Data warehouse: conceptual design

December 5, 2011

Outline

- Data Warehouse conceptual design
  - facts, dimensions, measures
  - attribute tree
  - fact schema
- **Exercise 1:** insurance company
- **Exercise 2:** international airport
- **Exercise 3:** wholesale furniture company

Introduction

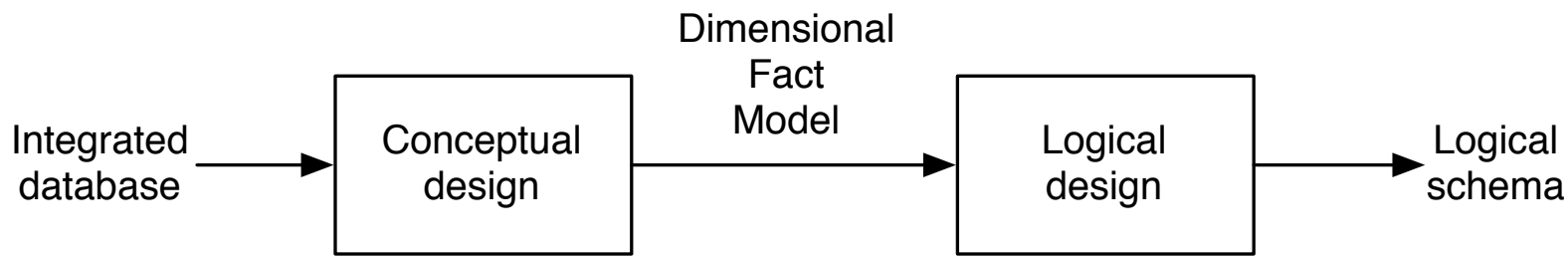
What is a Data Warehouse?

- It is a (usually huge) **collection of data**
- It is used primarily in **decision making processes**
- It is **integrated**: data comes from different sources
- It is **subject oriented**: it is used to study the dynamics of a specific topic
- It is **time varying**: it stores past and present data and the goal is to learn some information that could help in the future

Data Warehouse design process

Design steps

- Design process starts with the **integrated database**, usually represented by:
  - ER schema or
  - logical schema or
  - requirements



- The first step is **conceptual design**:
  - data is represented according to the **data cube model/fact model**
- The second step is **logical design**:
  - data is represented according to the **relational model**

Data cube model

Definitions

Fact

A **concept** that is relevant for the decisional process (e.g. sales)

- A fact is always represented by **frequently updated data**, not static archives!

Measure

A **numerical property** of a fact (e.g. sold quantity, total income)

Dimension

A **property** of a fact described with respect to a finite domain (e.g. product, time, zone)

- **Time** should always be a dimension!
- Dimensions can have **hierarchies** (e.g. Time: Day $\rightarrow$ Month $\rightarrow$ Year, Zone: City $\rightarrow$ Region $\rightarrow$ State)

Conceptual design

How to do it?

- It is the first step towards the design of a Data Warehouse
- It starts from the documentation related to the integrated database and consists of:
  1. Facts definition
  2. For each fact:
     - attribute tree definition
     - attribute tree editing
     - dimensions definition
     - measures definition
     - hierarchies definition
     - fact schemata creation
     - glossary definition

Exercise 1

Insurance company

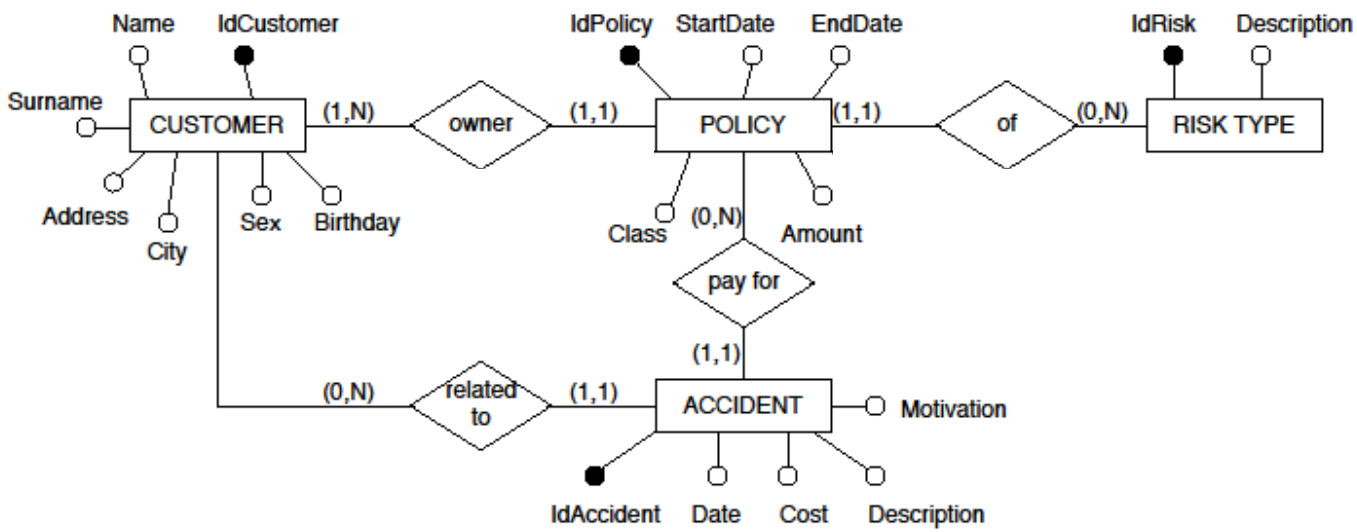
An insurance company requires the data warehouse design for accident analysis of its customers. In particular, the company requires to evaluate the type of accidents related to customers and type of policies.

- Goal:
  - Evaluate the history of accidents w.r.t. the policies and the customers of the insurance company
  - Evaluate the history of policies w.r.t. the customers of the insurance company by considering the risk type and the policy amount
- Questions: Design the Data Warehouse for the two problems (accident and risk analysis)
  - Choose facts, measures and dimensions
  - Define the attribute tree (and describe the editing phase)
  - Define the fact schemata for the two considered facts

Exercise 1

Insurance company

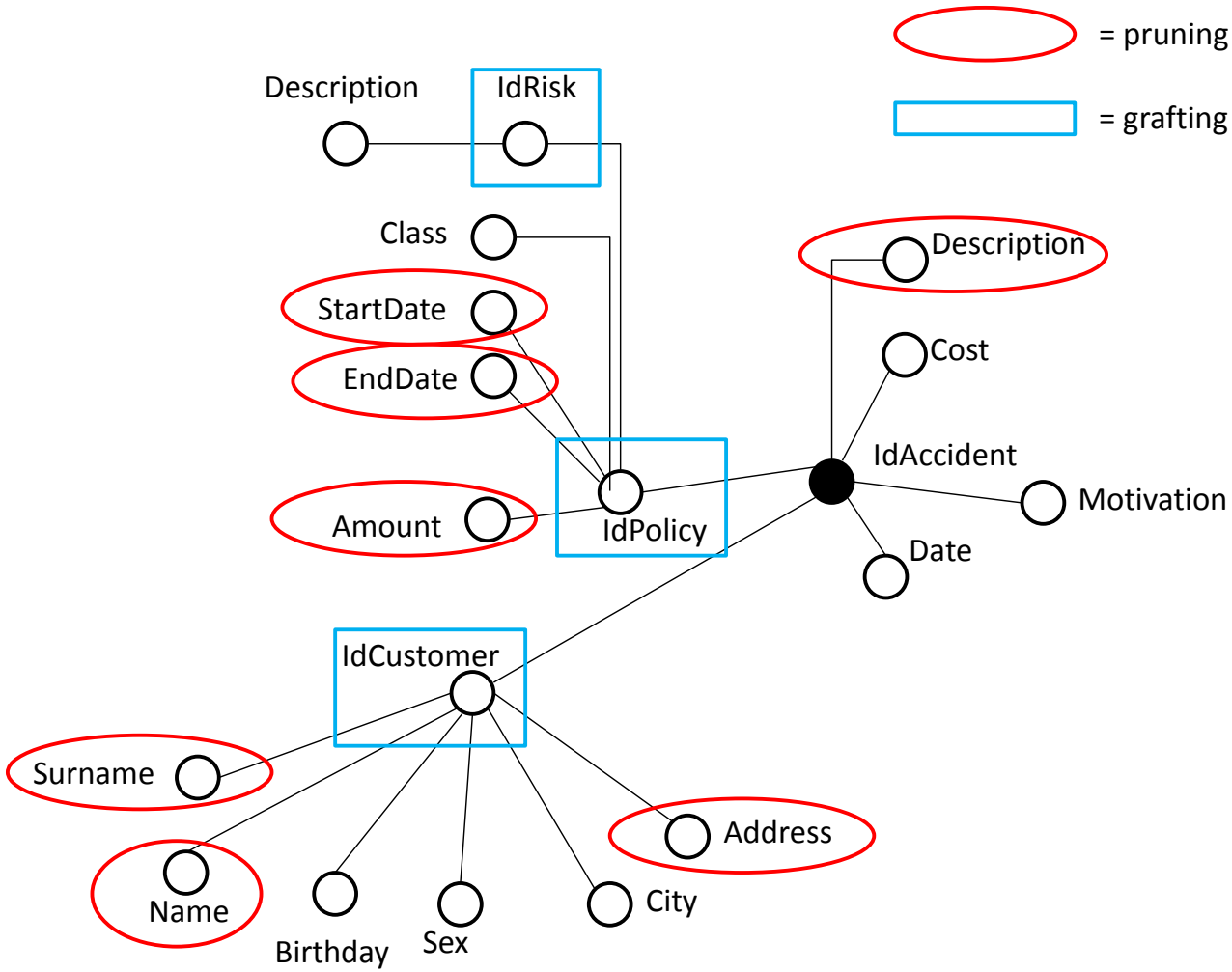
The ER schema related to the insurance company operational DB, which contains the information that has to be considered to design the required Data Warehouse, is:



Exercise 1: A possible solution

Attribute tree definition

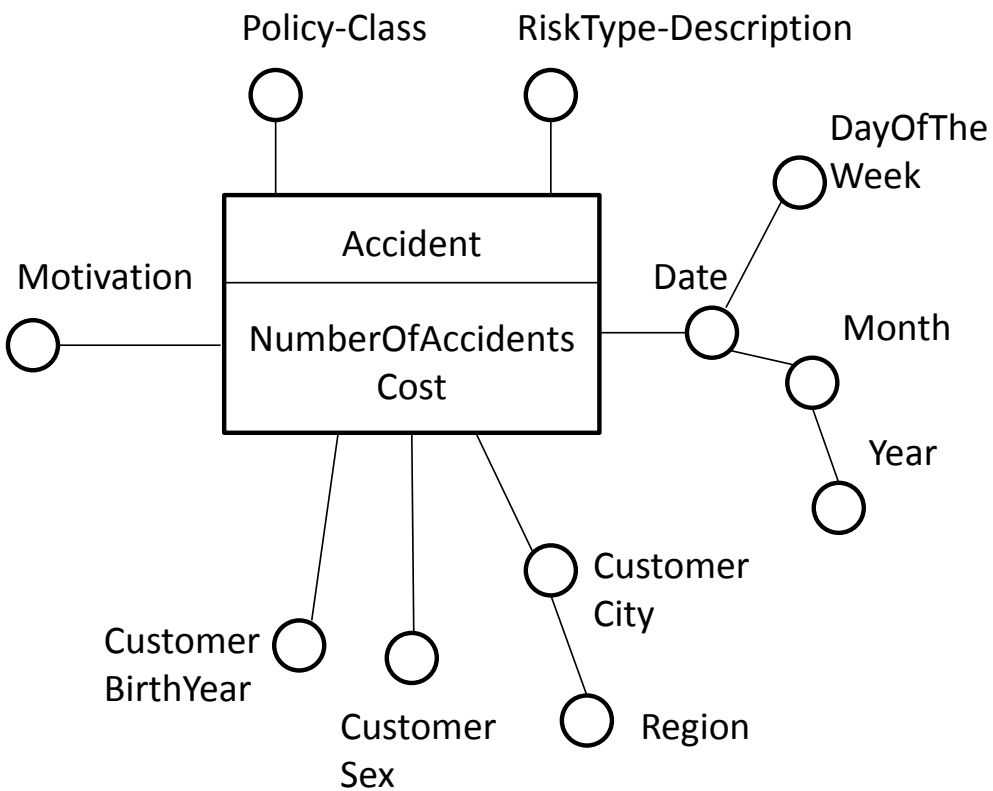
- *Fact:* **Accident**



Exercise 1: A possible solution

Fact model definition

- *Fact:* **Accident**



Exercise 1: A possible solution

Glossary definition

- **NumberOfAccidents**

  SELECT COUNT(*)
  FROM ACCIDENT A, POLICY P, RISK_TYPE R,
  CUSTOMER C
  WHERE - join conditions -
  GROUP BY A.Motivation, P.Class, R.Description,
  A.Date, C.City, C.Sex, Year(C.Birthday)

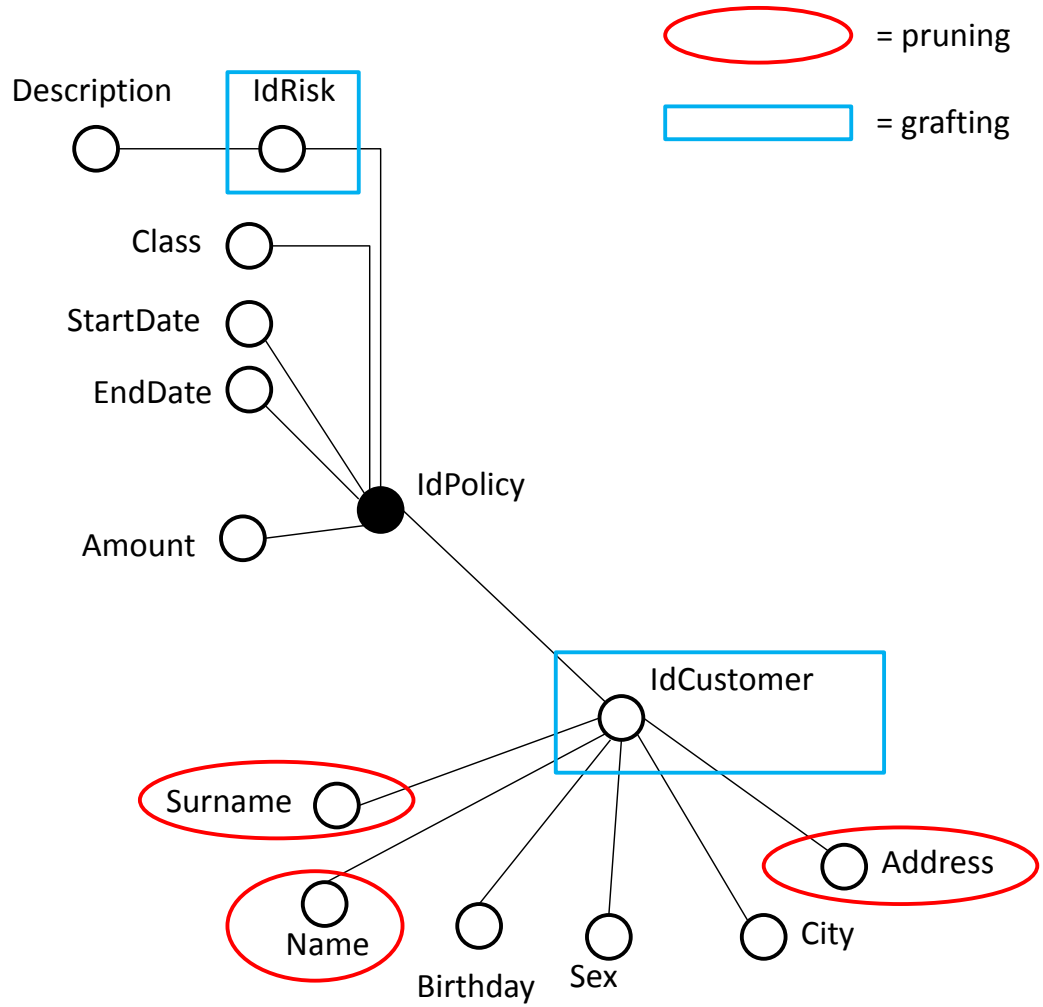
- **Cost**

  SELECT SUM(Cost)
  FROM ACCIDENT A, POLICY P, RISK_TYPE R,
  CUSTOMER C
  WHERE - join conditions -
  GROUP BY A.Motivation, P.Class, R.Description,
  A.Date, C.City, C.Sex, Year(C.Birthday)

Exercise 1: A possible solution

Attribute tree definition

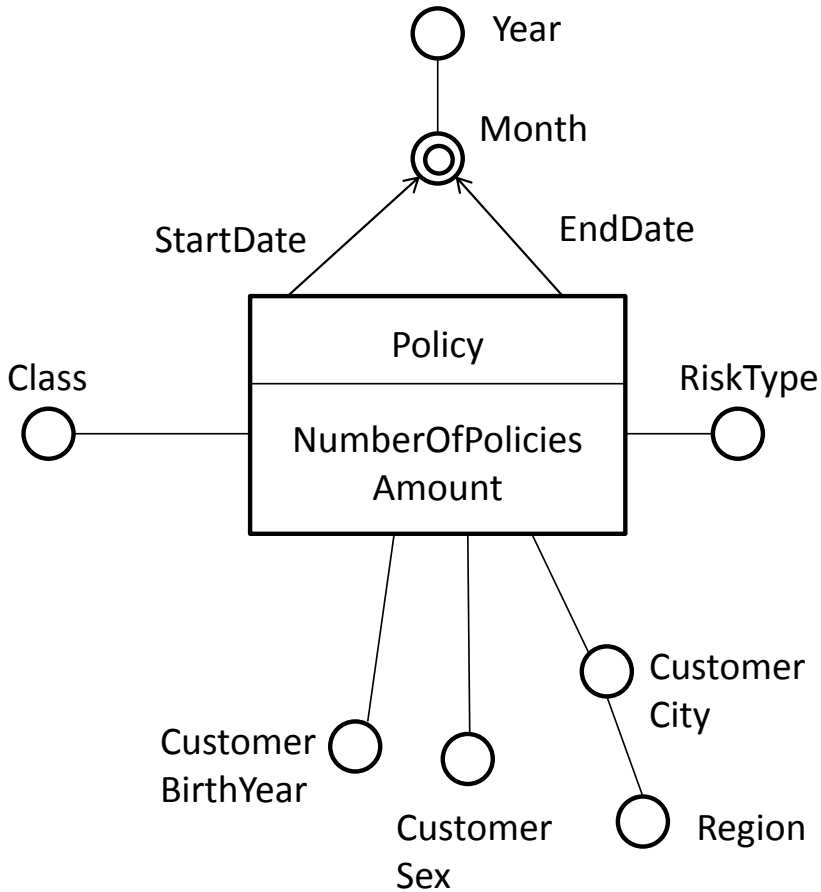
- *Fact:* **Policy**



Exercise 1: A possible solution

Fact model definition

- *Fact:* **Policy**



Exercise 1: A possible solution

Glossary definition

- **NumberOfPolicies**

  SELECT COUNT(*)
  FROM POLICY P, RISK_TYPE R, CUSTOMER C
  WHERE - join conditions -
  GROUP BY P.Class, P.StartDate, P.EndDate,
  R.Description, C.City, C.Sex, Year(C.Birthday)

- **Amount**

  SELECT SUM(Amount)
  FROM POLICY P, RISK_TYPE R, CUSTOMER C
  WHERE - join conditions -
  GROUP BY P.Class, P.StartDate, P.EndDate,
  R.Description, C.City, C.Sex, Year(C.Birthday)

Exercise 2

International airport

Consider the following relational database schema of an international airport.

FLIGHT	(IDF, Company, DepAirport, ArrAirport, DepTime, ArrTime)
FLYING	(IDFlight, FlightDate)
AIRPORT	(IDAirport, AirName, City, State)
TICKET	(Number, IDFlight, FlightDate, Seat, Rate, Name, Surname, Sex)
CHECK-IN	(Number, CheckInTime, LuggageNr)

Design the Data Warehouse for the analysis of tickets:

- Choose facts, measures and dimensions
- Define the attribute tree and the fact schema

Exercise 2: A possible solution

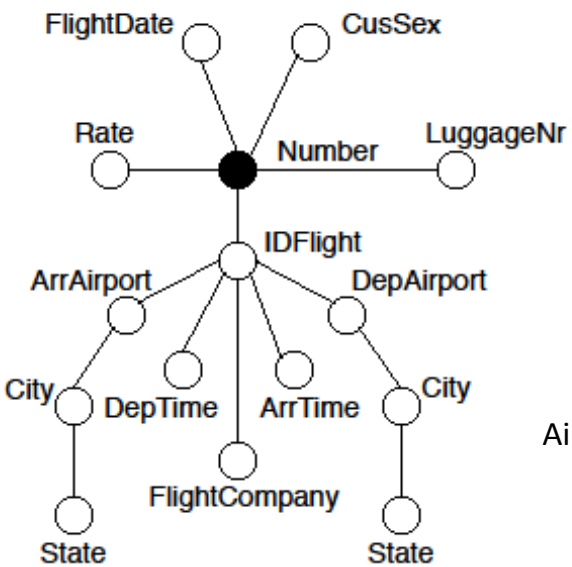
Fact definition, Attribute tree definition, Fact schemata creation

- *Facts:* Ticket analysis
- *Measures:* NumberOfTickets, NumberOfLuggage, TotalIncome
- *Dimensions:* Ticket characteristics (CusSex, FlightDate), Flight (FlightCompany, DepAirport, ArrAirport, DepTime, ArrTime)

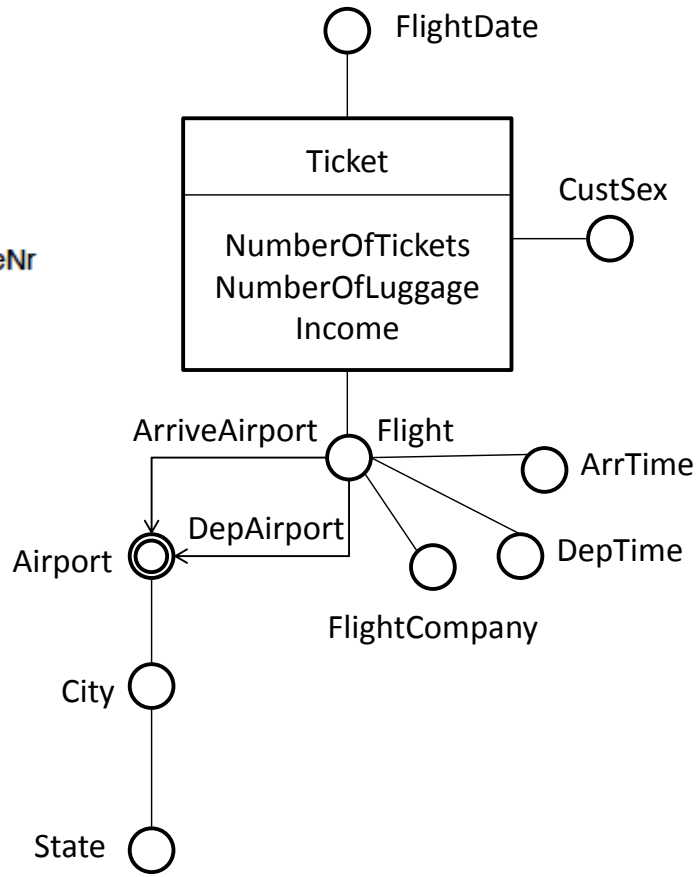
Exercise 2: A possible solution

Fact definition, Attribute tree definition, Fact schemata creation

- *Fact:* **Ticket**



Attribute tree



Exercise 2: A possible solution

Glossary definition

- **NumberOfTickets**

```
SELECT COUNT(*)
FROM TICKET
GROUP BY CustSex, IDFlight, FlightDate
```

- **NumberOfLuggage**

```
SELECT SUM(c.LuggageNr)
FROM TICKET t, CHECK-IN c
WHERE t.Number = c.Number
GROUP BY t.CustSex, t.IDFlight, t.FlightDate
```

- **TotalIncome**

```
SELECT SUM(Rate)
FROM TICKET
GROUP BY CustSex, IDFlight, FlightDate
```

Exercise 3

Wholesale furniture company

Design the data warehouse for a wholesale furniture company. The data warehouse has to allow to analyze the company's situation at least with respect to Furnitures, Customers and Time. Moreover, the company needs to analyze:

- the furniture with respect to its type (chair, table, wardrobe, cabinet...), category (kitchen, living room, bedroom, bathroom, office...) and material (wood, marble...)
- the customers with respect to their spatial location, by considering at least cities, regions and states

The company is interested in learning at least the quantity, income and discount of its sales:

- Choose facts, measures and dimensions
- Define the attribute tree and the fact schema

Exercise 3

Schema of the operational database

SALES (IDSale, Date, IDFurniture, IDCustomer, Quantity, Cost, Discount)

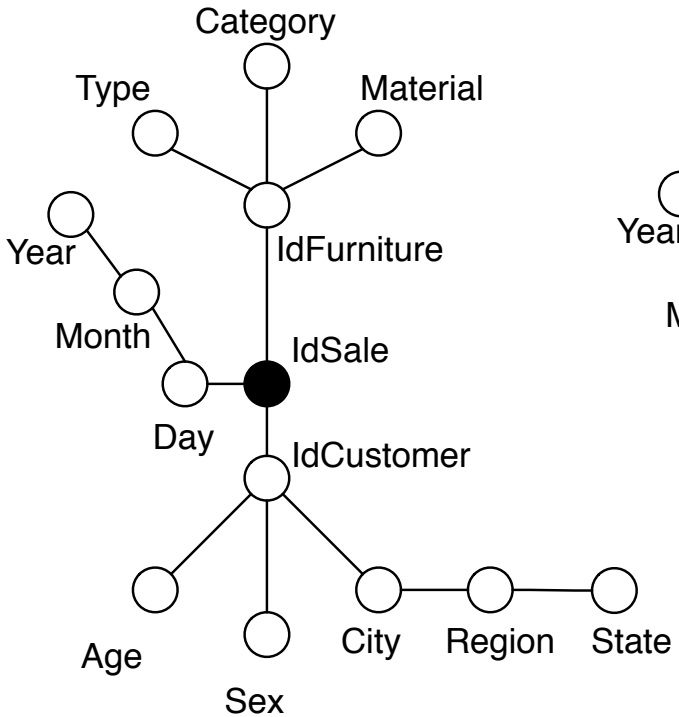
FURNITURE (IDFurniture, FurnitureType, FurnitureName, Category)

CUSTOMER (IDCustomer, Name, Surname, Birthdate, Sex, City)

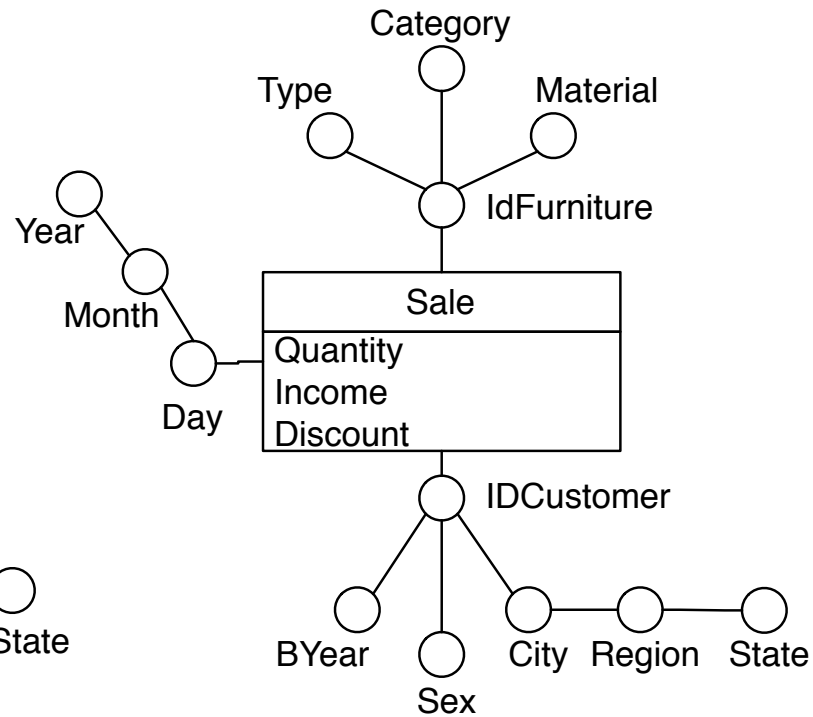
Exercise 3: A possible solution

Facts, dimensions, measures, attribute tree, fact schema

- *Fact:* **Sales**



Attribute tree



Fact schema

- *Measures:* Quantity, Income, Discount
- *Dimensions:* Furniture (Type, Category, Material)
  Customer (Age, Sex, City $\rightarrow$ Region $\rightarrow$ State)
  Time (Day $\rightarrow$ Month $\rightarrow$ Year)