611	1
SRep	D01	201612	1
SRep	D01	201701	1
SRep	D01	201702	2
SRep	D01	201703	2
SRep	D01	201704	2
SRep	D01	201705	3
SRep	D01	201706	3
Acc	D02	201705	1
Acc	D02	201706	1

The SQL command to create the Employee Fact table is as follows. Assume that the TimeDim table has been created, which consists of records showing each month between November 2016 and June 2017. When creating the Employee Fact table, we need to join the Employee table and the TimeDim table, and the join condition is Hire Date <= TimeID. This means that the months after the hire date will capture that that employee is still working. Consequently, when counting the number of employees on any month after the hire date, this employee will be counted.

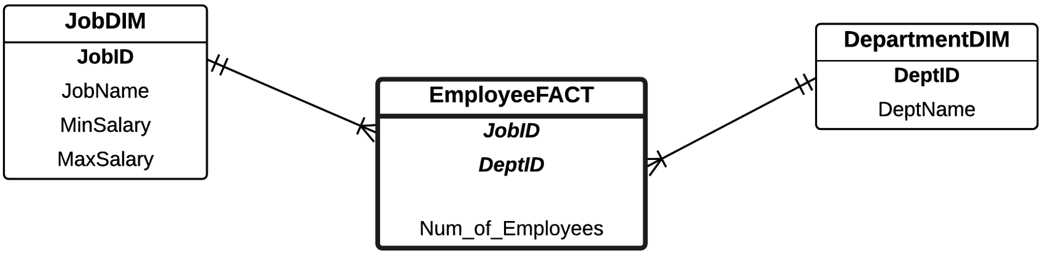
```
create table EmployeeFact
as select E.JobID, E.DeptID, T.TimeID, Count(*) as Num_of_Employees
from Employee E, TimeDim T
where to_char(E.HireDate, ‘YYYYMM’) <= T.TimeID
group by JobID, DeptID, TimeID;
```

When querying the Employee Fact about Number of Employees, the Time Dimension must always be used; hence the Time Dimension is a Determinant Dimension, as shown in the above star schema. One example is to ask “how many employees were working on March 2017?”, or “how many employees were working on June 2017?”. The answers are 2 and 4 respectively. The SQL for the second query is as follows:

```
select TimeID, sum(Num_of_Employees)
from EmployeeFact
where TimeID = ‘201706’
group by TimeID;
```

In the previous lesson we have learned that to increase the level of aggregation (to make the star schema higher level of granularity, or more general), you can add with a dimension to the star schema. The opposite is also applied. To lower down the level of aggregation, we can remove a dimension.

Suppose we want to take out the Time Dimension from the above star schema. The new star schema is shown below. The new star schema now has only two dimensions: Job and Department dimensions.



When moving up or down the level of aggregation, normally, the value in the fact measure is broken down (when we lower down the level of aggregation), or is aggregated or summed (when we increase the level of aggregation). For example, in the case of Num of Employees, when we remove the Time dimension, we expect that the Num of Employees of the same time record will simply be aggregated or summed. However, when we do this, we will get incorrect number of employees, as shown in the table below. There are no 15 employees doing a Sales Rep job in Department D01. Therefore, simply summing up records from the lower level of aggregation is not correct in this case study.

Employee Fact (**INCORRECT**)

JobID	**DeptID**	Num of Employees
SRep	D01	15
Acc	D02	2

The correct Employee Fact should be as follows. There are 3 employees (not 15) doing a Sales Rep job in Department D01.

Employee Fact (**CORRECT**)

JobID	**DeptID**	Num of Employees
SRep	D01	3
Acc	D02	1

The main question is why simply summing up the fact measure when we move to a higher level of aggregation does not work in this case? The answer is because the dimension that we remove is a "Determinant Dimension". Therefore, removing a determinant dimension implies that we need to recalculate the fact measure; simply aggregating up or summing up will not produce the correct results.

The SQL command to produce the new Employee Fact is as follows. Note that joining with the TimeDim is removed, because there is no TimeDim in this level of star schema.

```
create table EmployeeFact
as select E.JobID, E.DeptID, Count(*) as Num_of_Employees
from Employee E
group by JobID, DeptID;
```

The problem will be more complex if there is an attribute Cease Date for each employee record. For example, Employee 101 quits her job at the end of March 2017. If there is Time Dimension (as in the first star schema above), there will not be a problem in counting the Number of Employees, because the counting is done month-by-month. In this case, there were 2 employees in Feb and Mar 2017, but in Apr 2017, there was only 1 employee (not 2), because 1 employee has quit her job.

Employee Table

EmpID	LastName	FirstName	HireDate	CeaseDate	JobID	AnnualSalary	ManagerID	DeptID
101	Koh	Katie	1-Nov-2016	31-Mar-2017	SRep			D01
102	Li	Liam	1-Feb-2017	null	SRep			D01
103	Mao	Mary	1-May-2017	null	SRep			D01
104	Qi	Queeny	1-May-2017	null	Acc			D02
…	…	…	…					

The SQL to create the EmployeeFact table becomes like this. Note that it needs to incorporate the CeaseDate attribute in the join condition.

```
create table EmployeeFact
as select E.JobID, E.DeptID, T.TimeID, Count(*) as Num_of_Employees
from Employee E, TimeDim T
```

```
where to_char(E.HireDate, 'YYYYMM') <= T.TimeID
and to_char(E.CeaseDate, 'YYYYMM') >= T.TimeID
group by JobID, DeptID, TimeID;
```

However, if we remove the TimeDim, as in the second star schema above, the calculation of the Number of Employee can be difficult, if the star schema does not maintain the history of employees. Therefore, this higher level of aggregation star schema (the second star schema above) will not be accurate to reflect the status of number of employees, as the time dimension is missing.