Data warehousing logical design

Mirjana Mazuran
mazuran@elet.polimi.it

November 27, 2008

Outline

- Data Warehouse logical design
  - ROLAP model
    - star schema
    - snowflake schema
- **Exercise 1:** wine company
- **Exercise 2:** real estate agency
- **Exam 9/1/07:** travel agency

Introduction

Logical design

- Ex starting from the conceptual design it is necessary to determin the **logical schema** of data
- Ex we use **ROLAP** (Relational On-Line Analytical Processing) model to represent multidimensional data
- Ex ROLAP uses the relational data model, which means that **data is stored in relations**
- Ex given the DFM representation of multidimensional data, **two schemas** are used:
  - RO **star** schema
  - RO **snowflake** schema

ROLAP

Star schema

- Each dimension is represented by a relation such that:
  - the primary key of the relation is the primary key of the dimension
  - the attributes of the relation describe all aggregation levels of the dimension
- A fact is represented by a relation such that:
  - the primary key of the relation is the set of primary keys imported from all the dimension tables
  - the attributes of the relation are the measures of the fact

Pros and Cons

- few joins are needed during query execution
- dimension tables are denormalized
- denormalization introduces redundancy

ROLAP

Snowflake schema

- Each (primary) dimension is represented by a relation:
  - the primary key of the relation is the primary key of the dimension
  - the attributes of the relation directly depend by the primary key
  - a set of foreign keys is used to access information at different levels of aggregation. Such information is part of the secondary dimensions and is stored in dedicated relations
- A fact is represented by a relation such that:
  - the primary key of the relation is the set of primary keys imported from all and only the primary dimension tables
  - the attributes of the relation are the measures of the fact

Pros and Cons

- denormalization is reduced
- less memory space is required
- a lot of joins can be required if they involve attributes in secondary dimension tables

Exercise 1

Wine company

An online order wine company requires the designing of a **snowflake schema** to record the quantity and sales of its wines to its customers. Part of the original database is composed by the following tables:

CUSTOMER (Code, Name, Address, Phone)

WINE (Code, Name, Vintage, BottlePrice, CasePrice)

CLASS (Code, Name, Region)

AREA (Code, Description)

TIME (TimeStamp, Date, Year)

Note that the tables represent the main entities of the ER schema, thus it is necessary to derive the significant relashionships among them in order to correctly design the data warehouse.

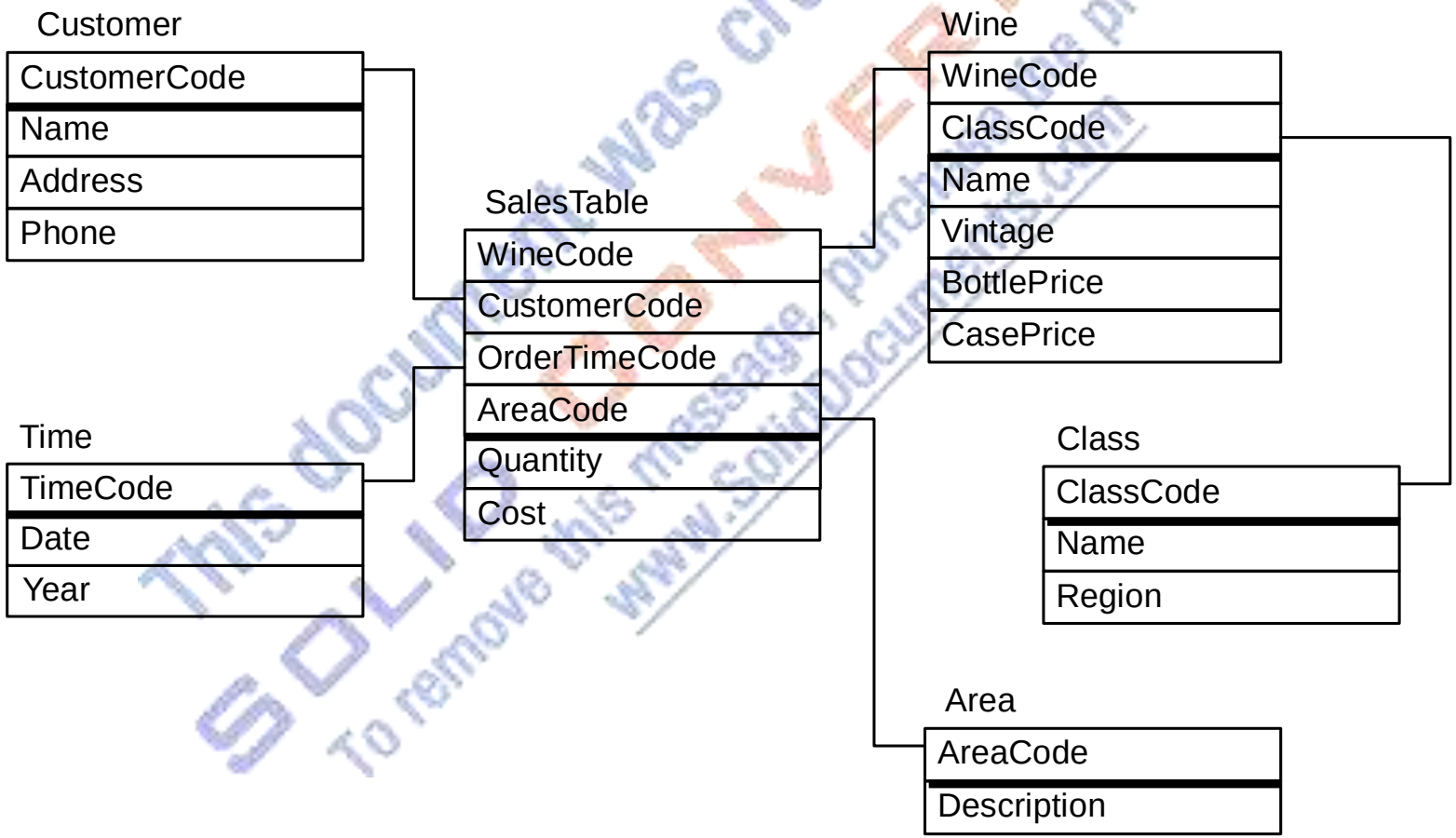
Exercise 1: A possible solution

Snowflake schema

FACT Sales

MEASURES Quantity, Cost

DIMENSIONS Customer, Area, Time, Wine → Class



Exercise 2

Real estate agency

Let us consider the case of a real estate agency whose database is composed by the following tables:

OWNER (IDOwner, Name, Surname, Address, City, Phone)

ESTATE (IDEstate, IDOwner, Category, Area, City, Province, Rooms, Bedrooms, Garage, Meters)

CUSTOMER (IDCust, Name, Surname, Budget, Address, City, Phone)

AGENT (IDAgent, Name, Surname, Office, Address, City, Phone)

AGENDA (IDAgent, Data, Hour, IDEstate, ClientName)

VISIT (IDEstate,IDAgent, IDCust, Date, Duration)

SALE (IDEstate,IDAgent, IDCust, Date, AgreedPrice, Status)

RENT (IDEstate,IDAgent, IDCust, Date, Price, Status, Time)

Exercise 2

Real estate agency

- Ex Goal:
  - RO Provide a supervisor with an overview of the situation. The supervisor must have a global view of the business, in terms of the estates the agency deals with and of the agents' work.
- Ex Questions:
  1. Design a conceptual schema for the DW.
  2. What facts and dimensions do you consider?
  3. Design a Star Schema or Snowflake Schema for the DW.
- Ex Write the following SQL queries:
  - RO How many customers have visited properties of at least 3 different categories?
  - RO What is the average duration of visits per property category?
  - RO Who has paid the highest price among the customers that have viewed properties of at least 3 different categories?
  - RO Who has bought a flat for the highest price w.r.t. each month?
  - RO What kind of property sold for the highest price w.r.t each city and month?

Exercise 2: A possible solution

Facts and dimensions

Points 1 and 2 are left as homework. In particular it is required to discuss the facts of interest (with respect to your point of view) and then define, for each fact, the attribute tree, dimensions and measures with the corresponding glossary.

The following ideas will be used during the solution of the exercise:

- Ex supervisors should be able to control the sales of the agency

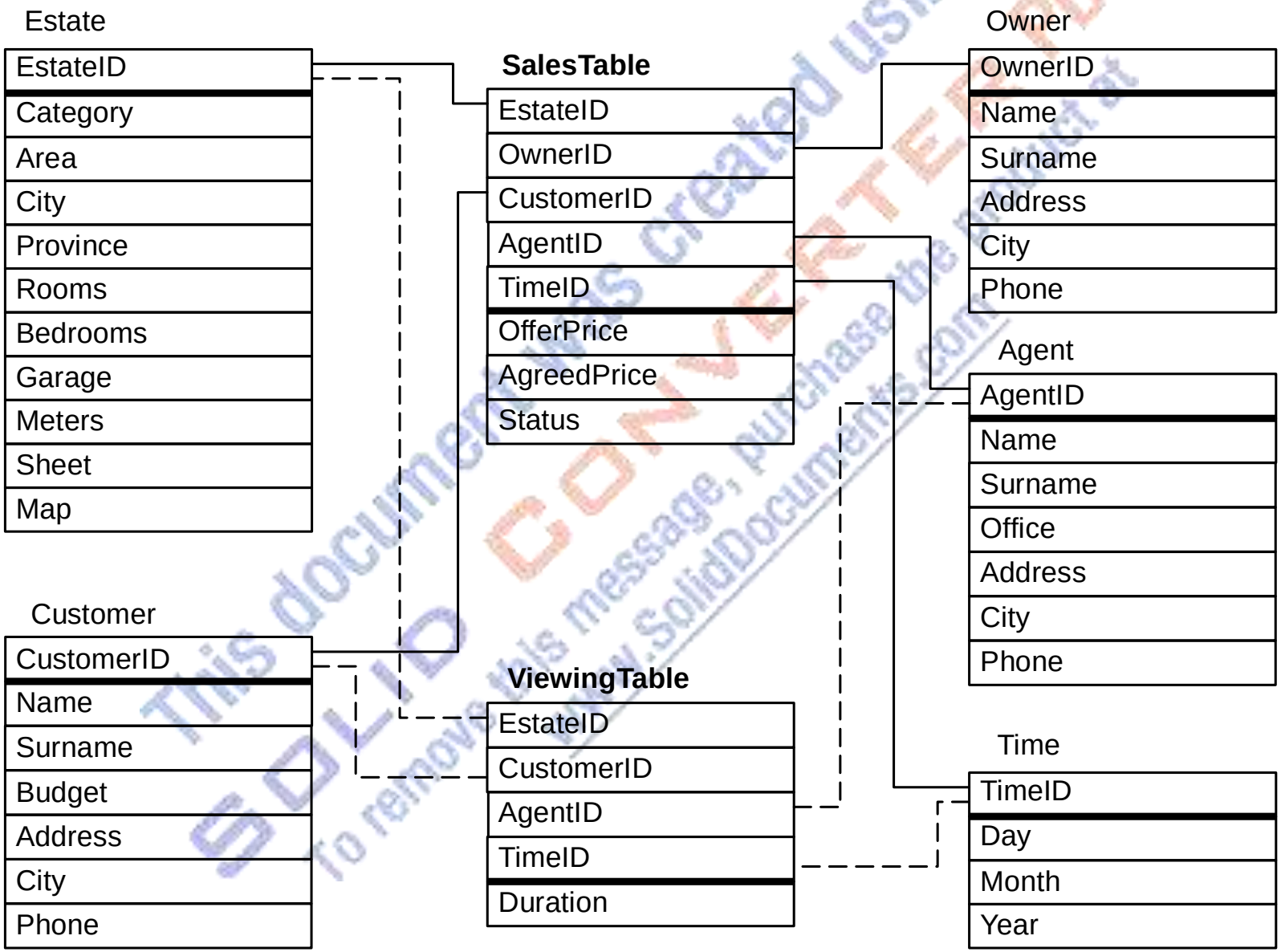
  FACT Sales
  MEASURES OfferPrice, AgreedPrice, Status
  DIMENSIONS EstateID, OwnerID, CustomerID, AgentID, TimeID

- Ex supervisors should be able to control the work of the agents by analyzing the visits to the estates, which the agents are in charge of

  FACT Viewing
  MEASURES Duration
  DIMENSIONS EstateID, CustomerID, AgentID, TimeID

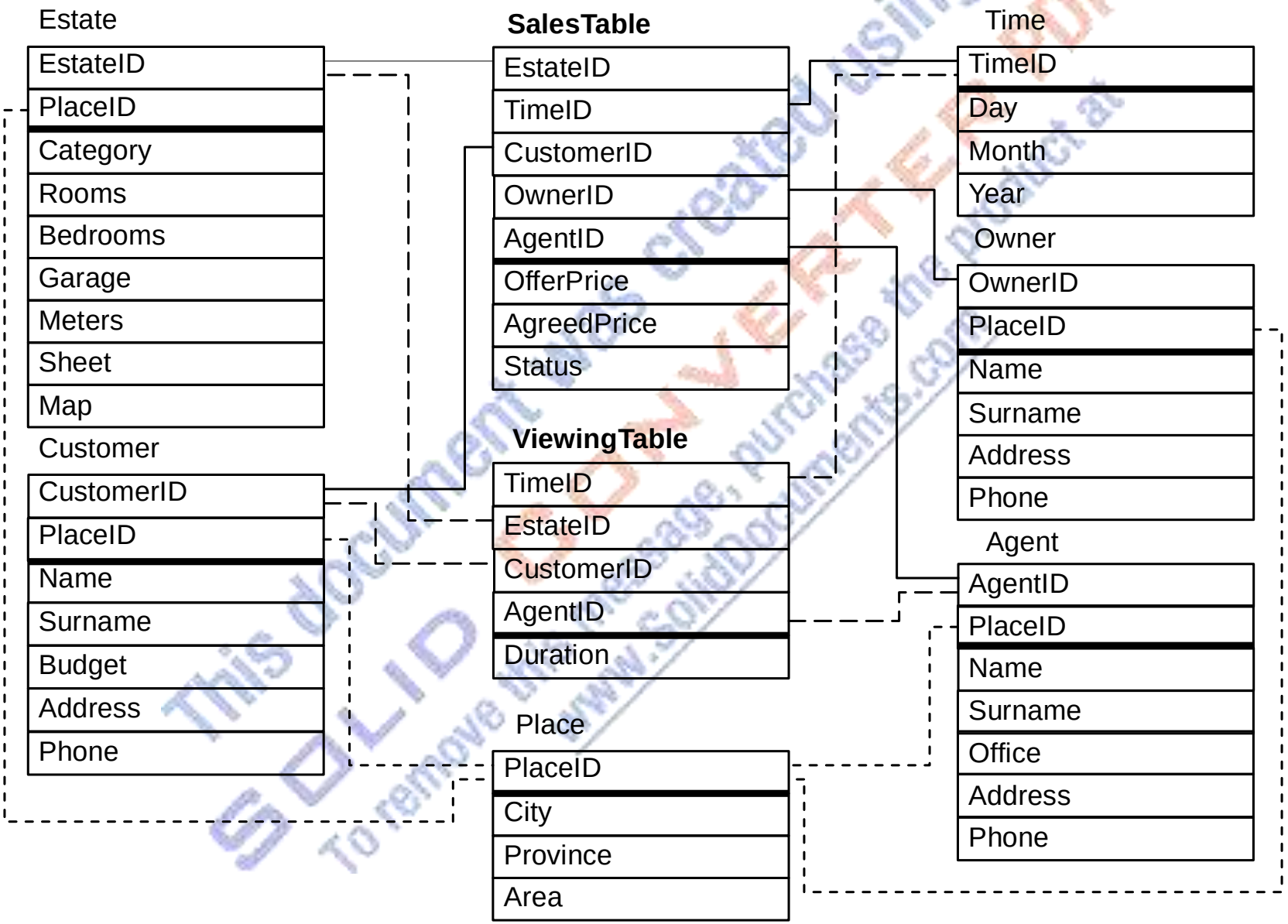
Exercise 2: A possible solution

Star schema



Exercise 2: A possible solution

Snowflake schema



Exercise 2: A possible solution

SQL queries wrt the star schema

Ex How many customers have visited properties of at least 3 different categories?

```
SELECT COUNT(*)
FROM ViewingTable V, Estate E
WHERE V.EstateID = E.EstateID
GROUP BY V.CustomerID
HAVING COUNT(DISTINCT E.Category) >= 3
```

Ex Average duration of visits per property category

```
SELECT E.Category, AVG(V.Duration)
FROM ViewingTable V, Estate E
WHERE V.EstateID = E.EstateID
GROUP BY E.Category
```

Exercise 2: A possible solution

SQL queries wrt the star schema

Ex Who has paid the highest price among the customers that have viewed properties of at least 3 different categories?

```
CREATE VIEW Cust3Cat AS
    SELECT V.CustomerID
    FROM ViewingTable V, Estate E
    WHERE V.EstateID = E.EstateID
    GROUP BY V.CustomerID
    HAVING COUNT(DISTINCT E.Category) >= 3


SELECT C.CustomerID
FROM Cust3Cat C, SalesTable S
WHERE C.CustID = S.CustID AND S.AgreedPrice IN
(SELECT MAX(S.AgreedPrice)
        FROM Cust3Cat C1, SalesTable S1
        WHERE C1.CustomerID = S1.CustomerID)
```

Exercise 2: A possible solution

SQL queries wrt the star schema

Ex Who has bought a flat for the highest price w.r.t. each month?

```
SELECT S.CustomerID, T.Month, T.Year, S.AgreedPrice
FROM SalesTable S, Estate E, Time T
WHERE S.EstateID = E.EstateID AND S.TimeID =
T.TimeID AND E.Category = "flat" AND (T.Month,
T.Year, S.AgreedPrice) IN (
    SELECT T1.Month, T1.Year,
    MAX(S1.AgreedPrice)
    FROM SalesTable S1, Estate E1, Time T1
    WHERE S1.EstateID = E1.EstateID AND
    S1.TimeID = T1.TimeID AND E1.Category =
    "flat"
    GROUP BY T1.Month, T1.Year
```

Exercise 2: A possible solution

SQL queries wrt the star schema

Ex What kind of property sold for the highest price w.r.t each city and month?

```
SELECT E.Category, E.City, T.Moth, T.Year,
                E.AgreedPrice

FROM SalesTable S, Time T, Estate E
WHERE S.TimeID = T.TimeID AND E.EstateID =
S.EstateID AND (P.AgreedPrice, P.City, T.month,
T.year) IN (
                SELECT MAX(E1.AgreedPrice), E1.City,
                T1.Month, T1.Year)
                FROM SalesTable S1, Time T1, Estate E1
                WHERE S1.TimeID = T1.TimeID AND
                E1.EstateID = S1.EstateID
                GROUP BY T.Month, T.Year, E.City)
```

Exam 9/1/07

Travel agency

A travel agency organizes guided trips for tourists of different nationalities. The agency wants to know the main trends about trips both with respect to the visited places and type of partecipant. The following relational schema contains the initial database:

TRIP (Code, Destination, Category, Guide, Duration, Date)

DESTINATIONS (Code, Name, Description, Type, Nation)

GROUP (Trip, Participant, Price)

PARTICIPANT (Code, Name, Surname, Address, Birthday, Nation)

NATION (Code, Name, Continent)

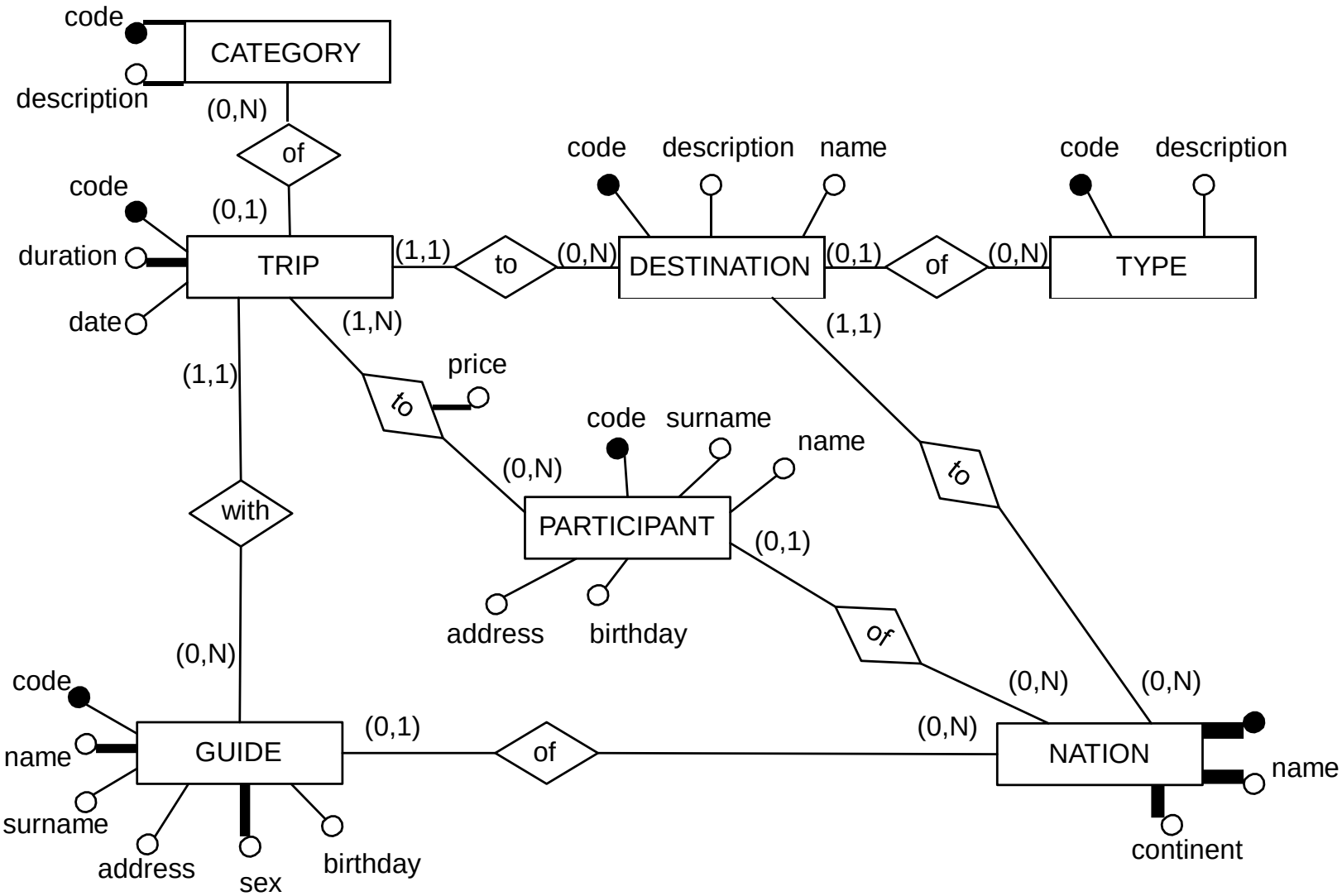
GUIDE (Code, Name, Surname, Address, Sex, Birthday, Nation)

TYPE (Code, Description)

CATEGORY (Code, Description)

Exam 9/1/07: A possible solution

Reverse engineering



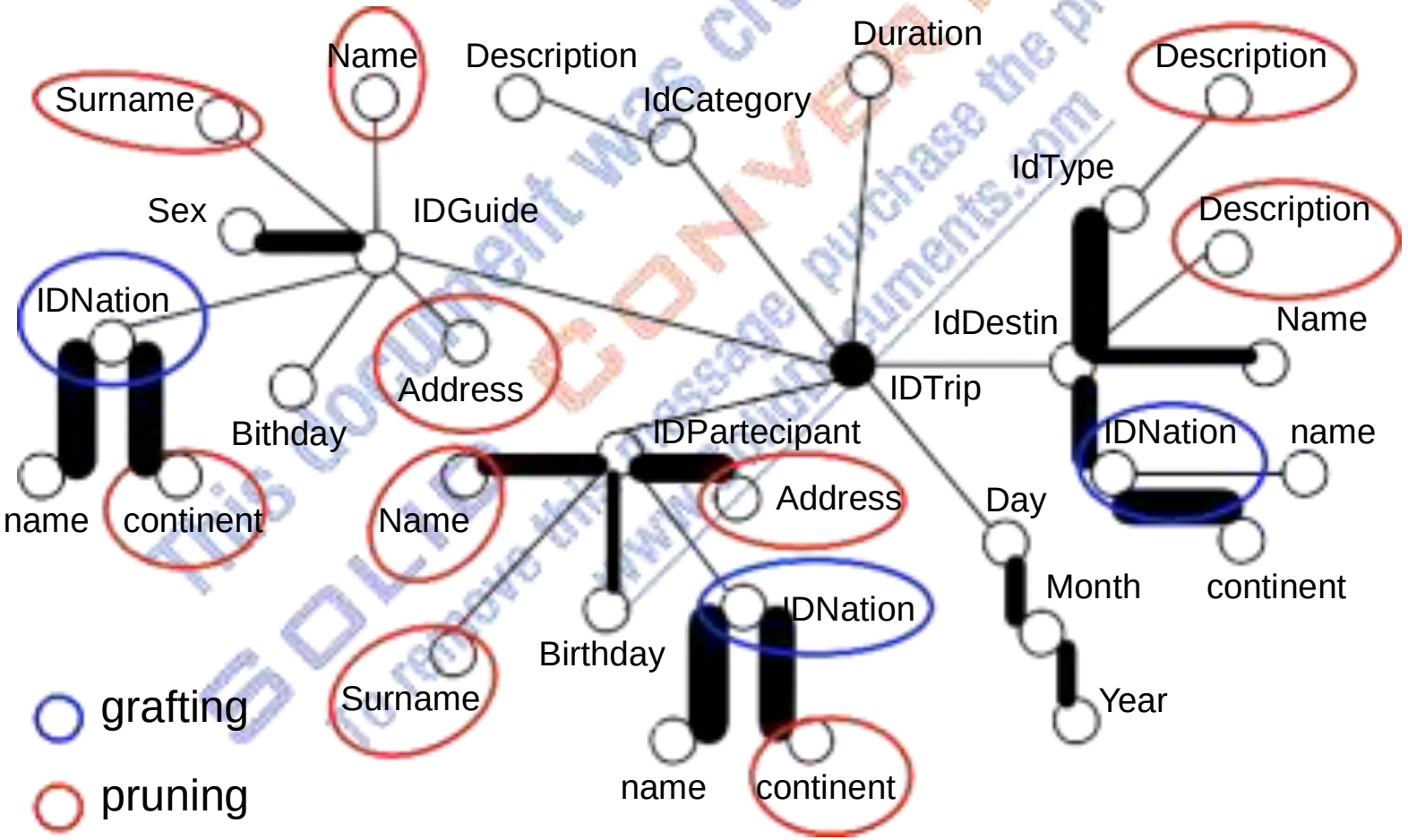
Exam 9/1/07: A possible solution

Facts, measures, dimensions, attribute tree

FACT Trip

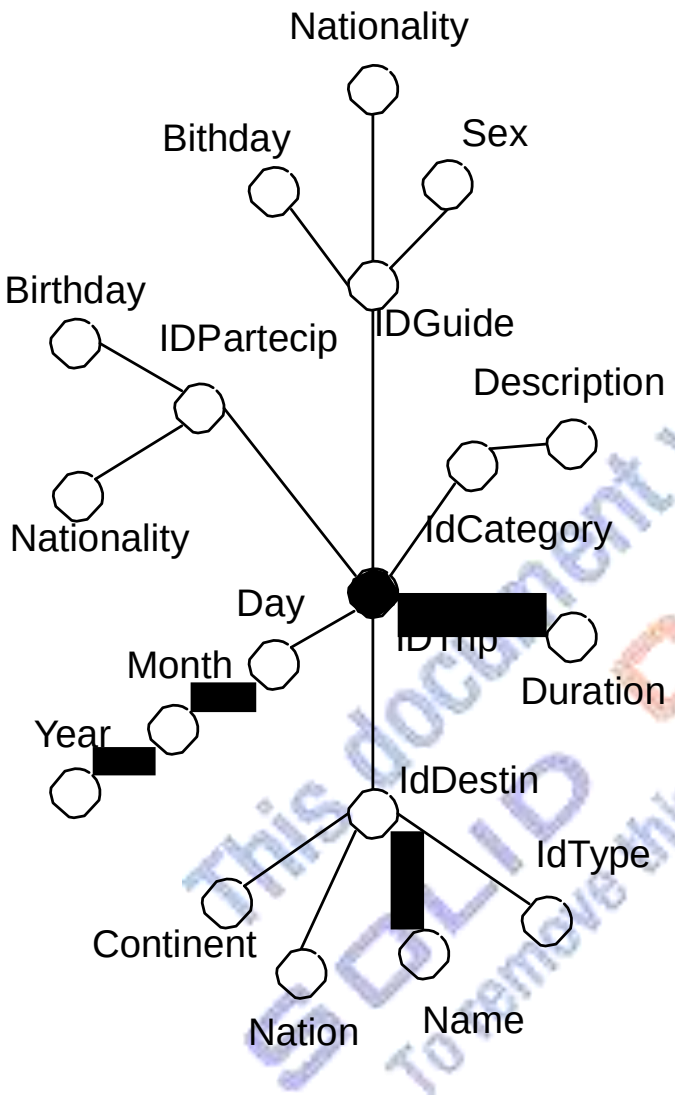
MEASURES PartecipantNr, Duration, Income

DIMENSIONS Partecipant, Place, Guide, Time, Category

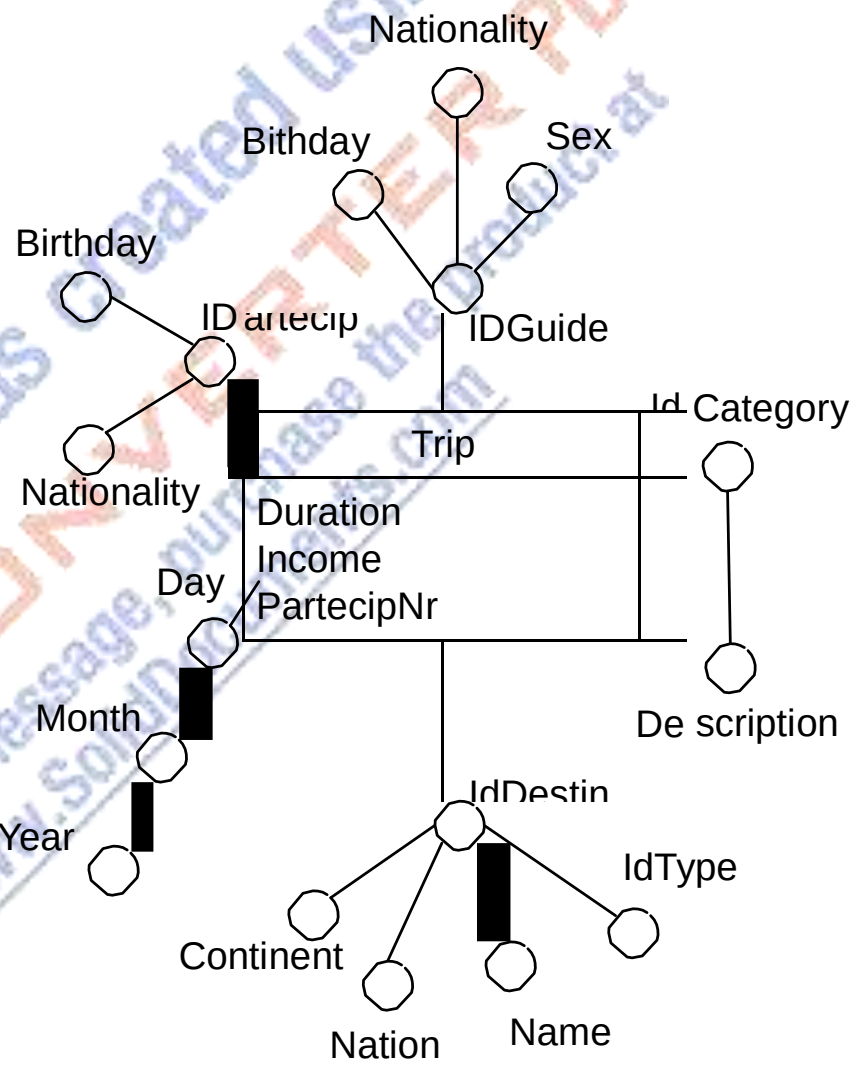


Exam 9/1/07: A possible solution

Attribute tree, fact schema



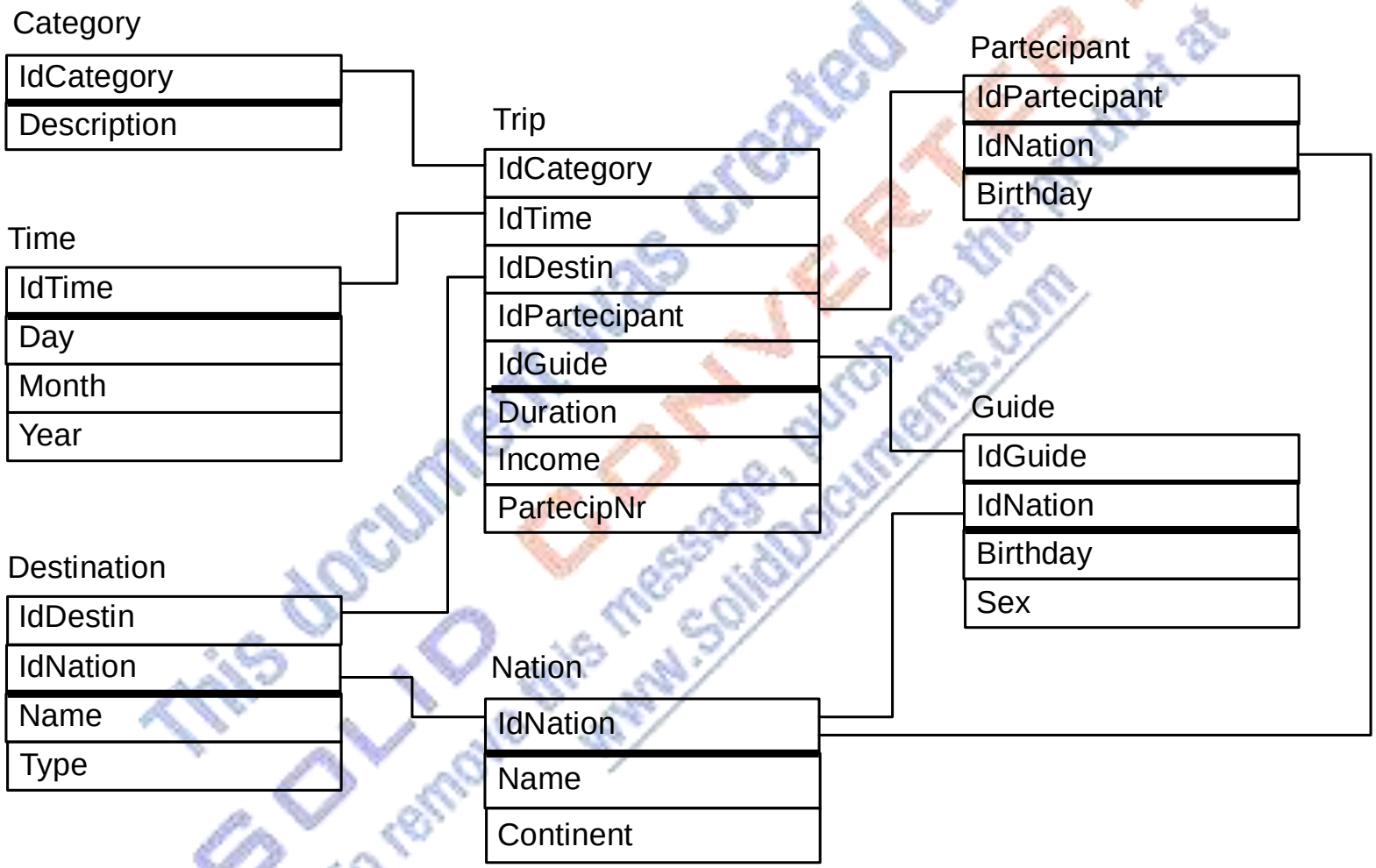
Attribute tree



Fact schema

Exam 9/1/07: A possible solution

Snowflake schema



Exam 9/1/07: A possible solution

SQL queries

- Ex Average trip duration for a given place

```
SELECT AVG(T.Duration)
FROM Trip T, Destination D
WHERE T.IdDestin = D.IdDstin AND D.Name = "Place"
```

- Ex Average trip duration for a given trip category and month

```
SELECT AVG(T.Duration)
FROM Trip T, Time Ti
WHERE T.IdTime = Ti.IdTime AND T.IdCategory =
"Category" AND Ti.Month = "Month"
```

- Ex Average trip price w.r.t. duration, type of place and year

```
SELECT T.Duration, D.Type, Ti.Year, AVG(T.Income)
FROM Trip T, Destination D, Time Ti
WHERE T.IdTim = Ti.IdTim AND T.IdDesti = D.IdDesti
GROUP BY T.Duration, D.Type, Ti.Year
```

Exam 9/1/07: A possible solution

SQL queries

Ex Number of trips w.r.t. type of place, month and guide's nationality

```
SELECT D.Type, Ti.Month, G.IdNation, COUNT(*)
FROM Trip T, Destination D, Time Ti, Guide G
WHERE T.IdDestin = D.IdDestin AND T.IdTime =
Ti.IdTime AND T.IdGuide = G.IdGuide
GROUP BY D.Type, Ti.Month, G.IdNation
```

Ex Average number of partecipants w.r.t. trip category and continent

```
SELECT T.IdCategory, N.Continent, AVG(T.PartecipNr)
FROM Trip T, Destination D, Nation N
WHERE T.IdDestin = D.IdDestin AND D.IdNation =
N.IdNation
GROUP BY T.IdCategory, N.Continent
```