Université libre de Bruxelles
Faculté des Sciences Appliquées

January 28, 2022

INFO-H-419 – Data Warehouses

First session examination

Question 1: Data Warehouse Design (8 points)

Consider the data warehouse of a university that contains information about teaching and research activities. On the one hand, the information about teaching activities is related to dimensions department, professor, course, and time, the latter at a granularity of academic semester. Measures for teaching activities are number of hours and number of credits. On the other hand, the information about research activities is related to dimensions professor, funding agency, project, and time, the latter twice for the start date and the end date, both at a granularity of day. In this case, professors are related to the department to which they are affiliated. Measures for research activities are the number of person months and amount.

1. Design a conceptual schema for the data warehouse. Propose dimension attributes and dimension hierarchies.
2. Translate the conceptual schema into a relational schema. Clearly indicate primary and foreign keys in your tables.
3. For the relational schema obtained in the previous question, write in SQL the following queries.
   (a) By department, total number of teaching hours during the academic year 2018–2019.
   (b) By department, total amount of research projects during the calendar year 2018.
   (c) By department, total number of professors involved in research projects during the calendar year 2018.
   (d) By professor, total number of courses delivered during the academic year 2018–2019.
   (e) By department and funding agency, total number of projects started in 2018.

Answer A MultiDim schema for this application is given in Fig. 1. The translation of this schema into a relational schema is given in Fig. 2. The SQL queries are given next.

1. Total number of kilometers made by Alstom trains during 2018 departing from French or Belgian stations.

```
SELECT SUM(Distance)
FROM   Segments F, Time T, Train TR, Model M,
       Constructor C, Station S, City CI, State ST, Country CO
WHERE F.FromTimeKey = T.TimeKey AND T.Year = '2018' AND
       F.TrainKey = TR.TrainKey AND
       TR.ModelKey = M.ModelKey AND
       M.ConstructorKey = C.ConstructorKey AND
       C.ConstructorName = 'Alstom' AND
       F.FromStationKey = S.StationKey AND
       S.CityKey = CI.CityKey AND
       CI.StateKey = ST.StateKey AND
       ST.CountryKey = CO.CountryKey AND
       ( CO.CountryName = 'France' OR
       CO.CountryName = 'Belgium' )
```

2. Total duration of international trips during 2018, that is, trips departing from a station located in a country and arriving at a station located in another country.

```
SELECT SUM(Duration)
FROM   Segments F, Time T, Station A1, City C1, State S1,
```

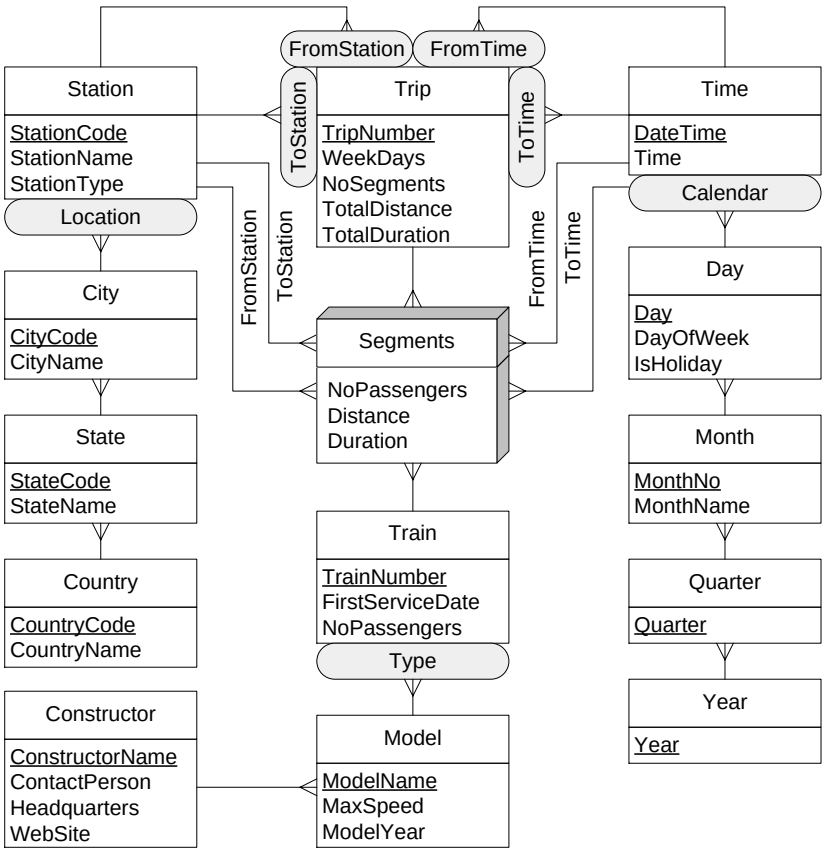

Figure 1: MultiDim schema for the train data warehouse

```
        Station A2, City C2, State S2
WHERE F.FromTimeKey = T.TimeKey AND T.Year = '2018' AND
        F.FromStationKey = A1.StationKey AND
        A1.CityKey = C1.CityKey AND
        C1.StateKey = S1.StateKey AND
        F.ToStationKey = A2.StationKey AND
        A2.CityKey = C2.CityKey AND
        C2.StateKey = S2.StateKey AND
        S1.CountryKey <> S2.CountryKey
```

3. Total number of trains that departed from or arrived at Paris during July 2018.

```
SELECT COUNT(*)
FROM   Segments F, Time T, Trip TR, Station S, City C
WHERE F.FromTimeKey = T.TimeKey AND
        T.Month = 'July' AND T.Year = '2018' AND
        ( F.FromStationKey = S.StationKey OR
        F.ToStationKey = S.StationKey ) AND
        S.CityKey = C.CityKey AND
        C.CityName = 'Paris'
```

4. Average duration of train segments in Belgium in 2018.

```
SELECT SUM(Duration)
FROM   Segments F, Time T, Station A1, City C1, State S1,
        Country CO1, Station A2, City C2, State S2, Country CO2
WHERE F.FromTimeKey = T.TimeKey AND T.Year = '2018' AND
        F.FromStationKey = A1.StationKey AND
        A1.CityKey = C1.CityKey AND
        C1.StateKey = S1.StateKey AND
        S1.CountryKey = CO1.CountryKey AND
```

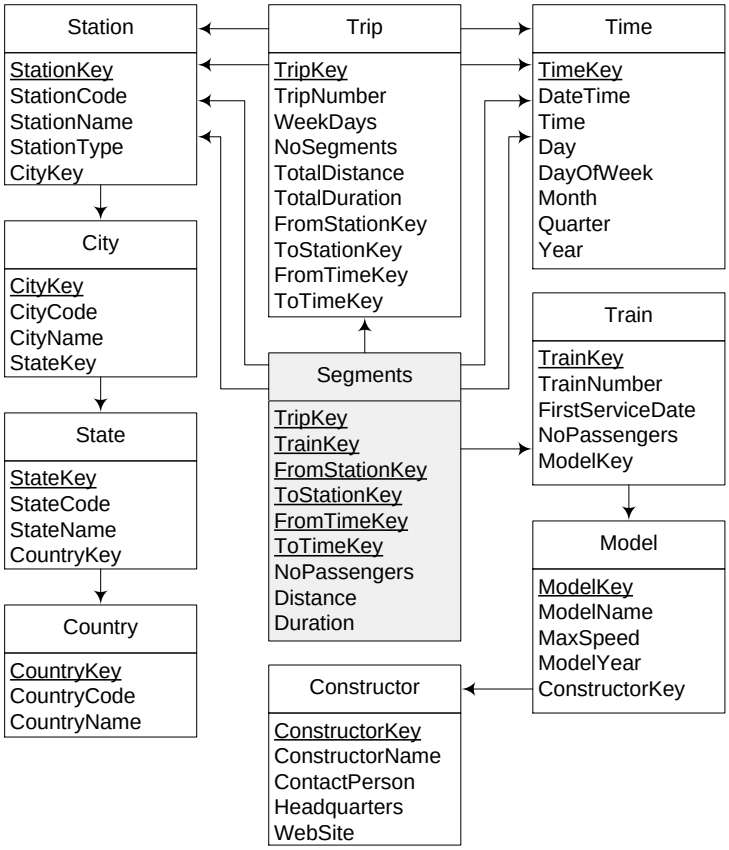

Figure 2: A snowflake schema for the data warehouse

```
CO1.CountryName = 'Belgium' AND
F.ToStationKey = A2.StationKey AND
A2.CityKey = C2.CityKey AND
C2.StateKey = S2.StateKey AND
S2.CountryKey = CO2.CountryKey AND
CO2.CountryName = 'Belgium'
```

5. For each trip, average number of passengers per segment, that means, take all the segments of each trip, and average the number of passengers.

```
SELECT    T.TripNumber, AVG(NoPassengers)
FROM      Segments F, Trip T
WHERE     F.FromTimeKey = T.TripKey
GROUP BY T.TripNumber
```

Question 2: View Materialization (4 points)

Consider the graph in Fig. 3, where each node represents a view and the numbers are the costs of materializing the view. Assume that the bottom of the lattice is materialized, and that the probability of the different views is as follows: B 5%, AC 10%, BC 15%, CD 20%, ABC 25%, ACD 25%, and 0% for the other views. Determine using the View Selection Algorithm the five views to be materialized first.

Answer The results of five iterations of the algorithm are shown in Fig. 4. Notice that in the fourth iteration there is a tie between views BCD and CD. We have chosen arbitrarily to materialize the latter.

Question 3: Indexing (4 points)

Consider the tables Sales, Employee, and Department given below.

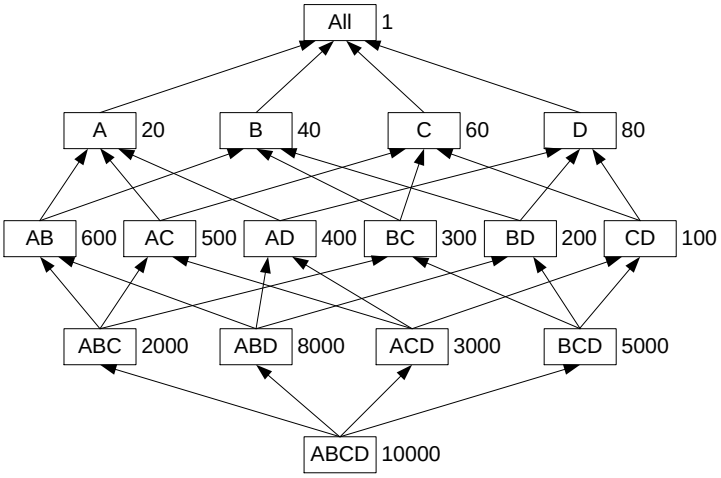

Figure 3: A data cube lattice

View	Iteration				
	1	2	3	4	5
ABC	**64000**				
ABD	14000	8000	4000	2000	2000
ACD	49000	**28000**			
BCD	35000	20000	10000	5000	**5000**
AB	37600	5600	5600	2800	2800
AC	38000	6000	6000	4500	3000
AD	38400	22400	8400	4200	4200
BC	38800	6800	6800	3400	1700
BD	39200	23200	**16200**		
CD	39600	23600	9600	**5000**	
A	19960	3960	3960	2160	2060
B	19920	3920	3920	320	220
C	19880	3880	3880	2080	80
D	19840	19840	5840	240	44
All	9999	1999	1999	199	99

Figure 4: Result of the selection of the five views to be materialized

RowID Sales	Product Key	Customer Key	Employee Key	Date Key	Sales Amount
1	p1	c1	e1	d1	100
2	p1	c2	e3	d1	100
3	p2	c2	e4	d2	100
4	p2	c3	e5	d2	100
5	p3	c3	e1	d3	100
6	p4	c4	e2	d4	100
7	p5	c4	e2	d5	100

Employee Key	Employee Name	Title	Address	City	Department Key
e1	Peter Brown	Dr.	...	Brussels	d1
e2	James Martin	Mr.	...	Wavre	d1
e3	Ronald Ritchie	Mr.	...	Paris	d2
e4	Marco Benetti	Mr.	...	Versailles	d2
e5	Alexis Manoulis	Mr.	...	London	d3
e6	Maria Mortsel	Mrs.	...	Reading	d3
e7	Laura Spinotti	Mrs.	...	Brussels	d4
e8	John River	Mr.	...	Waterloo	d4
e9	Bert Jasper	Mr.	...	Paris	d5
e10	Claudia Brugman	Mrs.	...	Saint-Denis	d5

Department Key	Department Name	Location
d1	Management	Brussels
d2	Production	Paris
d3	Marketing	London
d4	Human Resources	Brussels
d5	Research	Paris

Propose an indexing scheme for the tables, including any kind of index you consider it necessary. Discuss possible alternatives according to several query scenarios. Discuss the advantages and disadvantages of creating the indexes.

Question 4: Aggregate Computation (4 points)

1. Draw the data cube lattice of a three-dimensional cube with dimensions A, B, and C.
2. Extend the lattice to take into account the hierarchies $A \rightarrow A_1 \rightarrow All$ and $B \rightarrow B_1 \rightarrow B_2 \rightarrow All$. Since the lattice is complex to draw, represent it by giving the list of nodes and the list of edges.

Answer