Lecture Note on Dimensional Modeling

Data Warehousing and Foundations of Business Intelligence (California State University, Fullerton)

Lecture Note on Dimensional Modeling for Data Warehouses

ISDS 556
Prof. Turel

At this point, you probably know that operational databases, while optimized for data insertions, can be inadequate for analytical purposes. To mention a few deficiencies, the table structure of OLTP systems, as represented by the Entity-Relationship-Diagram is complex, not easy to comprehend and makes it difficult to run queries against. Here are some of the problems with OLTP data:

(1) We typically take a nice big table with data repeats (e.g., a table the records employees' assignments to projects; because an employee can be assigned to many projects, his or her information, such as name and extension, is repeated), and break it down into several interconnected tables (Employee, Project, Employee_To_Project_Assignment). We do this to adhere to the second normal form (i.e., avoid partial dependencies on parts of the composite key; you have a review of normal forms in the enclosed PPT slides). These multiple interconnected tables are difficult (some may even say impossible) to query.

(2) Furthermore, we do not save calculated data in the operational systems such that we adhere to third-normal form rules (reminder: no functional dependencies; see enclosed PPT). When we run queries, however, we typically look for summarized data (e.g., quarterly sales, by store by line of products). It can take a lot of time, and CPU power, as well as sophisticated "joins", to retrieve transaction-level data from the operational systems, and summarize it on the fly. It may be more efficient, for analytical purposes, to store pre-computed summary (aggregate) data.

(3) Also, historical data are often not saved in the operational systems (written-over every 3-12 months). Thus, operational data does not allow longitudinal analyses (e.g., sales trend).

Analytical systems, on the other hand, are (should be) optimized for data retrieval and analysis. We would like such data retrievals to be simple, and fast. This is why the star-schema concept was created --- it allows understanding the data structure, makes queering simple, and allows fast access to aggregated data.

The key components of a star schema are outlined in the textbook (in chapters 6-7) as well as in the enclosed PowerPoint slides. These include the fact table, which contains (1) the measures we are interested in (e.g., $ sales, units sold), and (2) foreign keys for connecting it to dimension tables (the foreign keys combined create a composite key); and dimension tables which contain descriptions of the dimensions we would like to use for slicing and dicing the data (e.g., geography, time, customer). But how do we come up with these dimensions? Facts? Decide on their granularity? Choose a variation of the star-schema?

To better understand how these star schemas are built; please go over the enclosed PowerPoint slides. The content of the PPT is as follows:

(1) Review of data modeling / entity-relationship diagrams

(2) Review of normalization forms → the source of problems for analytical systems

(3) Overview of the star schema concept and its components

(4) Overview of star-schema variations (e.g., multiple fact-tables, snow-flake)

(5) Discussion on granularity, and the time-dimension

(6) A structured approach a-la Ralph Kimball for developing star schemas (4 steps)

The approach we will be using is the one that Ralph Kimball describes (4-step approach), and is based on user needs. That is, the measures and dimensions we record are the ones that users identified as needed for their queries. It is worth mentioning that there is another approach, according to which ERD models can be translated directly to star-schemas. This approach is described in the attached articles. In realty, a combination of these approaches is used (i.e., understand the available operational data and its structure, combine it with knowledge of what queries users would like to run, and generate based on both sources, a star-schema). In the exam, we will focus on the first (i.e., you will not need to convert ERDs into dimensional models).

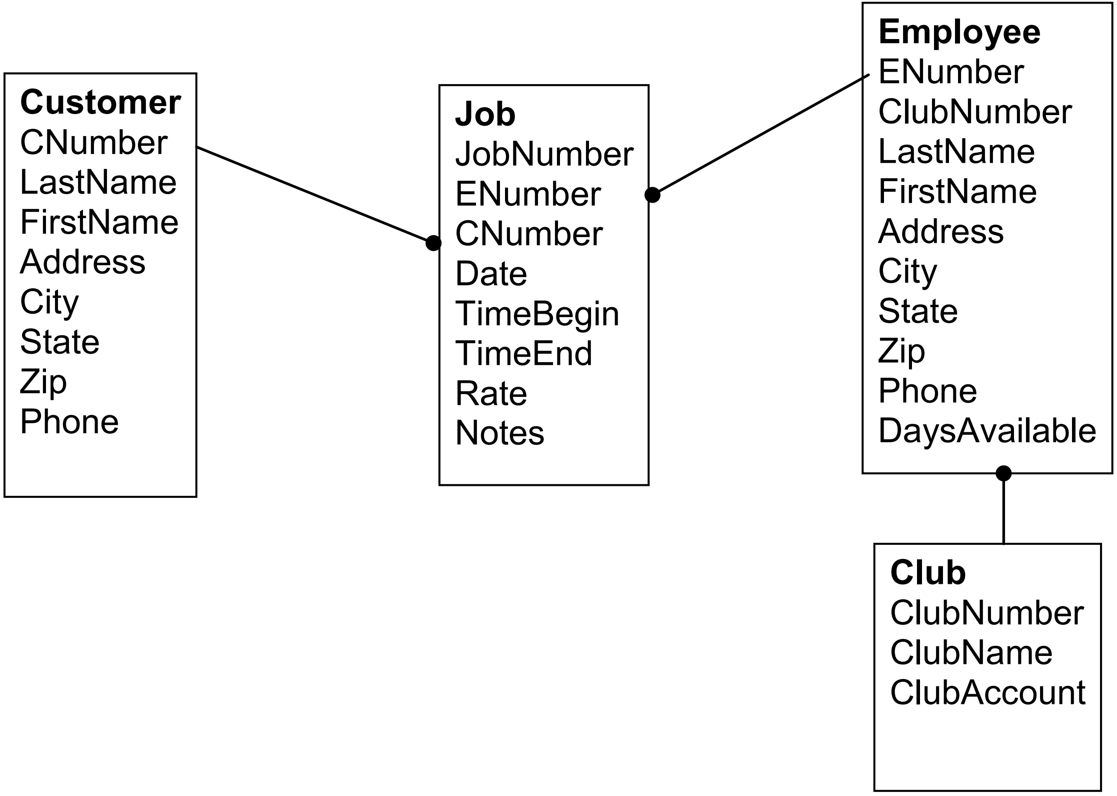
To further assist you with dimensional modeling, below is a set of exercises that will help you build-up your dimensional modeling skills. While the solutions follow the questions, try to initially solve the problems without consulting with the solutions.

1. Babysitter service

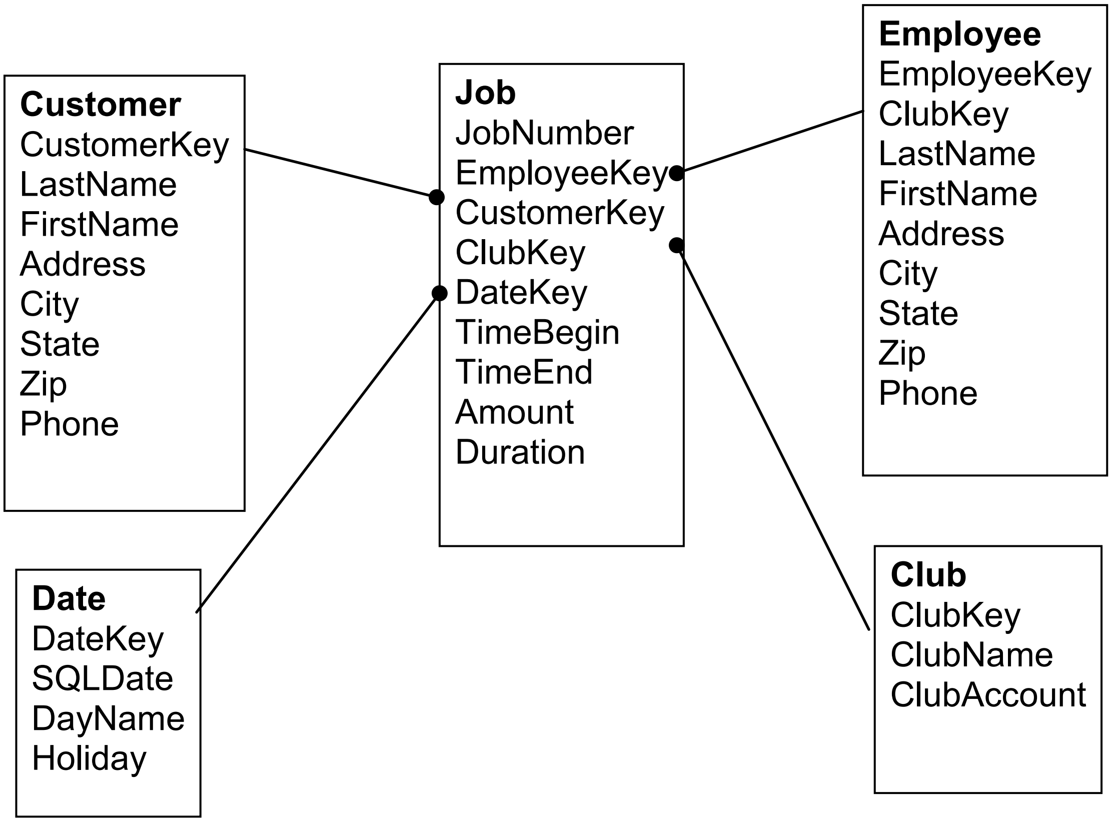
The CSUF Business Association is conducting a babysitter service as a fundraiser for different clubs in the college. When a customer is entered into the system, the CSUF Club coordinator gets name, address, and phone. The coordinator also records each babysitting job, the amount paid for it and the sitter assigned to the job. Each person may sign up to credit only one club and the system keeps the contact person and phone number for each participating club.

The treasurer wants a data warehouse for this. He would like to be able to determine how much each customer was billed by week, month or year. How much each employee (sitter) earned, also summed by time periods. He is interested in how much work is done on weekends, holidays or other special days. Develop a data warehouse to provide this information.

The transaction E-R diagram is below.



One Potential Solution:



2. Rental Agency

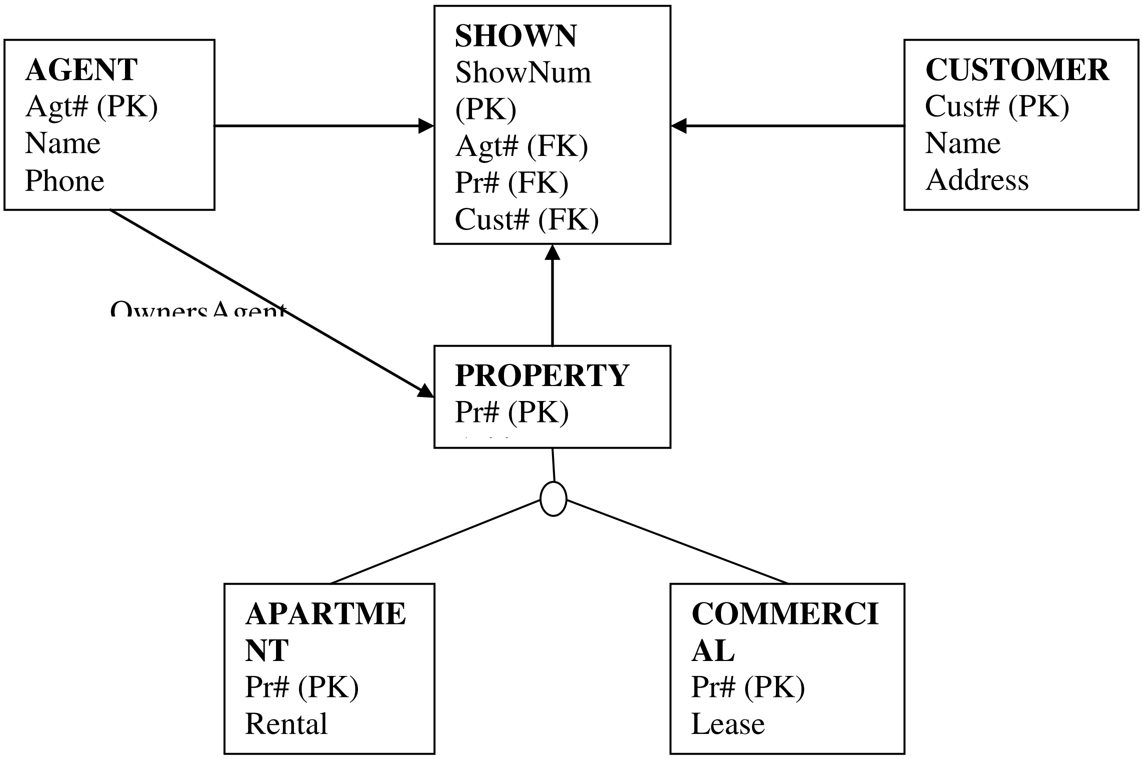
The Schooner Rental Agency rents apartments and commercial property in the Fullerton area. Each property has an address and owner. Apartments rent by the month and the system needs the monthly rental, the approximate utility bills, and an available date. Commercial property rents by the year and the system provides yearly lease cost, square feet, and number of parking spaces for each property.

Schooner Rental employs agents to rent the property. The system keeps name, address and phone number for each agent. Each property has a single "owners agent" who is responsible for working with the owner to assure that the property is maintained. There are a number of agents who are "renters agents" who show the property and try to obtain renters for it. When an agent shows a property, the system must show each date and time and potential customer that a given agent showed a given property to. Agents may be both owner and renter agents for a given property. Each property may have many renters' agents and each renter agent can show many properties several times to the same customer.

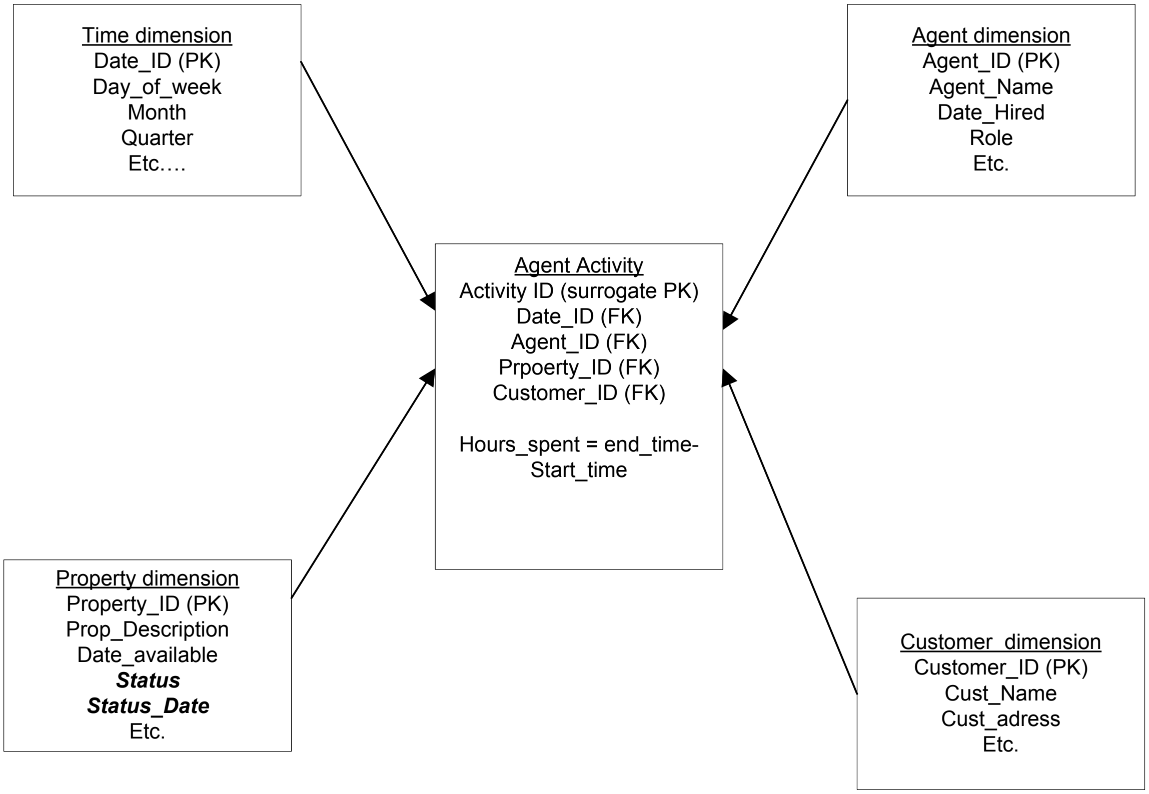
The system keeps track of name, address, and phone for each customer. See blow ERD.

Management wants to analyze agent activities. Produce an appropriate dimensional model.

Sooner Rental Agency ERD



Potential Solution (1):



Time dimension
Date_ID (PK)
Day_of_week
Month
Quarter
Etc.…

Agent dimension
Agent_ID (PK)
Agent_Name
Date_Hired
Role
Etc.

Agent Activity
Activity ID (surrogate PK)
Date_ID (FK)
Agent_ID (FK)
Prpoerty_ID (FK)
Customer_ID (FK)

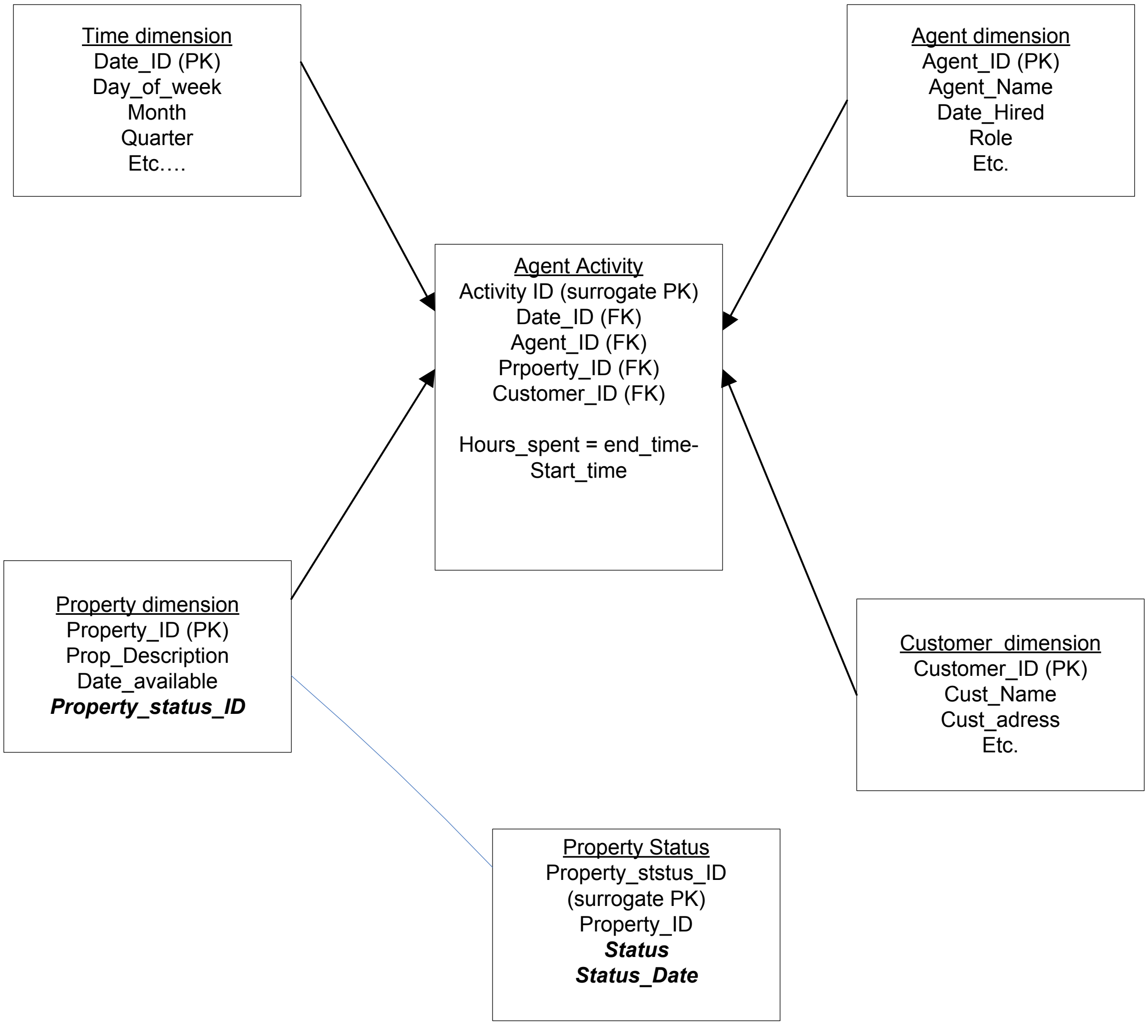
Hours_spent = end_time-Start_time

Property dimension
Property_ID (PK)
Prop_Description
Date_available
Status
Status_Date
Etc.

Customer_dimension
Customer_ID (PK)
Cust_Name
Cust_adress
Etc.

Potential Solution (2):

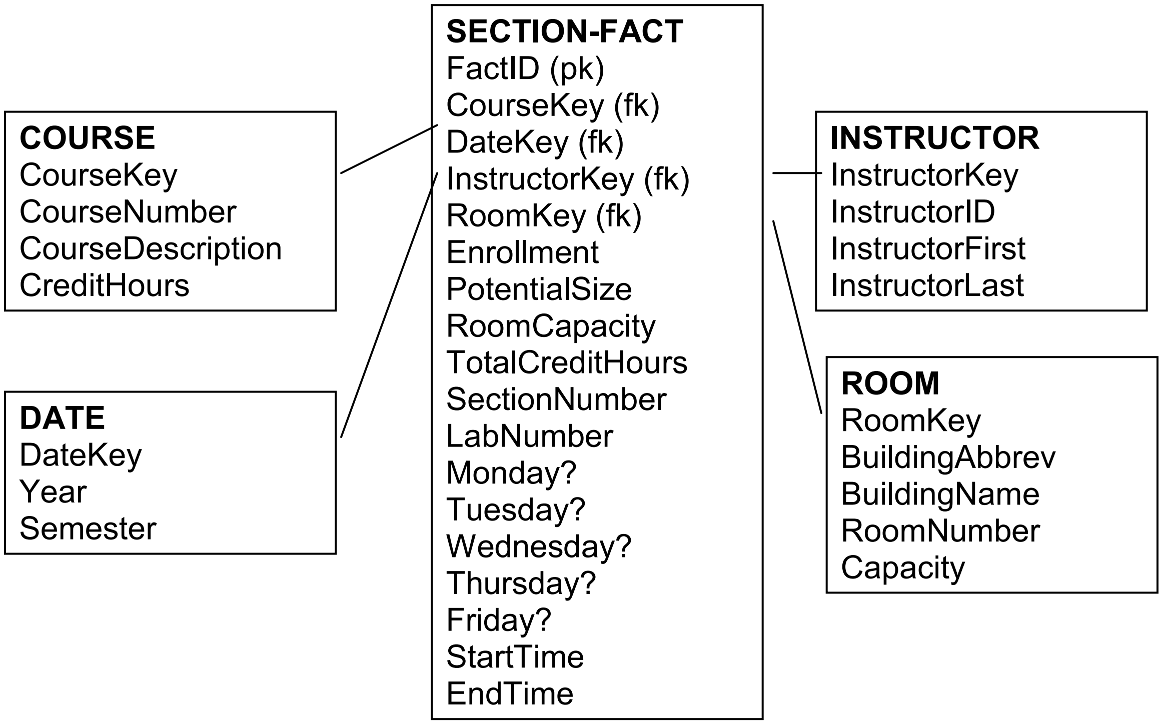
Note that this solution allows for recording multiple statuses for each property (i.e., property history), as opposed to the first solution, that allowed recording only the recent status.



3. Course Offerings System

- Goal (Business need): Analyze courses offered by different instructors to understand enrollment trends and adjust loads for potential classes.
- Sections are Offered Each Semester- Facts are organized by sections
- Dimensions:
    - Course
    - Instructor
    - Room
    - Date
- Fact:
    - Capacity
    - Enrollment

One Potential Solution:



SECTION-FACT
FactID (pk)
CourseKey (fk)
DateKey (fk)
InstructorKey (fk)
RoomKey (fk)
Enrollment
PotentialSize
RoomCapacity
TotalCreditHours
SectionNumber
LabNumber
Monday?
Tuesday?
Wednesday?
Thursday?
Friday?
StartTime
EndTime

COURSE
CourseKey
CourseNumber
CourseDescription
CreditHours

DATE
DateKey
Year
Semester

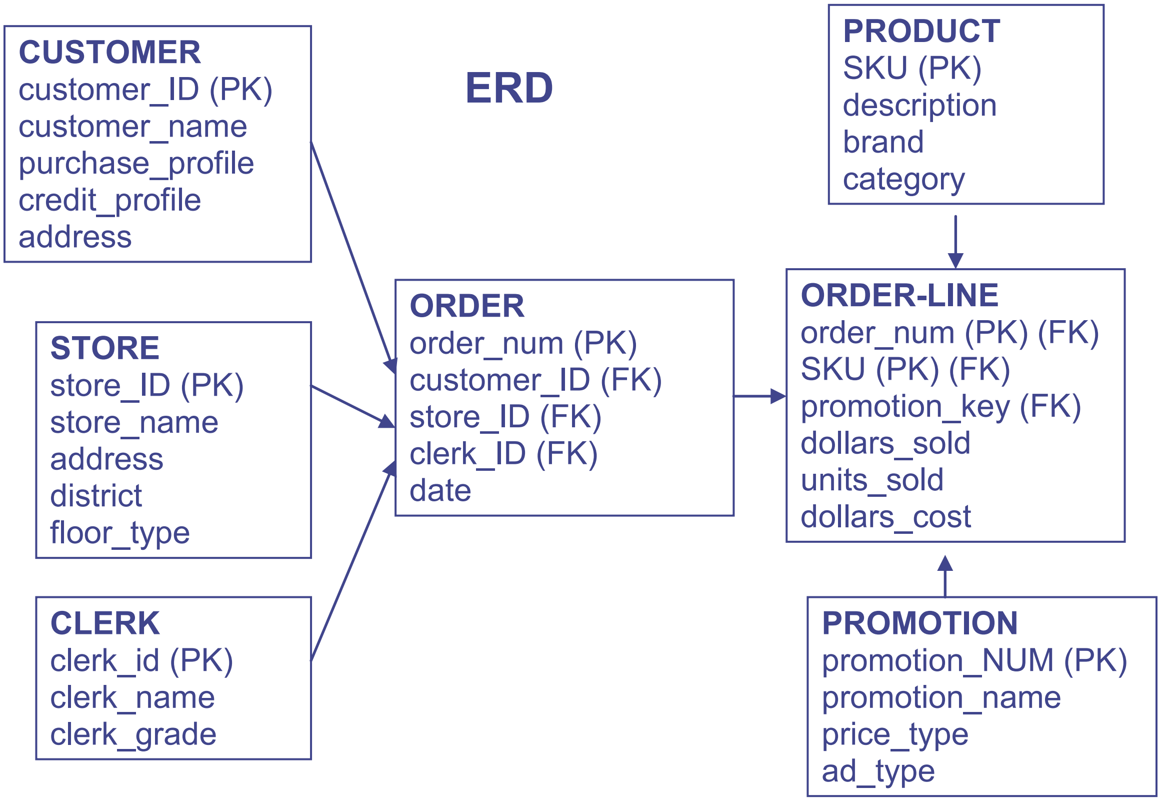
INSTRUCTOR
InstructorKey
InstructorID
InstructorFirst
InstructorLast

ROOM
RoomKey
BuildingAbbrev
BuildingName
RoomNumber
Capacity

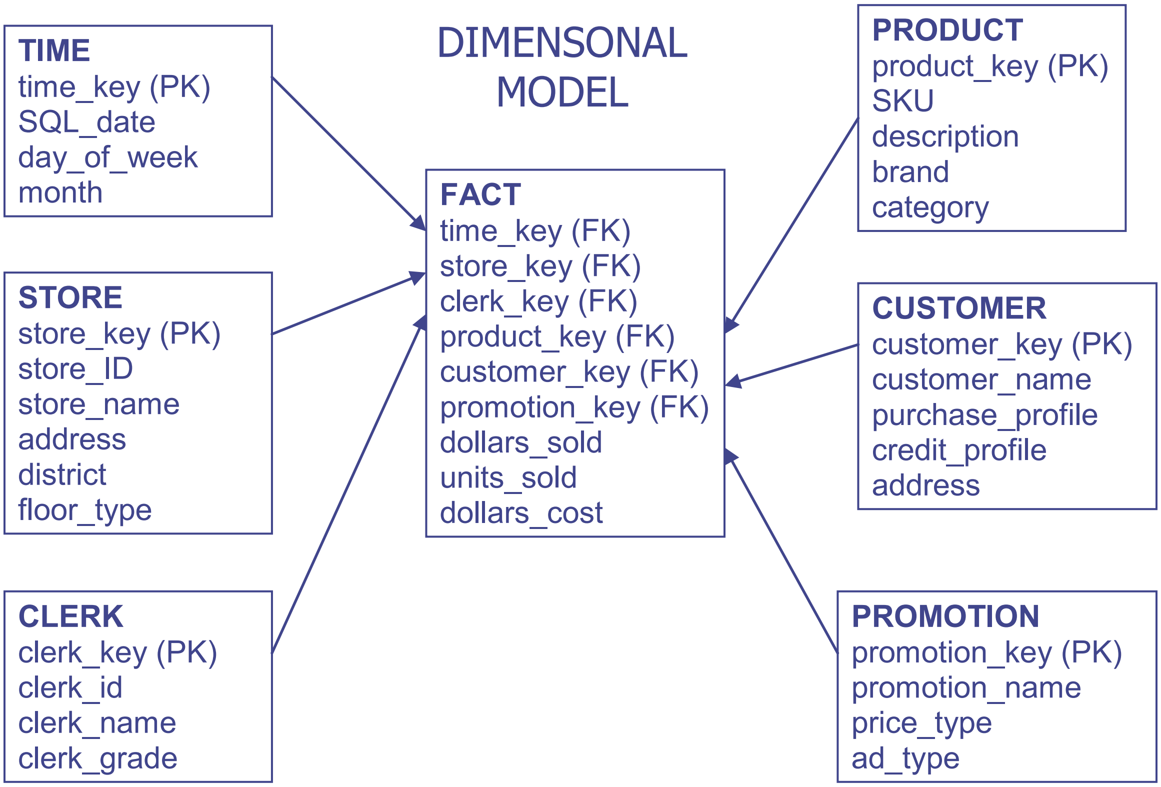
Note: another viable alternative is to include day_of_week field in the date (time) dimension.

4. Selling Office Supplies

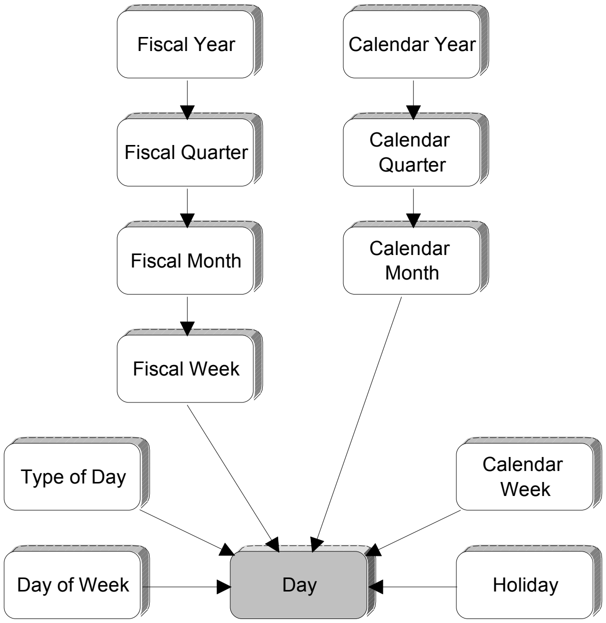
The following ERD is given for a retail chain that sells office supplies. Management would like to be able to analyze the profitability of the different products sold, taking into account the time of sale, customers, promotions, stores, and clerks. A potential report would, for example, compare the average sale prices for a certain product in different stores in a certain range of dates.



One Potential Solution:



Note 1: The time dimension may be decomposed into time-elements for facilitating faster and more precise analysis along the time dimension. Here is an example:

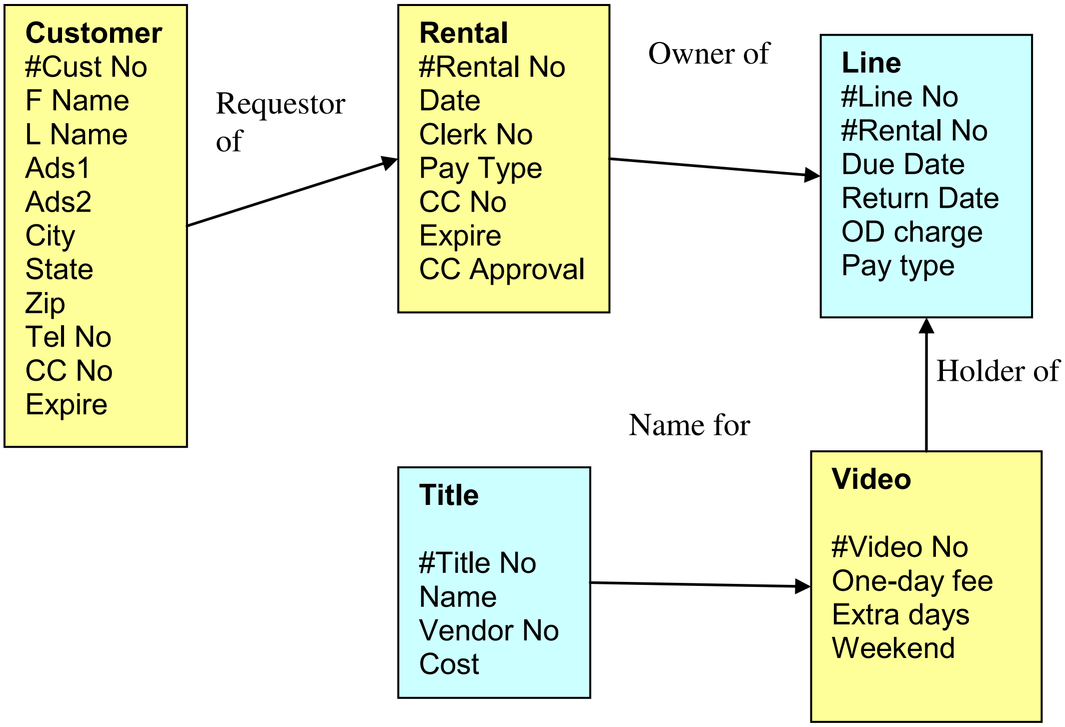


Attribute Name	Attribute Description	Sample Values
Day	The specific day that an activity took place.	06/04/1998; 06/05/1998
Day of Week	The specific name of the day.	Monday; Tuesday
Holiday	Identifies that this day is a holiday.	Easter; Thanksgiving
Type of Day	Indicates whether or not this day is a weekday or a weekend day.	Weekend; Weekday
Calendar Week	The week ending date, always a Saturday. Note that WE denotes	WE 06/06/1998; WE 06/13/1998
Calendar Month	The calendar month.	January,1998; February, 1998
Calendar Quarter	The calendar quarter.	1998Q1; 1998Q4
Calendar Year	The calendar year.	1998
Fiscal Week	The week that represents the corporate calendar. Note that the F	F Week 1 1998; F Week 46 1998
Fiscal Month	The fiscal period comprised of 4 or 5 weeks. Note that the F in the data	F January, 1998; F February, 1998
Fiscal Quarter	The grouping of 3 fiscal months.	F 1998Q1; F1998Q2
Fiscal Year	The grouping of 52 fiscal weeks / 12 fiscal months that comprise the financial year.	F 1998; F 1999

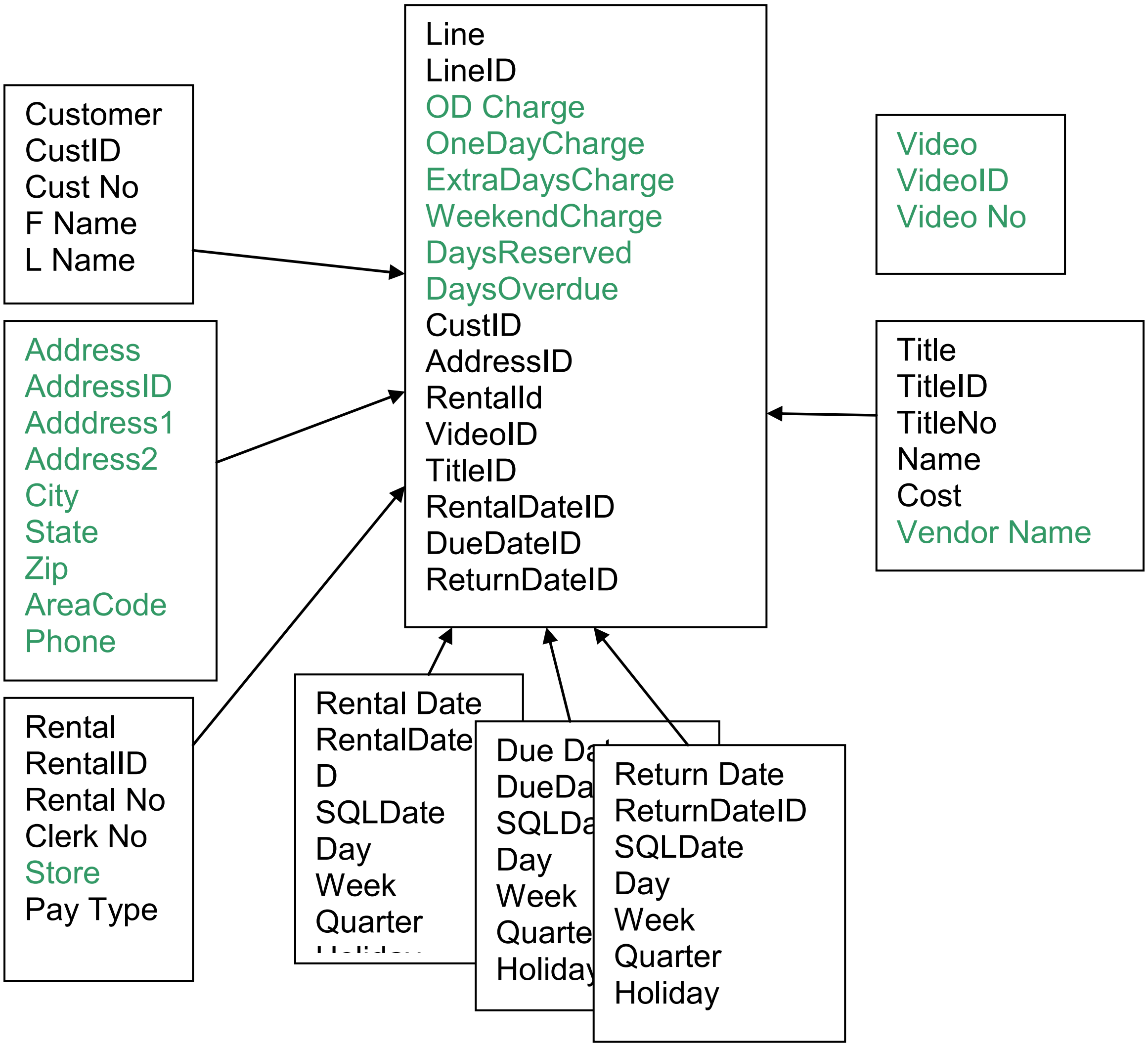
Note 2: The portrayed solution does not allow for analysis of issues such as the number of line-items per order (e.g., do we have more line-items per order on holidays?). For this, we need to include “order_number” in the fact table. This is an example of a “degenerate dimension”; i.e., a dimension key that is included in the fact table, but is not connected to any dimension table. See the enclosed PPT for more detail.

5. Video Rentals

MovieRental is interested in analyzing its video rental data. The data are recorded in the following operational ERD structure (a single title may have many copies/ videos; each rental may include multiple line-items):



One Potential Solution:



Note: This solution allows analyzing rentals along several time dimensions (return date, due date, rental date)

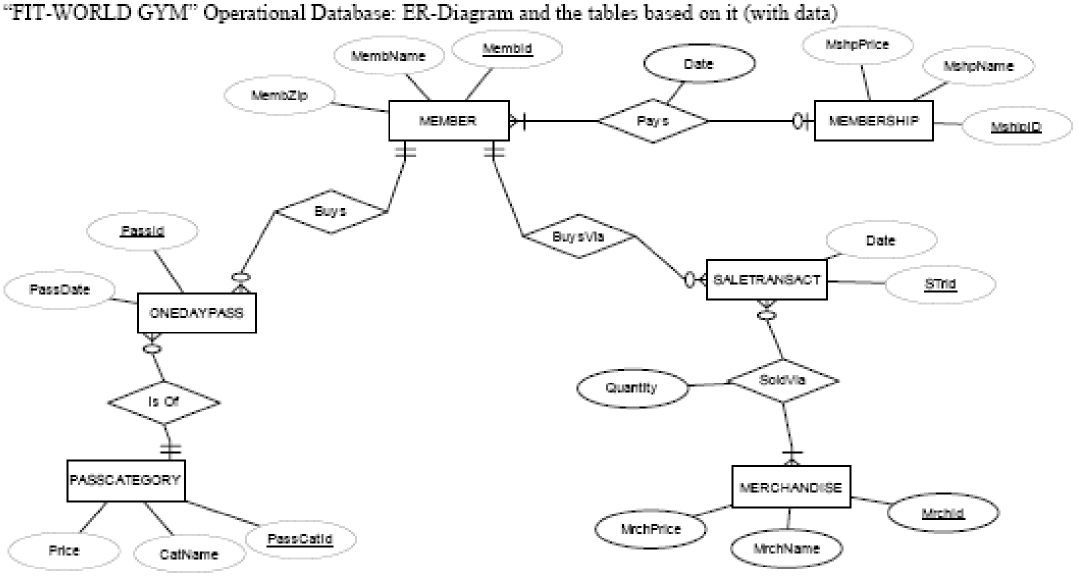
6. FIT-WORLD GYM

Create a star schema diagram that will enable FIT-WORLD GYM INC. to analyze their revenue.

- The fact table will include: for every instance of revenue taken – attribute(s) useful for analyzing revenue.
- The star schema will include all dimensions that can be useful for analyzing revenue.
- The only data sources available are shown bellow

SOURCE 1

"FIT-WORLD GYM" Operational Database: ER-Diagram and the tables based on it (with data)



MEMBER

MembId	MembName	MembZip	MchpID	McDatePayed
111	Joe	60611	M1	1-Jan-04
222	Mary	60640	M3	1-Jan-04
333	Sue	60611	M3	1-Jan-04

MEMBERSHIP

MshpID	MshpName	MshpPrice
M1	Platinum	$1,000
M2	Gold	$800
M3	Value	$300

ONE DAY PASS CATEGORY

* PassCatId	CatName	Price
PSA	Adult	$20
PSS	Senior	$10
PSK	Kid	$3

MERCHANDISE

MrchID	MrchName	MrchPrice
AP1	T-shirt	$11
AP2	Hat	$9
EQ1	Jump Rope	$12

* ONE DAY GUEST PASS

PassId	PassDate	PassCatId	MembId
1-001	1-Jan-04	PSA	111
1-002	1-Jan-04	PSA	333
1-003	2-Jan-04	PSK	333

SALE TRANSACTION

STrId	Date	MembId
11111	1-Jan-04	333
11112	2-Jan-04	222
11113	3-Jan-04	111

SOLD VIA

STrId	MrchId	Quantity
11111	AP1	1
11111	AP2	1
11112	AP2	1
11113	EQ1	3

* Members can bring in non-member guests. For each non-member guest, a member buys a one-day-guest-pass of a certain pass category.

SOURCE 2

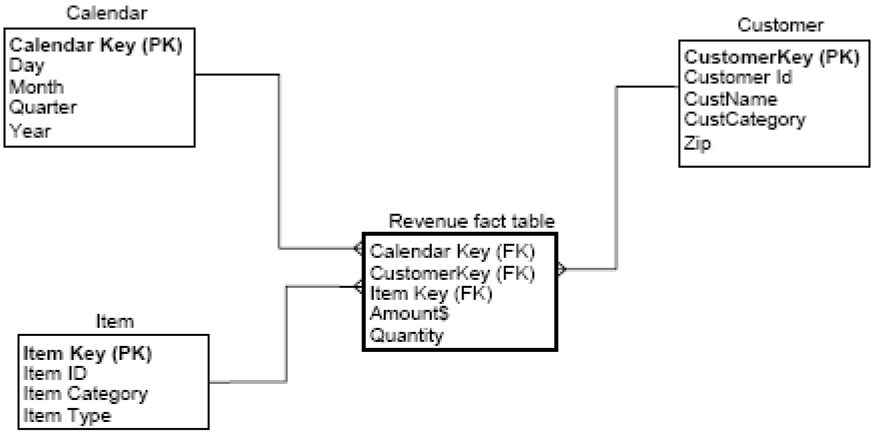
"FIT-WORLD GYM" Special Events Table

Corporate Customer ID	Corporate Customer Name and Location	Event Type Code	Event Type	Event Date	Amount Charged
CC1	Sears, Chicago 60640	L-A	All Day Rental,	January 4, 2004	$3500
CC2	Boeing, Chicago 60611	L-H	Half Day Rental,	January 5, 2004	$2200

One Potential Solution:

SOLUTION

Dimensional Model



Populated Tables

CALENDAR DIMENSION

Calendar Key	Day	Month	Quarter	Year
1	1	Jan	1	2004
2	2	Jan	1	2004
3	3	Jan	1	2004
4	4	Jan	1	2004
5	5	Jan	1	2004

ITEM DIMENSION

Item Key	Item Id	Category	Type
1	M1	Memship	Platinum
2	M2	Memship	Gold
3	M3	Memship	Value
4	PSA	One Day P	Adult
5	PSS	One Day P	Senior
6	PSK	One Day P	Kid
7	AP1	Mrch	T-Shirt
8	AP2	Mrch	Hat
9	EQ3	Mrch	Jump Rope
10	L-A	Spec. Ev.	All Day
11	L-H	Spec. Ev.	Half Day

CUSTOMER DIMENSION

Cust Key	Cust ID	CustName	CCategory	Zip
1	111	Joe	Ind	60611
2	222	Mary	Ind	60640
3	333	Sue	Ind	60611
4	CC1	Sears	Corp	60640
5	CC2	Boeing	Corp	60611

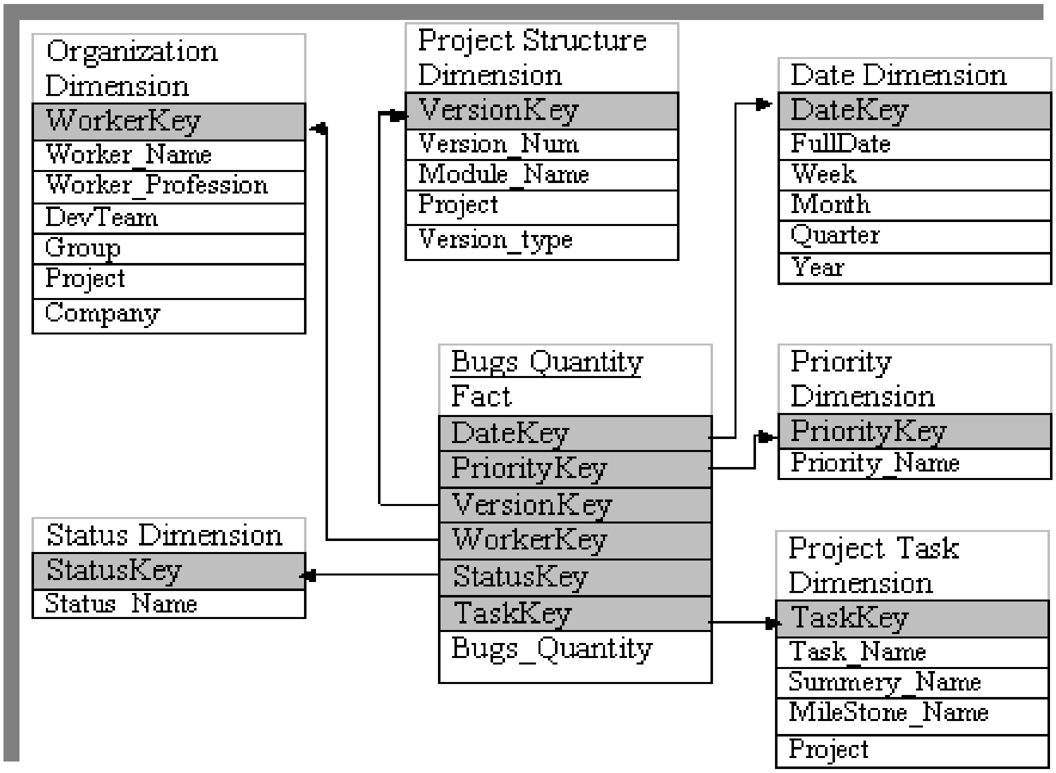
FACTREVENUE

Calendar Key	Cust Key	Item Key	Amount	Quant
1	1	1	$1,000	1
1	2	3	$300	1
1	3	3	$300	1
1	1	4	$20	1
1	3	4	$20	1
2	3	6	$3	1
1	3	7	$11	1
1	3	8	$9	1
2	2	8	$9	1
3	1	9	$36	3
4	4	10	$3,500	1
5	5	11	$2,200	1

7. Software Projects at CSUF

The IT department at CSUF wants to monitor software quality issues, as measured by number of bugs. Projects may have different priorities, can pertain to different tasks, have different structures, and different statuses. Assume that each project is assigned to a single employee. Draw a potential star schema that will facilitate analyses of software bugs.

Potential Solution:



Organization Dimension
WorkerKey
Worker_Name
Worker_Profession
DevTeam
Group
Project
Company

Project Structure Dimension
VersionKey
Version_Num
Module_Name
Project
Version_type

Date Dimension
DateKey
FullDate
Week
Month
Quarter
Year

Bugs Quantity Fact
DateKey
PriorityKey
VersionKey
WorkerKey
StatusKey
TaskKey
Bugs_Quantity

Priority Dimension
PriorityKey
Priority_Name

Status Dimension
StatusKey
Status_Name

Project Task Dimension
TaskKey
Task_Name
Summery_Name
MileStone_Name
Project