DATA WAREHOUSING AND MINING

T.E. CSE-Data Science , Sem V
Academic Year: 2022-23

Data Warehousing Fundamentals: ER vs Dimensional Modelling, IPD
Lecture 4

Poonam Pangarkar

Designing a Data Warehouse

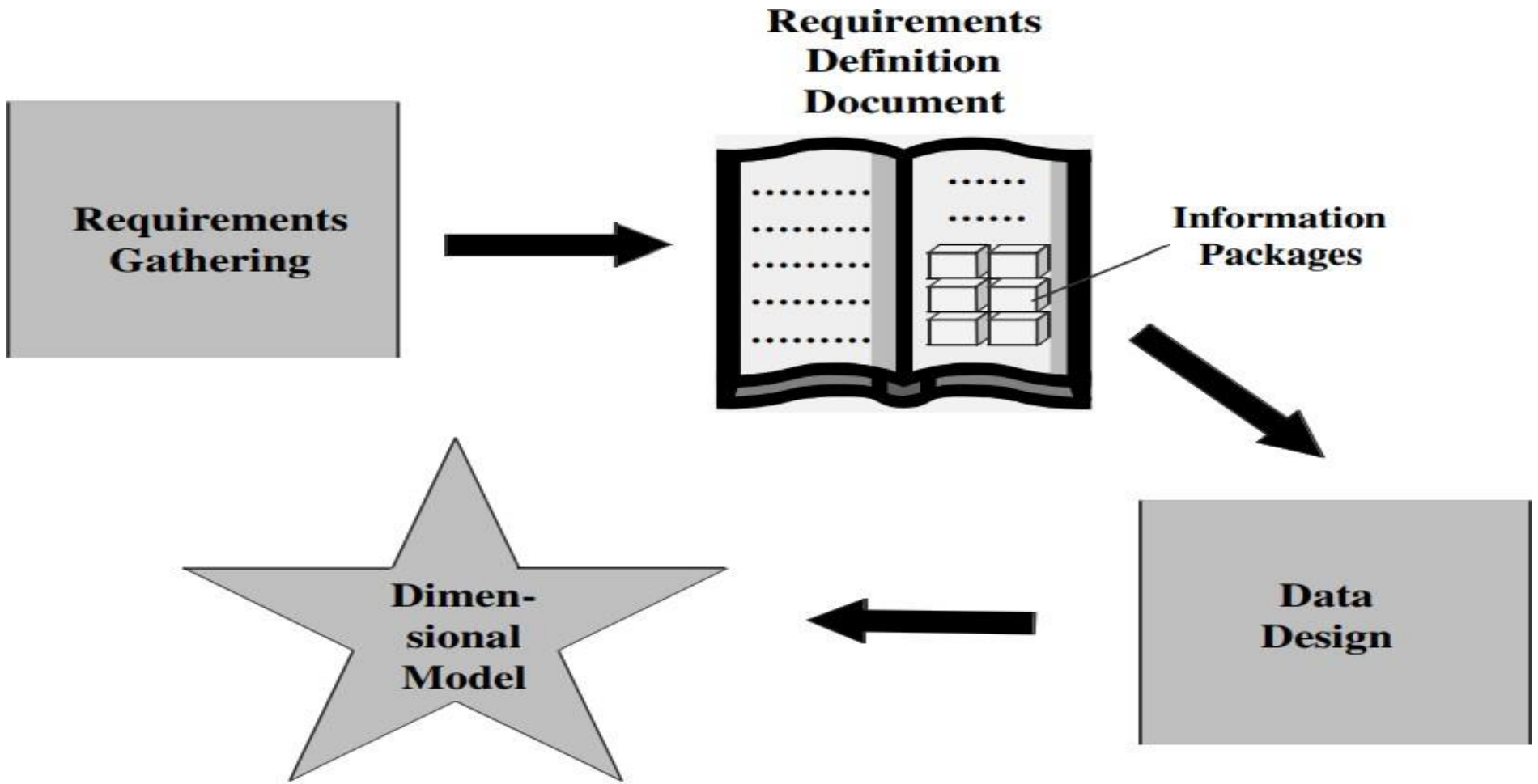


Figure 10-1 From requirements to data design.

Facts and Measures

- What exactly gets analysed?
- Facts or Metrics - measure the success of the business
- The requirement must represent two elements of the analysis

  ✓What is being analysed? - Dimensions
  ✓Evaluation criteria for what is being analysed - measures

Fact : a number or an integer value e.g. total sale of a day

Information Package Diagrams(IPD)

IPD enables the DW project team to define the subject area’s, key business metrics, the way in which the user will aggregate the data, the data granularity, and the size of the DW

It is a technique to gather requirements for the DW system that is based on the metrics and business dimensions.

For every specific subject we form an IPD

E.g. 1. IPD for student monitoring system

2. IPD for sales analysis

3. IPD for hotel business

IPD for student's performance

Subject: Student's performance					
Facts: Attendance, aggregate					
Time dimension	Day	Week	Month	Semester	Year
Professor dimension	Name	Professional qualification	Experience	Grade	No. of subjects
Subject dimension	Name	Semester	Theory/ practical		
Student dimension	Name	Grad/PG	Family income	Division in class XII	Grade
Course dimension	Name	Duration	Type	University	

Figure 4.3 IPD for a student monitoring system

IPD for sales

Subject: Sales					
Facts: Actual sales, forecast sales, price, discount					
Time dimension	Day	Week	Month	Quarter	Year
Product dimension	Name	Brand	Category	Colour	Price
Customer dimension	Name	Age	Income	Gender	Marital status
Store dimension	Name	City	State	Country	Operational from year
Payment method dimension	Payment type	Interest rate			

Figure 4.4 IPD for sales analysis

IPD for Hotel Occupancy

Subject: Hotel occupancy						
Facts: Occupied rooms, vacant rooms, unavailable rooms, occupants, revenue						
Time dimension	Day	Week	Month	Quarter	Year	Holiday flag
Hotel dimension	Branch name	Branch code	Region	City	Construction year	Renovation year
Room dimension	Room type	Room size	No. of beds	Bed size	Max No. of occupants	Price
Facilities dimension	Gym	Side view	Swimming pool	Cultural activities	Out door games	SPA

Figure 4.5 IPD of the hotel business

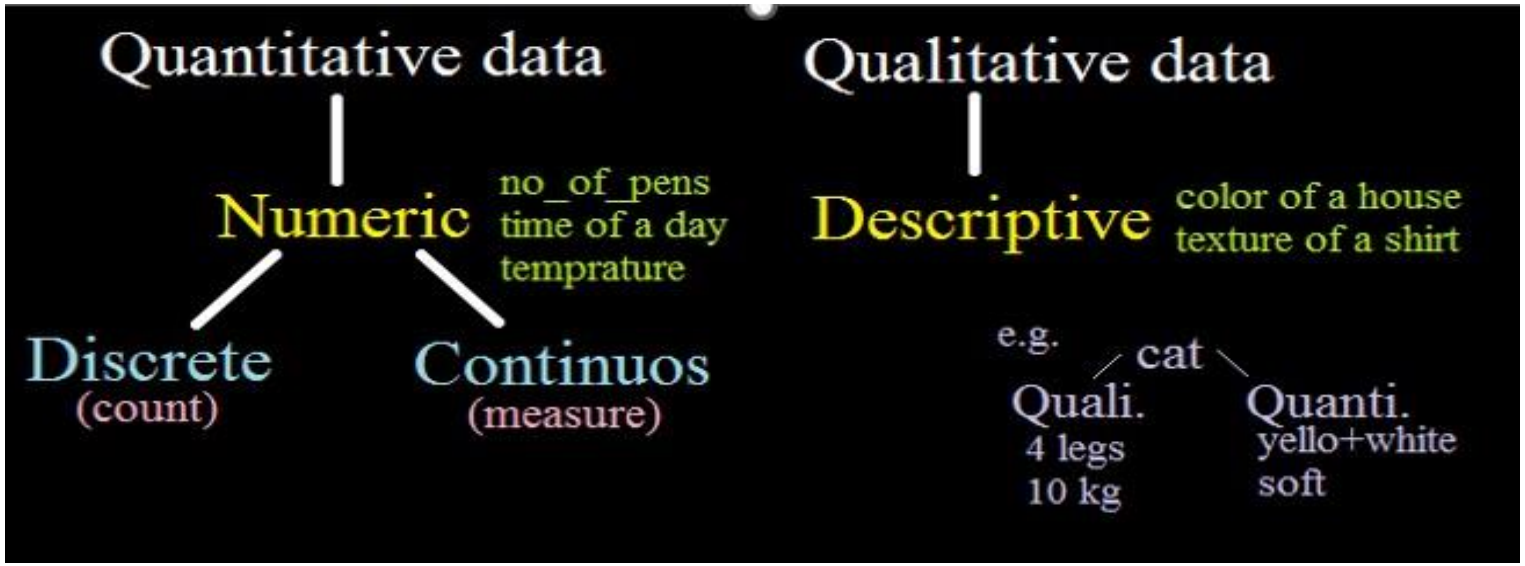
Why use IPD

1. IPD can help us identify Dimension Tables and Fact Table.
2. For a given IPD there will be n number of dimension tables where n is the number of dimensions identified in IPD.
3. The hierarchies /categories in the IPD will become the attributes of the dimension tables.
4. One more attribute will included apart from the attributes from IPD – the key attribute.
5. It represents the primary key of every individual record that will be stored in the dimension tables.

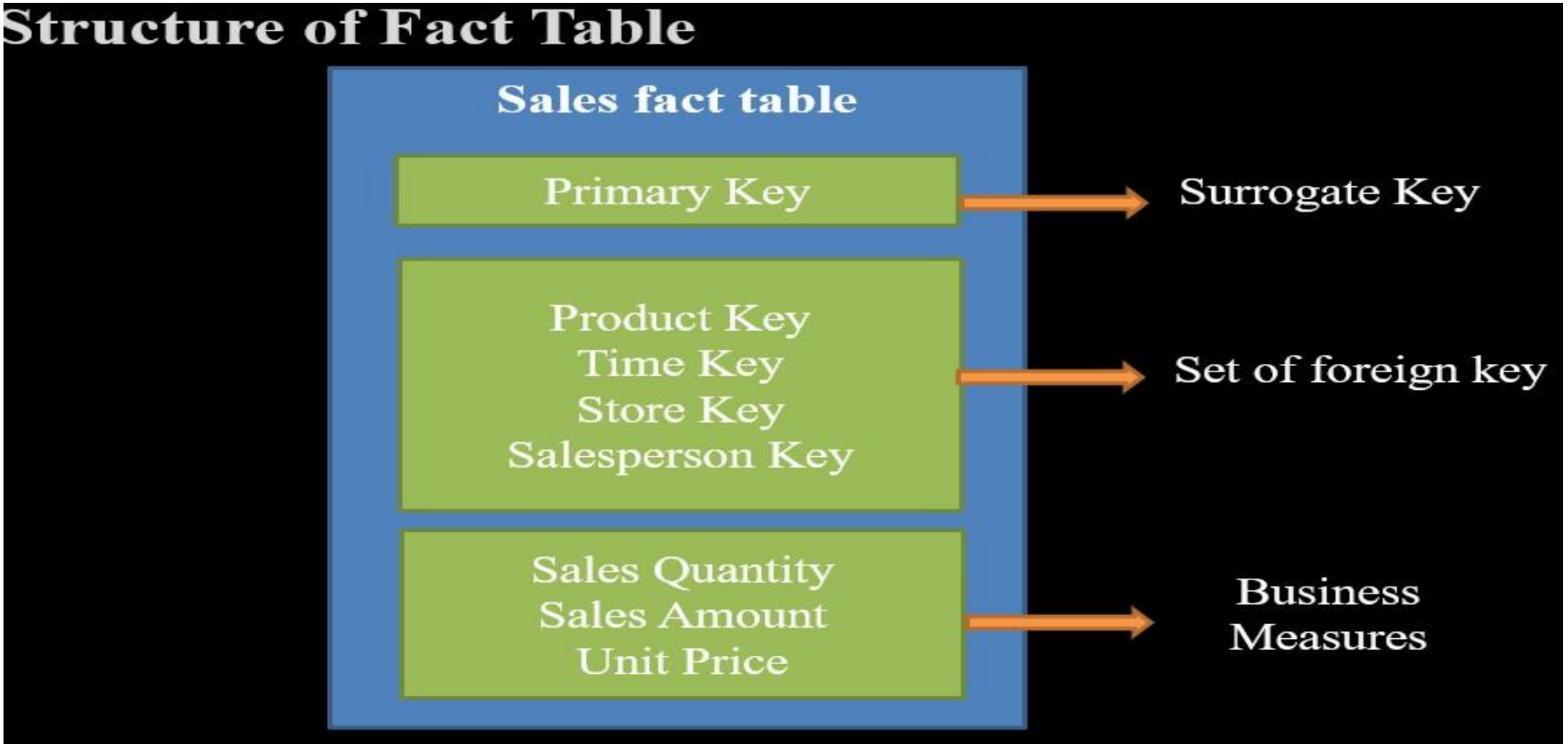
Fact Tables:

1. Fact is an entity that provides quantitative information about process. It is also called as measurements or matrices, they are mostly numeric in nature
2. E.g. profit, margin, total earning, sales amount, total turnover etc.

what is quantitative and qualitative data



Structure of Fact Table



Exercise

1. Generate fact table and Dimension tables for retail sales using IPD created earlier

Time Dimension Table
Time Key
Year
Quarter
Month
Year

Product Dimension Table
Product Key
Name
Brand
Category
Color
Price

Sales Fact Table
Time key
Customer key
Product key
Store key
Actual Sale
Forecast Sale
Price

Customer Dimension Table
Customer Key
Name
Age
Income
gender
Marital Status

Store Dimension Table
Store key
Name
city
state
country
year

	Sales	Profit	Qty
16	2000	200	1
17	1000	50	1
18	2500	250	1

	Sales	Profit	Qty	Customer	Product	Store
16	2000	200	1	Cx	Px	Sx
17	1000	50	1	Cy	Py	Sx
18	2500	250	1	Cz	Pz	Sx
22				Cx		

Fact / Measures(Metrics/KPIs)

Fact is a Numeric Value which have impact on business.

Ex : Sales, Profit, Quantity, Discount

Measure is an aggregation of Fact/Dimension.

Ex : Sum of sales,
Avg of Profit,
Count of Items.
Count of Customers

Primary Key

- A primary key is a field in a table which uniquely identifies each row/record in a database table.
- Primary keys must contain unique values
- Primary key column cannot have NULL values.
- A table can have only one primary key, which may consist of single or multiple fields. When multiple fields are used as a primary key, they are called a composite key.

Foreign Key

- A Foreign Key is a key used to link two tables together.
- Foreign key can contain duplicate values.
- Foreign key column can have NULL values.
- A FOREIGN KEY is a field (or collection of fields) in one table that refers to the PRIMARY KEY in another table.
- The table containing the foreign key is called the child table, and the table containing the primary key is called the referenced or parent table.

Primary Key / Foreign Key Examples

```
CREATE TABLE CUSTOMERS(
CUST_ID INT NOT NULL,
NAME VARCHAR (20) NOT NULL,
AGE INT NOT NULL,
ADDRESS CHAR (25) ,
SALARY DECIMAL (18, 2),
PRIMARY KEY (CUST_ID) );

CREATE TABLE ORDERS
( ORDER_ID INT NOT NULL,
DATE DATETIME,
CUSTOMER_ID INT references CUSTOMERS(CUST_ID),
AMOUNT double,
PRIMARY KEY (ORDER_ID) );
```

Surrogate Key vs Primary Key

1. Both keys contain a unique value for a record in a table.
2. Primary keys are used in OLTP whereas surrogate keys are used in OLAP schemas.
3. Primary keys hold some business meaning whereas surrogate does not hold any business meaning.
4. Primary key may contain numeric as well as non-numeric values whereas surrogate keys contain only (simple)numeric values(auto-increment).

Customer_id (PK)	Customer_name	Phone_number	State
Q7A55Q8A	Leonard F.	855-966-566	California
A8DG8B2H	Sherlock H	967-968-563	NewYork
Q8W2C26F	Monica Bing	852-563-988	New York
D2B2GNHY	Rachel	859-563-556	New York
C56T89THE	Sheldon Cooper	859-898-456	California

Customers (dim table)

Product_id (PK)	Product Name	Product Description
ABHEUMSHR	ABC Laptop	ABC Laptop, 8BG, 256GB
OLRGIDTBRM	ABC Mobile	ABC Mobile, 6.1, 4GB, 64 GB
DHBVDVADV	ABC Laptop Charger	ABC laptop charger with adaptor
LKMTNYKOEF	Headphones	Wireless headphones
ZDVBFSGBGD	Tony TV	56' Smart TV

Products (dim table)

Customer_id

Product_id

Transaction_id	Product_id (FK)	Customer_id (FK)	Transaction_date	Quantity	Amount
123456789	ABHEUMSHR	Q7A55Q8A	24th Mar 2021	1	30000
547896124	OLRGIDTBRM	A8DG8B2H	01st Nov 2021	1	10000
475896125	DHBVDVADV	Q8W2C26F	11th Feb 2021	5	500
859641237	LKMTNYKOEF	D2B2GNHY	31st Dec 2021	1	1500
178594025	ZDVBFSGBGD	A8DG8B2H	14th Jan 2021	12	1200
685247830	OLRGIDTBRM	C56T89THE	06th April 2021	10	10000
965214738	DHBVDVADV	D2B2GNHY	25th May 2021	15	15000
605853459	DHBVDVADV	Q7A55Q8A	8th Oct 2021	6	600
259035721	ABHEUMSHR	Q7A55Q8A	19th Jun 2021	2	56000
741256780	OLRGIDTBRM	A8DG8B2H	01ST Dec 2021	2	10000

Transactions (fact table)

Fact tables and Dimension tables

- Fact tables:
  - It contains measurements, facts or metrics of the attributes
  - Hold no meaning in itself
  - Numeric and quantifiable
  - Created or loaded after dimensions are loaded
  - Primary Key is a new column identifying the unique row, references dim tables with FK
- Dimension tables:
  - Gives context to facts, holds attributes for the facts
  - Created or loaded before facts are loaded
  - Primary key is referenced by fact tables

Note: All numeric column that has impact on business called as fact. doesn’t have impact on business is called as dimensions.

Operational database modelling Vs. Datawarehouse modelling

Operational data modeling

- Data collected in terms of entities, attributes and relationship among them
- Database users thinks in terms of entities
- In requirement analysis users are able to give enough details of the required functions , information content and usage patterns of data
- It is application oriented

Datawarehouse data modeling

- Data collected in terms of dimensions, facts or measures
- DW users thinks in terms of dimensions
- users are unable to define their requirements clearly
- It is not limited to certain applications

Operational database modelling Vs. Datawarehouse modelling

Operational data modeling

- Users have an idea about certain activities need to be performed for the various tasks they wish to accomplish with the system
- Can visualize the scope and benefits of data
- E-R diagrams are used to gather information and define schemas

Datawarehouse data modeling

- Users don’t have an idea about what information they want from the DW, even they can not express how they would like to use the DW information
- Can not visualize the scope and benefits of data
- Information Package Diagrams(IPD) are used to determine requirements for a DW system(dimension tables and facts tables are important components of IPD’s)

ER vs Dimensional

ER

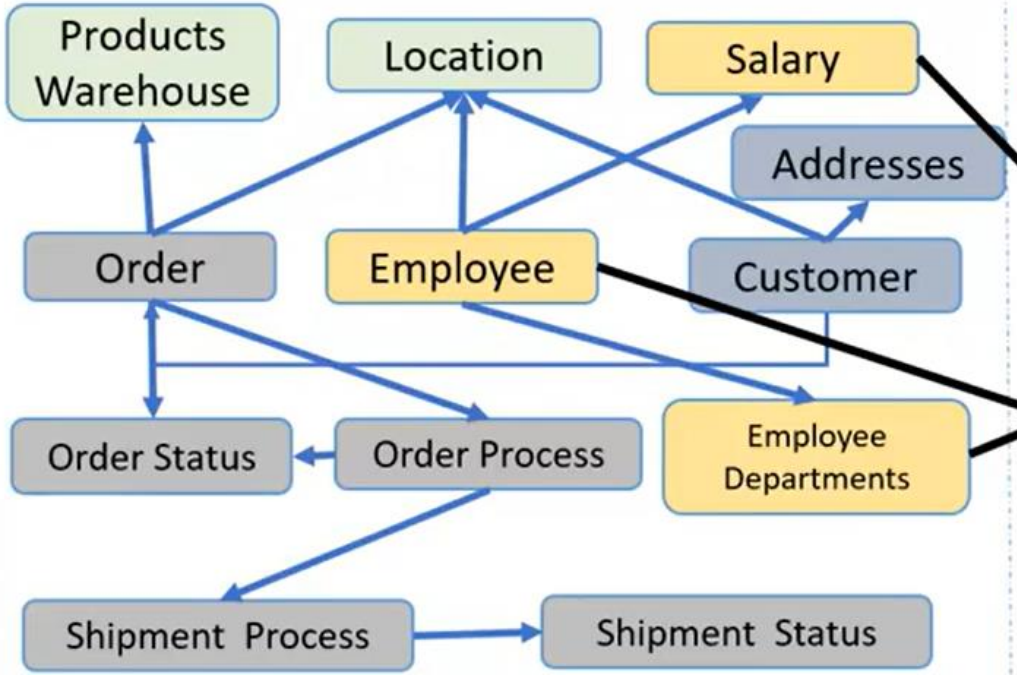
Data Oriented

No Data Redundancy

Entities

Need Expertise and understand of the model

Optimized for transactions (Update/Delete/Insert)



Dimensional

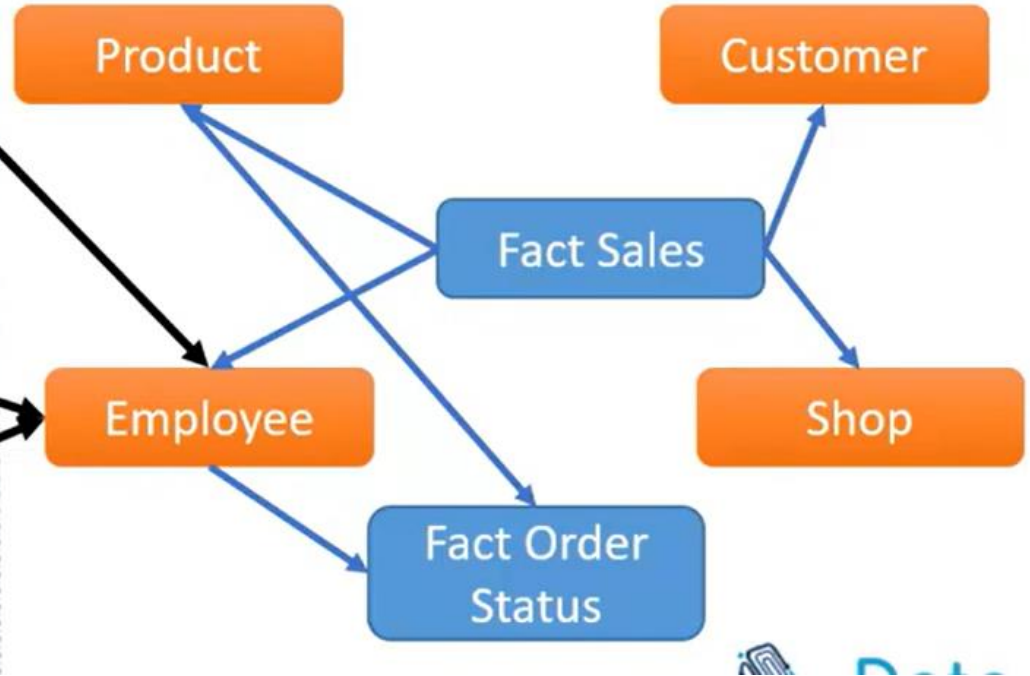
Business Oriented

Data Redundancy

Facts and Dimensions

Easy to Understand by Normal Users

Optimized for Query Performance and Analytics



ER Model vs Dimensional Model

E-R Model

- E-R modeling is used the designing tables for operational systems
- E-R diagrams are complex diagrams that are used to represent multiple processes
- E-R model is designed to express microscopic relationship between the data elements

Dimensional Model

- Dimensional modeling is used for designing tables for datawarehouse
- A single E-R diagram can be broken down into multiple dimensional modeling diagrams
- Key idea behind this model is to capture business measures

ER Model vs Dimensional Model

E-R Model

- E-R model is well suited to answer queries at transaction level
- Type/example of queries that we can execute: 1D, 2D
- How many units of product_X were sold?

Dimensional Model

- A dimensional model is designed to answer queries on the overall business process to reveal trends and to grow business
- Type/example of queries that we can execute: 3D, Multidimensional
- How many units of product_X were sold on 17 Jan 2018 in Thane branch?

Features of a good Dimensional modelling

1. Best data access
2. Optimized for queries and analysis
3. Depict the way in which the fact table interacts with the dimension table
4. Allow equal interaction of every dimensional table with the fact table
5. Enable the users to perform roll up and drill down operations along dimension hierarchies