Sub: DWM Sem : 8

Course Owner : Prof I.R. jamkhandikar

Academi Year : 2017/ 18

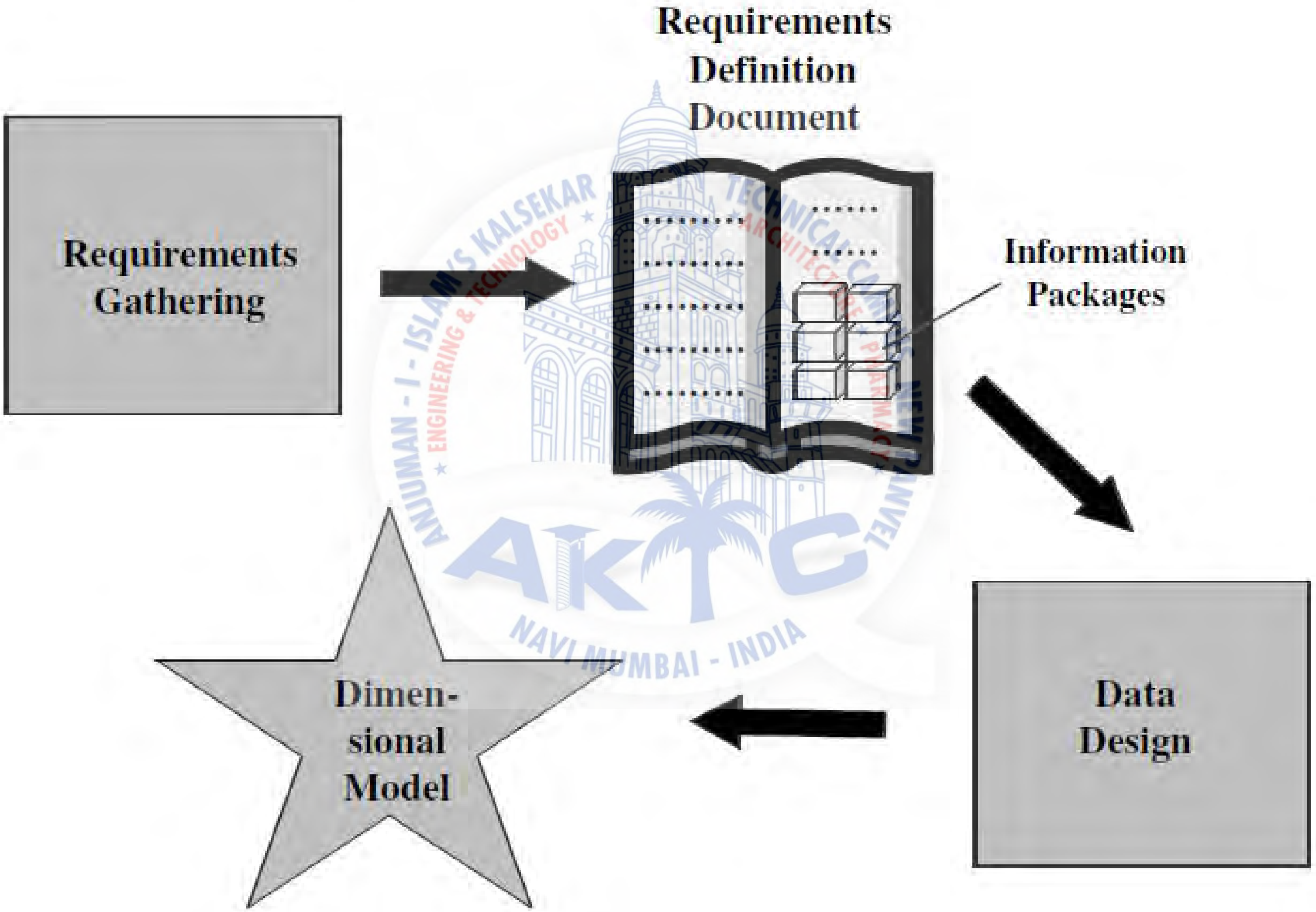
Principles of Dimensional Modeling

I.R. Jamkhandikar

Objectives

- Understand how requirements definition determines data design
- Introduction of dimensional modeling /contrast with E/ R modeling
- Basics of star schema
- Contents of fact/dimension tables
- Advantages of star schema for DW

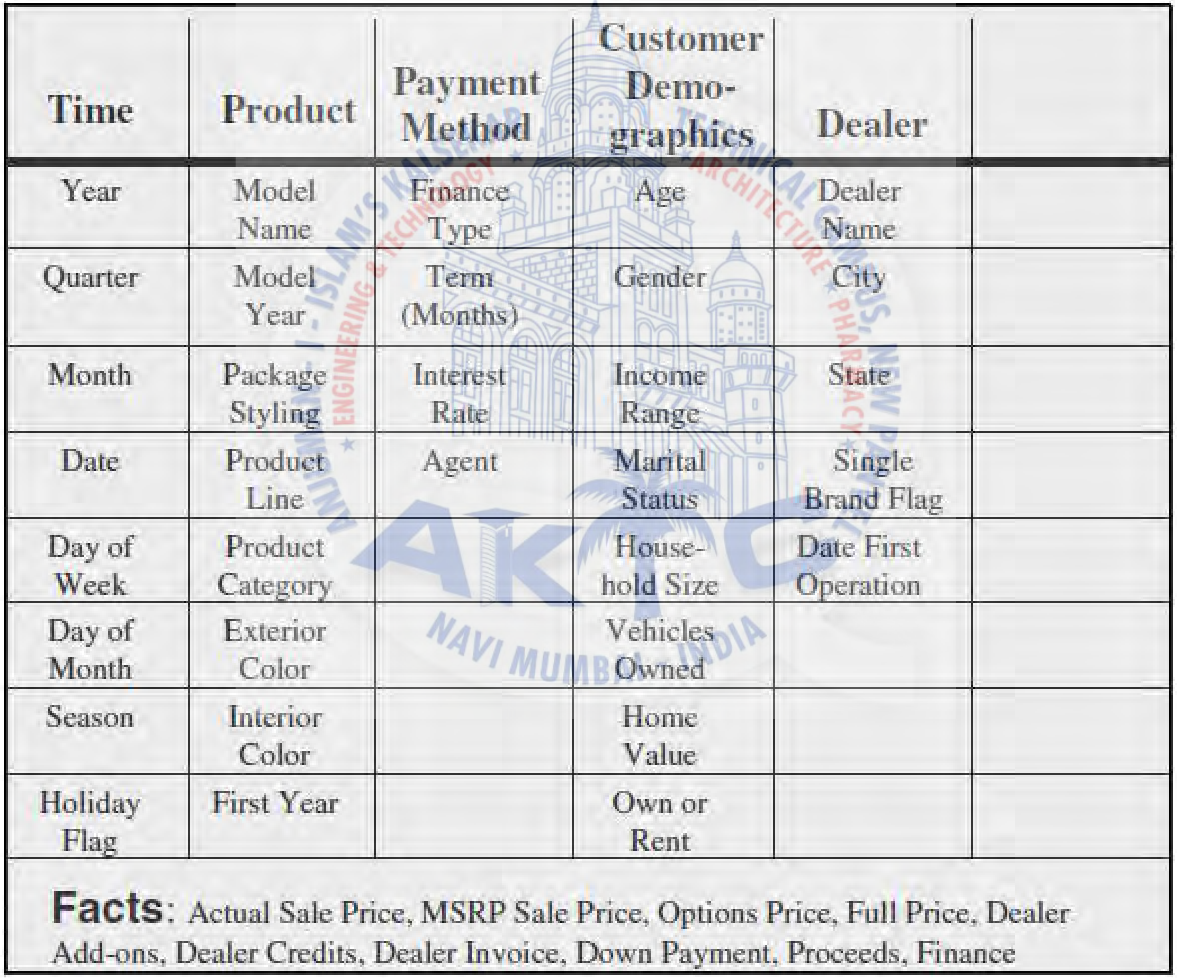
Requirements to Design



Design decisions to be taken

- Choosing the process:/ deciding subjects
- Choosing the grain
- Identifying and confirming dimensions
- Choosing the facts
- Choosing the duration of the database

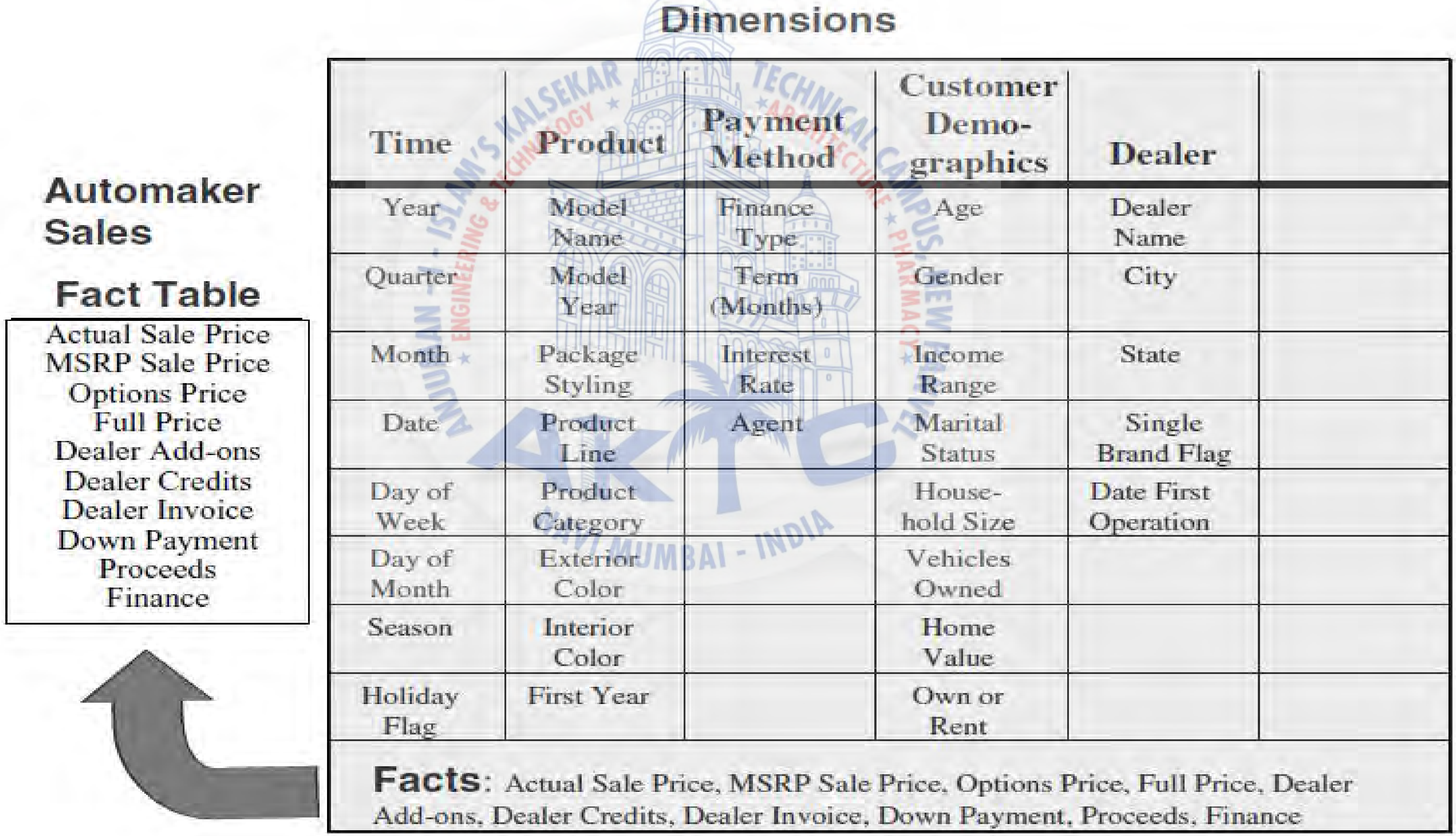
Dimensional modeling basics



Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

Formation of the automaker sales fact table



Dimensions

Automaker Sales

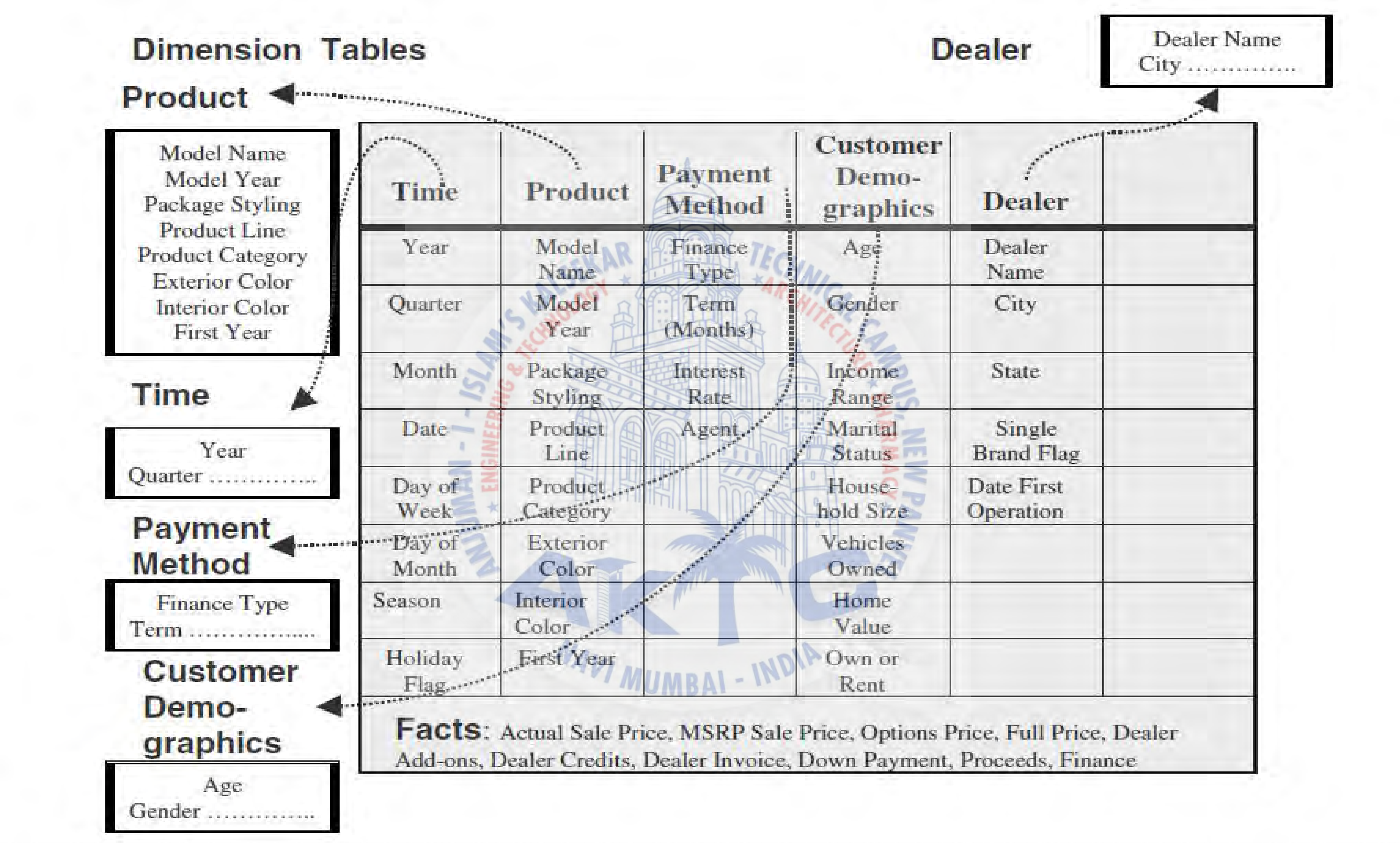
Fact Table

Actual Sale Price
MSRP Sale Price
Options Price
Full Price
Dealer Add-ons
Dealer Credits
Dealer Invoice
Down Payment
Proceeds
Finance

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

Formation of the automaker dimension tables



Dimension Tables

Product

Model Name
Model Year
Package Styling
Product Line
Product Category
Exterior Color
Interior Color
First Year

Time

Year
Quarter …………..

Payment Method

Finance Type
Term ……………....

Customer Demo-graphics

Age
Gender …………..

Dealer

Dealer Name
City …………..

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

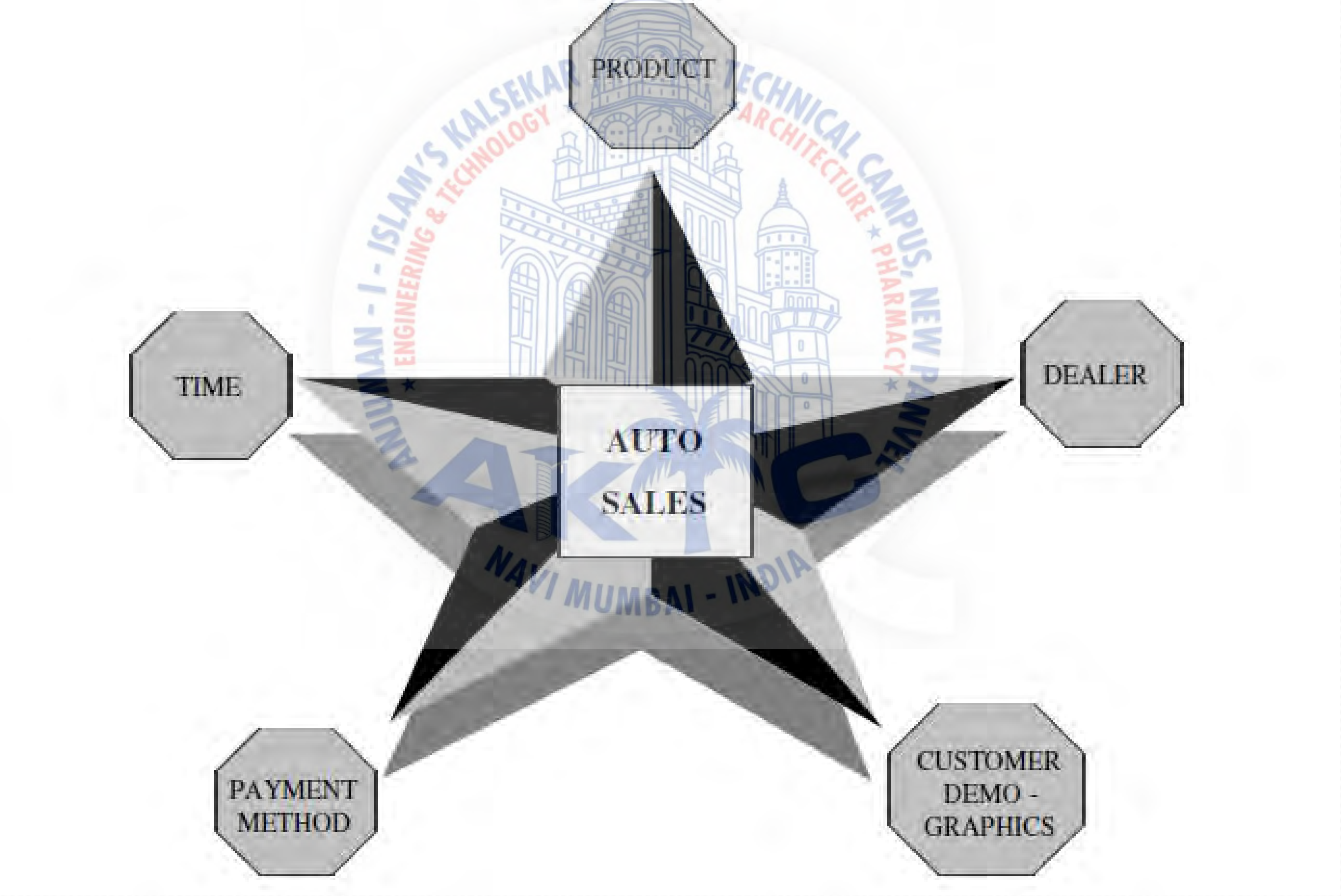
Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

How much sales proceeds did the jeep tata mahindra, 2009 model with vxi options, generate in january 2000 at spectra auto dealership for buyers who owned their homes, financed by icici prudential financing ?

Tips for combining data into dimensional model

- Provide best data access
- Model should be query/ centric
- Model should be optimized for queries and analyses
- Model should reveal the interactions between the dimension and fact tables
- There should be drilling down or rolling up along dimension hierarchies

STAR SCHEMA for automaker sales



ER Model V/s Dimension Model

- ER diagram is a complex diagram, used to represent multiple processes. A single ER diagram can be broken down into several DM diagrams.
- In DM, we prefer keeping the tables de/ normalized, whereas in a ER diagram, our main aim is to remove redundancy
- ER model is designed to express microscopic relationships between elements. DM captures the business measures
- DM is designed to answer queries on business process, whereas the ER model is designed to record the business processes via their transactions.

Entity-Relationship vs. Dimensional Models

E- R DIAGRAM

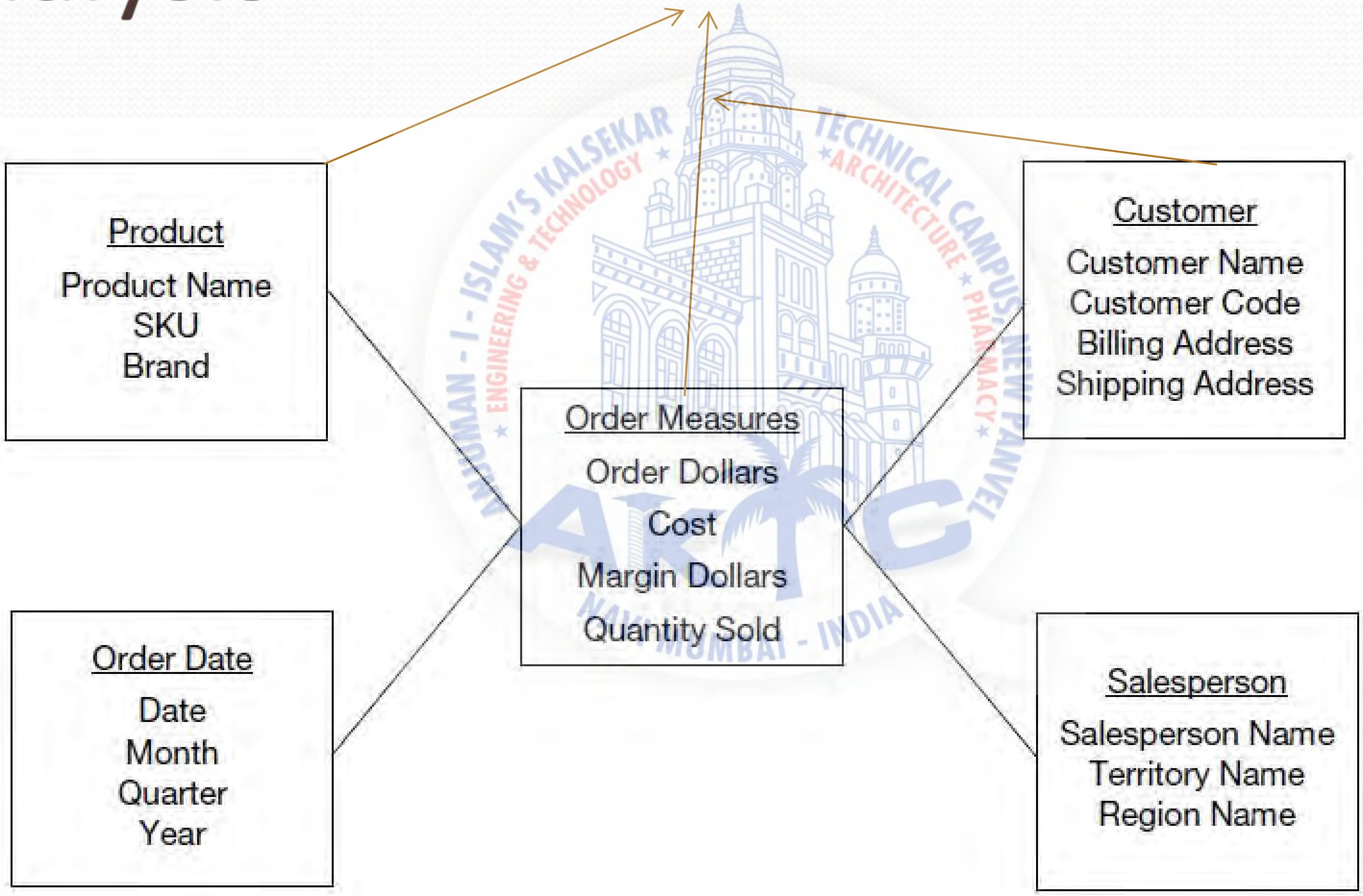
- One table per entity
- Minimize data redundancy
- Optimize update
- The Transaction Processing Model

DIMENSIONAL MODEL

- One fact table for data organization
- Maximize understandability
- Optimized for retrieval
- The data warehousing model

Star Schema-example of order analysis

Query result



Product
Product Name
SKU
Brand

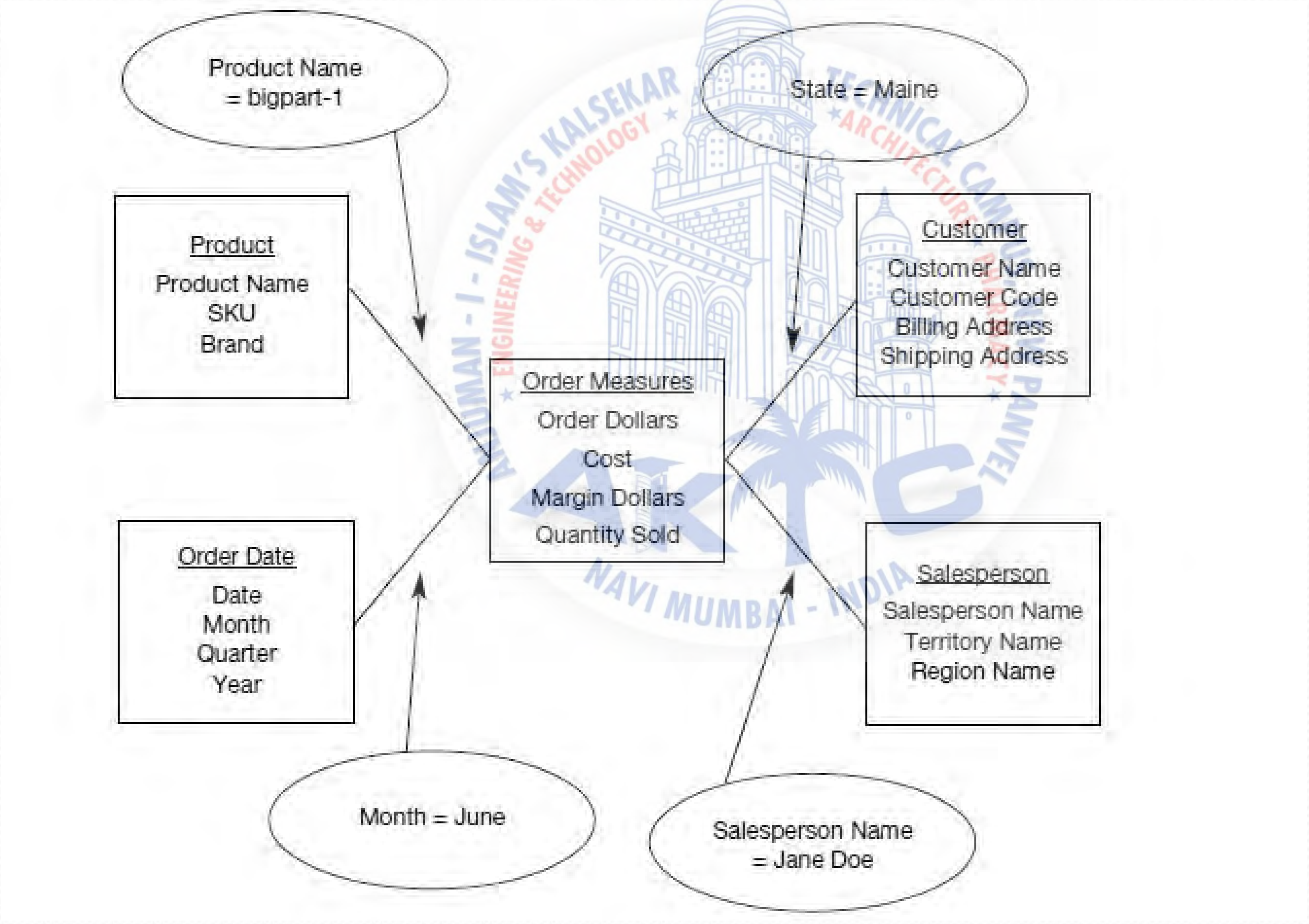
Customer
Customer Name
Customer Code
Billing Address
Shipping Address

Order Measures
Order Dollars
Cost
Margin Dollars
Quantity Sold

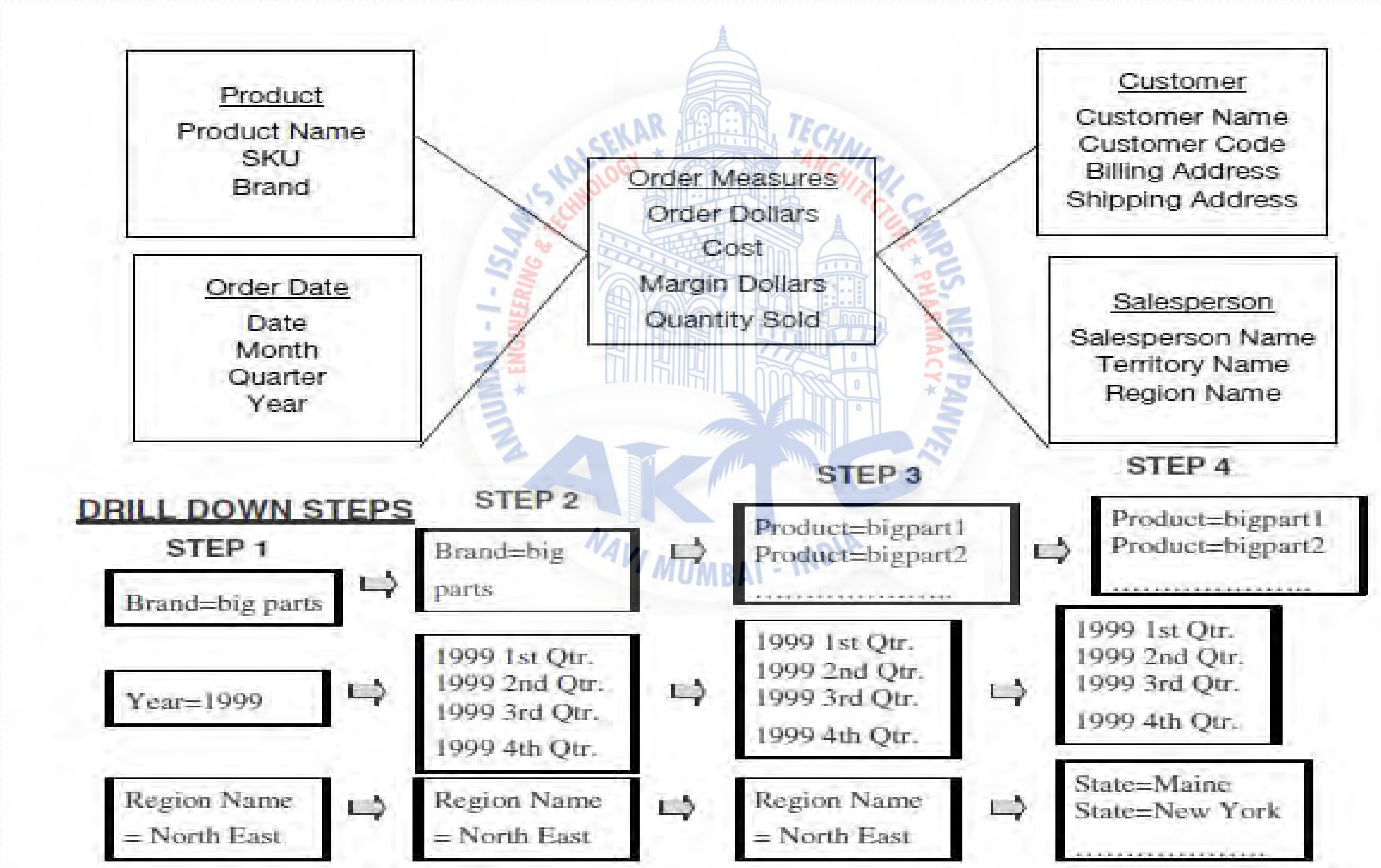
Order Date
Date
Month
Quarter
Year

Salesperson
Salesperson Name
Territory Name
Region Name

Understanding query from the star schema



Understanding drill down analysis from the star schema



Product
Product Name
SKU
Brand

Order Date
Date
Month
Quarter
Year

Order Measures
Order Dollars
Cost
Margin Dollars
Quantity Sold

Customer
Customer Name
Customer Code
Billing Address
Shipping Address

Salesperson
Salesperson Name
Territory Name
Region Name

DRILL DOWN STEPS

STEP 1	STEP 2	STEP 3	STEP 4

Brand=big parts	Brand=big parts	Product=bigpart1
		Product=bigpart2
		....
			Product=bigpart1
			Product=bigpart2
			....

Year=1999	1999 1st Qtr. 1999 2nd Qtr. 1999 3rd Qtr. 1999 4th Qtr.	1999 1st Qtr. 1999 2nd Qtr. 1999 3rd Qtr. 1999 4th Qtr.	1999 1st Qtr. 1999 2nd Qtr. 1999 3rd Qtr. 1999 4th Qtr.

Region Name = North East	Region Name = North East	Region Name = North East	State=Maine
			State=New York
			....

Dimension table

- Contain information about a particular dimension.
  - Dimension table key
  - Table is wide
  - Textual attributes
  - Attributes not directly related
  - Not normalized
  - Drilling down, rolling up
  - Multiple hierarchies
  - Fewer number of records

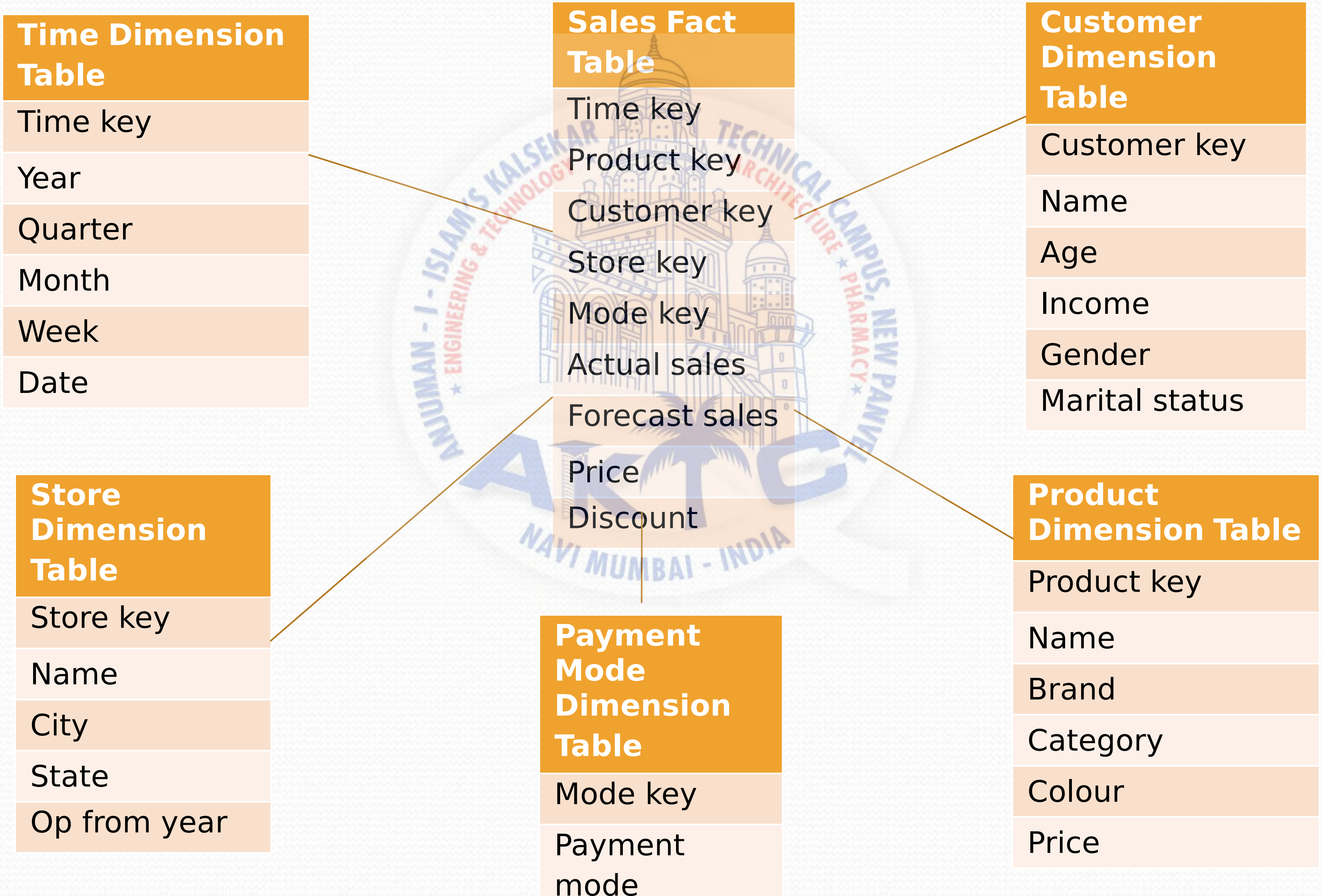
Facts

- Numeric measurements (values) that represent a specific business aspect or activity
- Stored in a fact table at the center of the star scheme
- Contains facts that are linked through their dimensions
- Can be computed or derived at run time
- Updated periodically with data from operational databases

Fact table

- Contains primary information of the warehouse
  - Concatenated key
  - Data grain
  - Fully additive measures
  - Semi/ additive measures(derived attributes)
  - Table deep, not wide
  - Sparse data
  - Degenerate dimensions(attributes which are neither fact or a dimension)

Star schema for a retail chain



Time Dimension Table
Time key
Year
Quarter
Month
Week
Date

Sales Fact Table
Time key
Product key
Customer key
Store key
Mode key
Actual sales
Forecast sales
Price
Discount

Customer Dimension Table
Customer key
Name
Age
Income
Gender
Marital status

Store Dimension Table
Store key
Name
City
State
Op from year

Payment Mode Dimension Table
Mode key
Payment mode

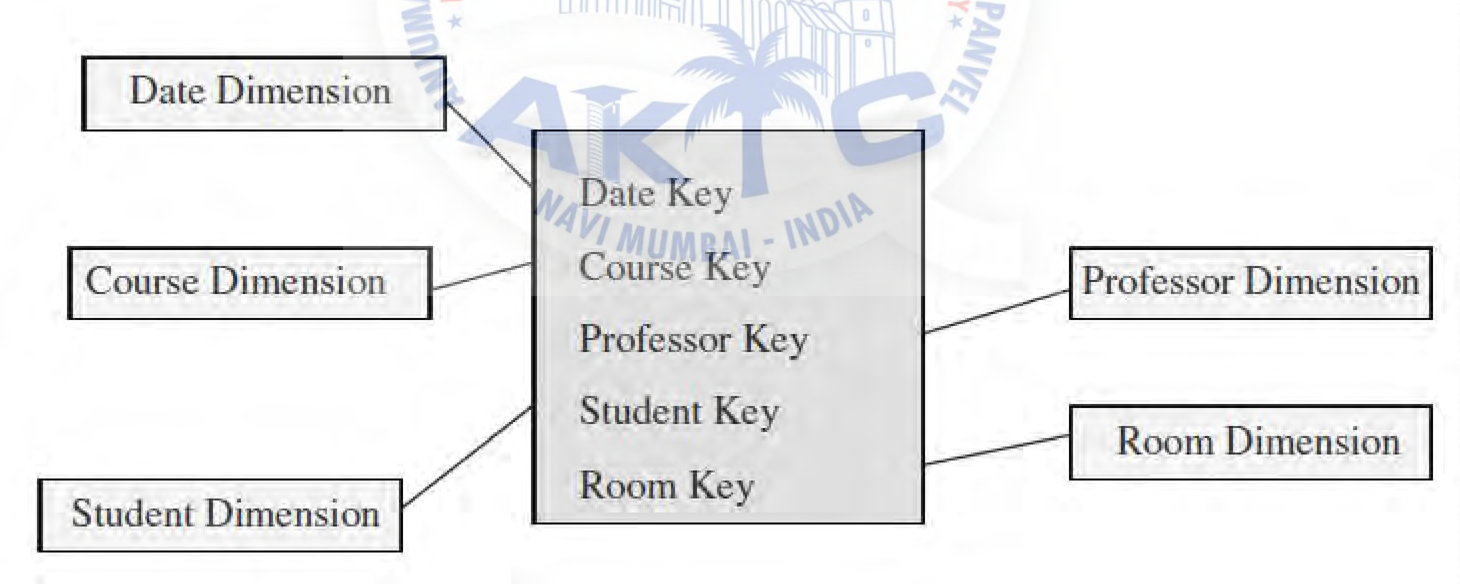
Product Dimension Table
Product key
Name
Brand
Category
Colour
Price

Star Schema characteristics

- Star schema is a relational model with one/ to/ many relationship between the fact table and the dimension tables.
- De/ normalized relational model
- Easy to understand. Reflects how users think. This makes it easy for them to query and analyse the data.
- Optimizes navigation.
- Enhances query extraction.
- Ability to drill down or roll up.

Factless fact table

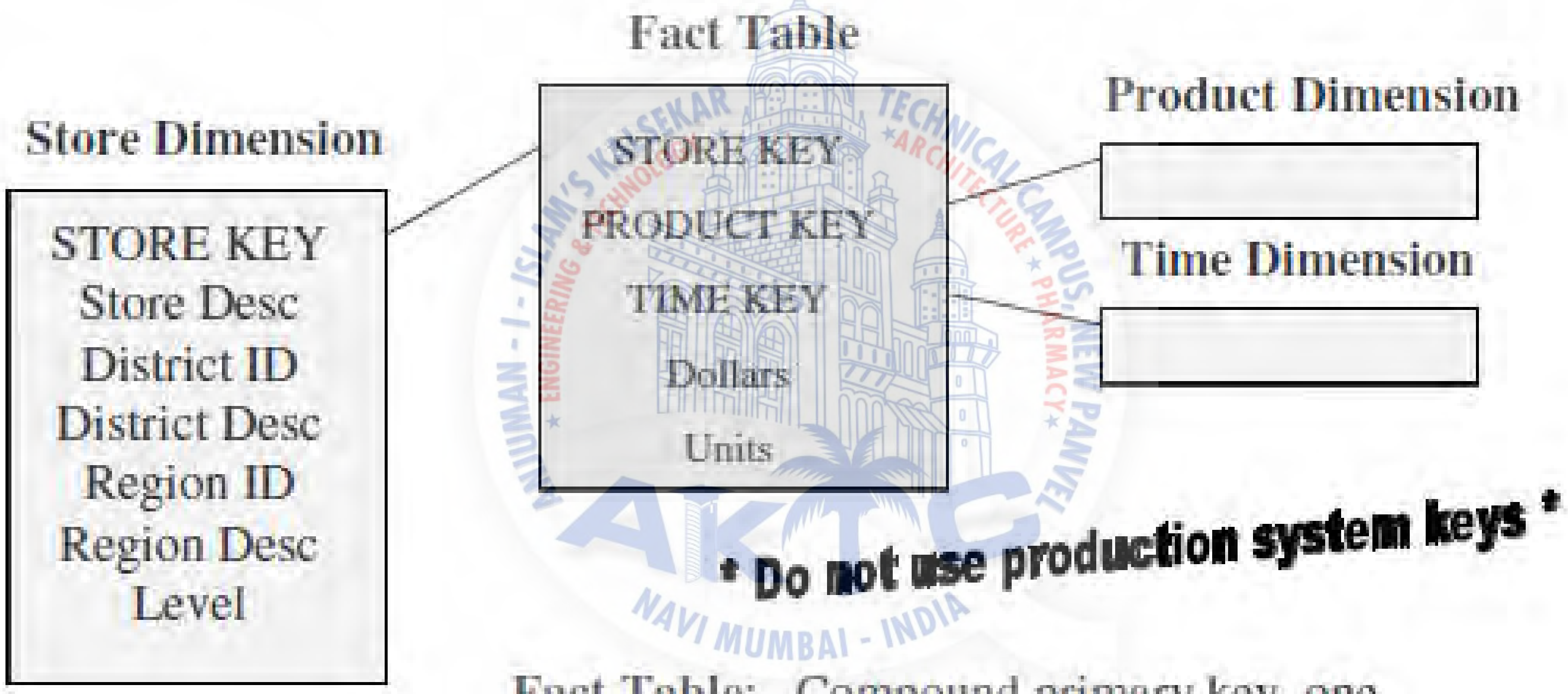
- A fact table is said to be empty if it has no measures to be displayed. Fact table represents **events**
- Contains no data, only keys.



Data Granularity

- When fact table at the lowest grain, the users can as well drill down to the lowest grain of details
- But when data is kept till the lowest level of data, we have to compromise on the storage and maintenance of DW
- Advantages
  - Easier to extract from operational data and load into DW
  - Can be feed directly to the DM application

Star Scheme Keys



Fact Table: Compound primary key, one segment for each dimension

Dimension Table: Generated primary key

Star schema keys contd…

- Primary keys: should not be same as production system
- Surrogate keys: System generated sequence numbers having no built/ in meanings
- Foreign keys: primary key of each dimension table must be a foreign key in the fact table.

Primary key for Fact table

- A single compound primary key whose length is the total length of the keys of the individual dimension tables
- **Concatenated primary key that is the concatenation of all the primary keys of the dimension tables,**
- A generated primary key independent of the keys of the dimension tables.

Advantages of the star schema

- Easy for users to understand
- Optimizes navigation
- Most suitable for query processing

Starjoin and Starindex

- Star join:/ high/ speed, single pass parallelizable, multitable join.
    - Boots query performance
- Star index:/ specialized index to accelerate join performance
    - Speed up joins between the dimension tables and fact tables

Summing up

- Derived from the information packages in the requirements definition.
- The STAR schema used for data design is a relational model consisting of fact and dimension tables.
- The fact table contains the business metrics or measurements; the dimensional tables contain the business dimensions. Hierarchies within each dimension table are used for drilling down to lower levels of data.
- STAR schema advantages are: easy for users to understand, optimizes navigation, most suitable for query processing, and enables specific

Objectives

- Slowly changing dimensions
- Large dimensions
- Snowflake schema
- Aggregate tables
- Family of starts and their applications

Updating the Dimension table

- Dimension tables are non/ volatile and mostly read/ only.
- More rows are added to the Dimension tables over time.
- Changes to certain attributes of a row become eminent at times.
- There are many types of changes that affect the dimension tables.

Slowly changing dimensions

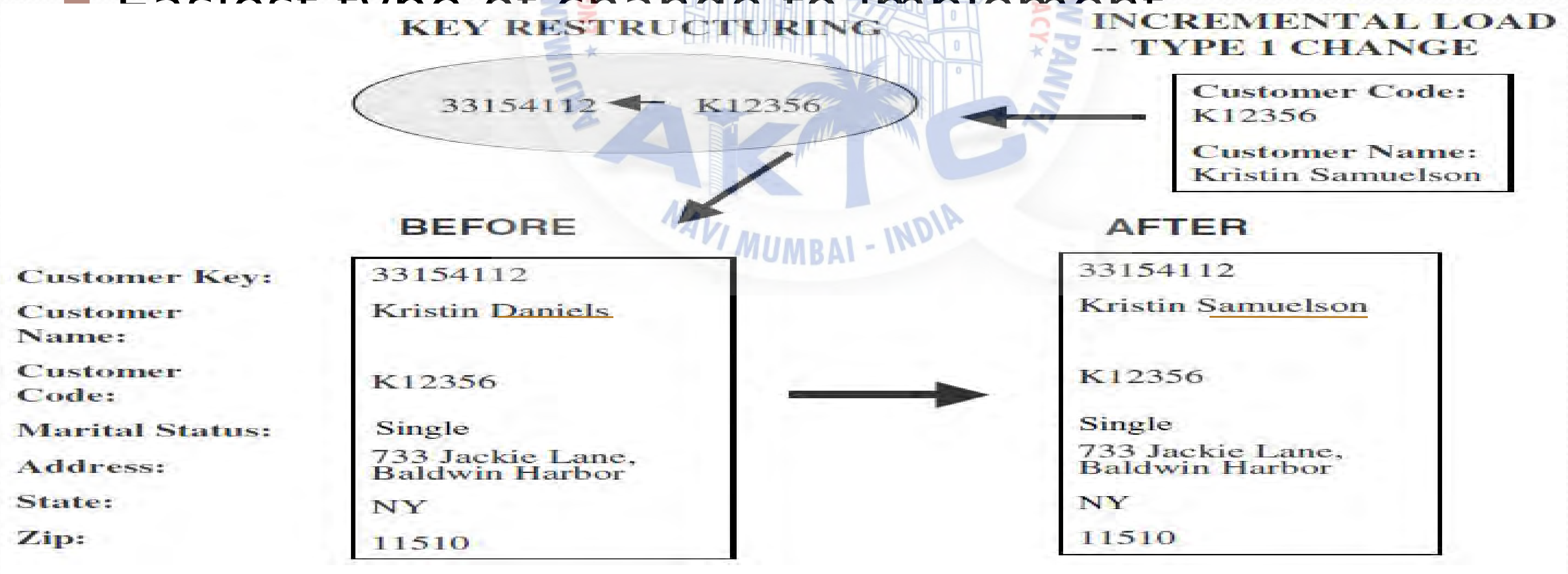
- Most dimensions are generally constant over time
- Many dimensions change slowly
- Though the key does not change other description and attributes change slowly over time
- Dimension table attributes are not overwritten
- The ways changes are made in dimension tables depend on the types of changes and what information must be preserved.

Type 1: Correction of errors

- Usually relate to correction of errors in the source systems.
- E.g., spelling error in customer names; change of names of customers;
- There is no need to preserve the old values here.
- The old value in the source system needs to be discarded.
- The changes made need not be preserved or noted.

Type 1.. continued

- Overwrite attribute value in the dimension table row with new value
- No other changes are made to the dimension table row.
- The key is not disturbed
- Easiest type of change to implement



KEY RESTRUCTURING

33154112 ← K12356

INCREMENTAL LOAD -- TYPE 1 CHANGE

Customer Code: K12356

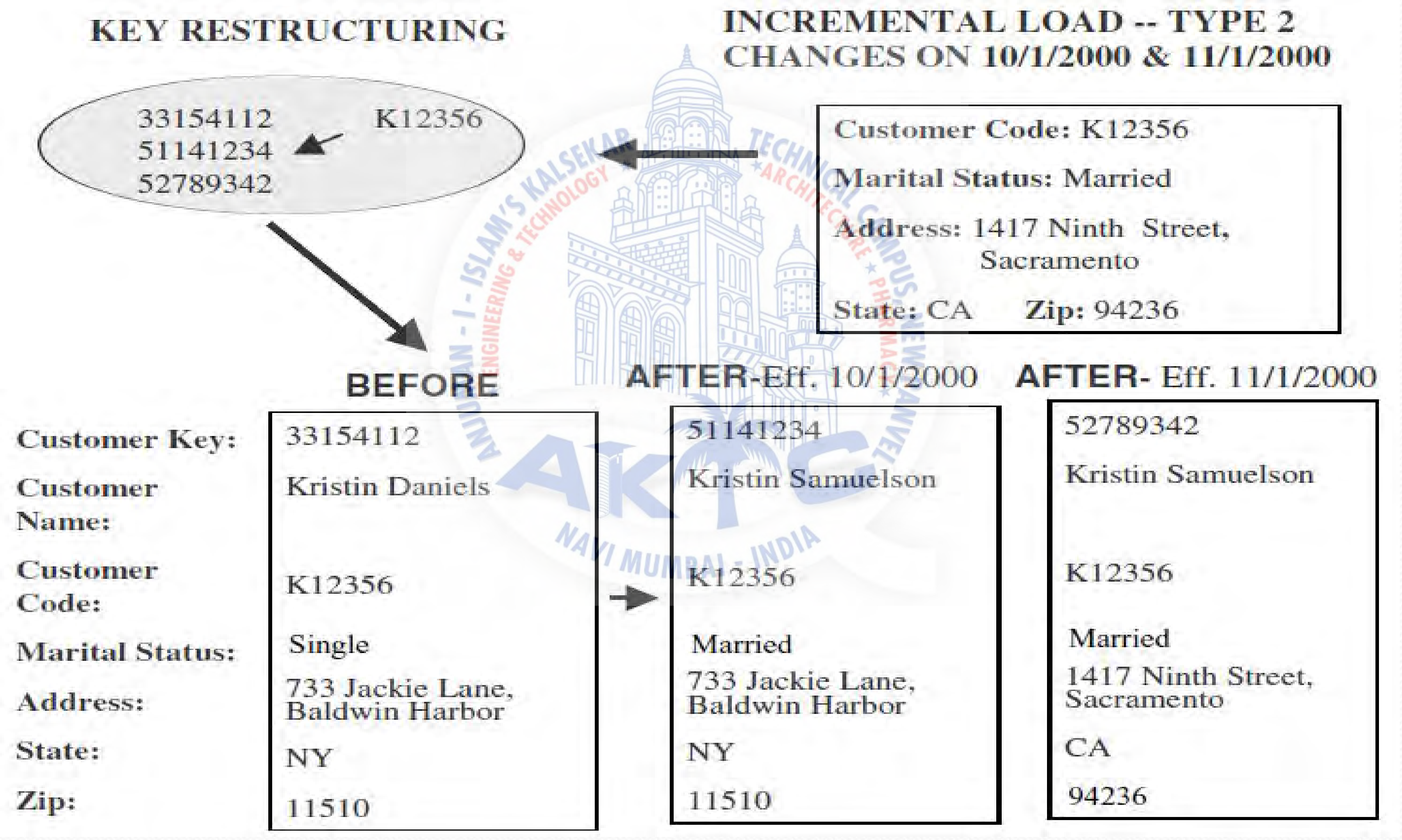
Customer Name: Kristin Samuelson

	BEFORE	AFTER
Customer Key:	33154112	33154112
Customer Name:	Kristin Daniels	Kristin Samuelson
Customer Code:	K12356	K12356
Marital Status:	Single	Single
Address:	733 Jackie Lane, Baldwin Harbor	733 Jackie Lane, Baldwin Harbor
State:	NY	NY
Zip:	11510	11510

Type 2: preservation of history

- True changes in the source systems.
- E.g., change of marital status; change of address
- There is a need to preserve history
- This type of changes partition the warehouse
- Every change for the same attribute must be preserved.
- Applying these changes:
  - Add a new dimension table row with new value of the changed attribute
  - No changes are made to the existing row.
  - New rows are inserted with a new surrogate key.

Type 2.. continued



KEY RESTRUCTURING

33154112
51141234
52789342

K12356

INCREMENTAL LOAD -- TYPE 2
CHANGES ON 10/1/2000 & 11/1/2000

Customer Code: K12356

Marital Status: Married

Address: 1417 Ninth Street,
Sacramento

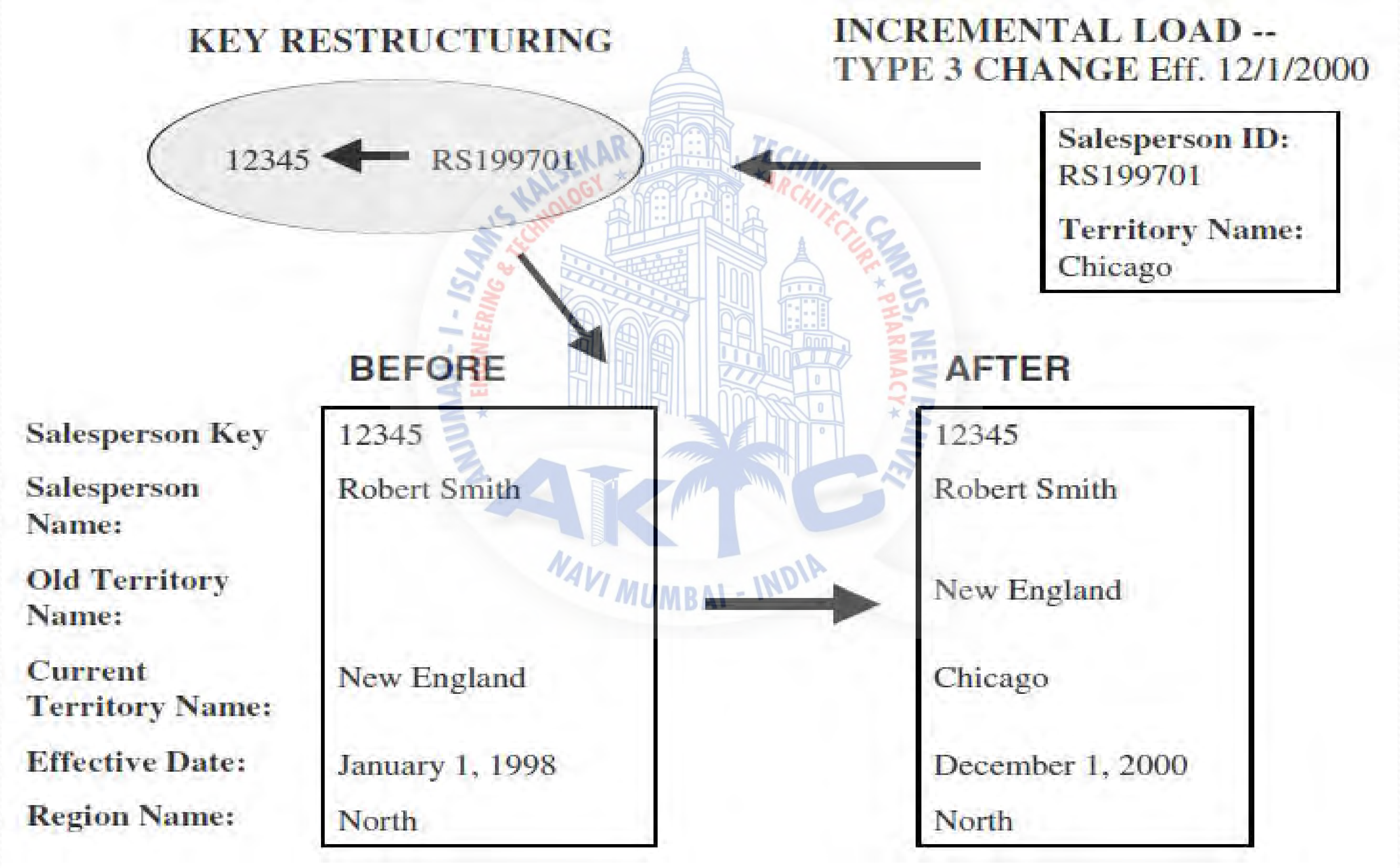
State: CA Zip: 94236

	BEFORE	AFTER-Eff. 10/1/2000	AFTER- Eff. 11/1/2000
Customer Key:	33154112	51141234	52789342
Customer Name:	Kristin Daniels	Kristin Samuelson	Kristin Samuelson
Customer Code:	K12356	K12356	K12356
Marital Status:	Single	Married	Married
Address:	733 Jackie Lane, Baldwin Harbor	733 Jackie Lane, Baldwin Harbor	1417 Ninth Street, Sacramento
State:	NY	NY	CA
Zip:	11510	11510	94236

Type 3: tentative soft revision

- Tentative changes in the source system
- E.g., if an employee will get posted for a short period to a different location
- Need to keep track of history with old and new values
- Used to compare performances across the transition
- Applying these changes
  - An "old" field is added in the dimension table
  - Push existing value of attribute from "current" to "old"

Type 3.. continued



KEY RESTRUCTURING

12345 ← RS199701

INCREMENTAL LOAD -- TYPE 3 CHANGE Eff. 12/1/2000

Salesperson ID: RS199701

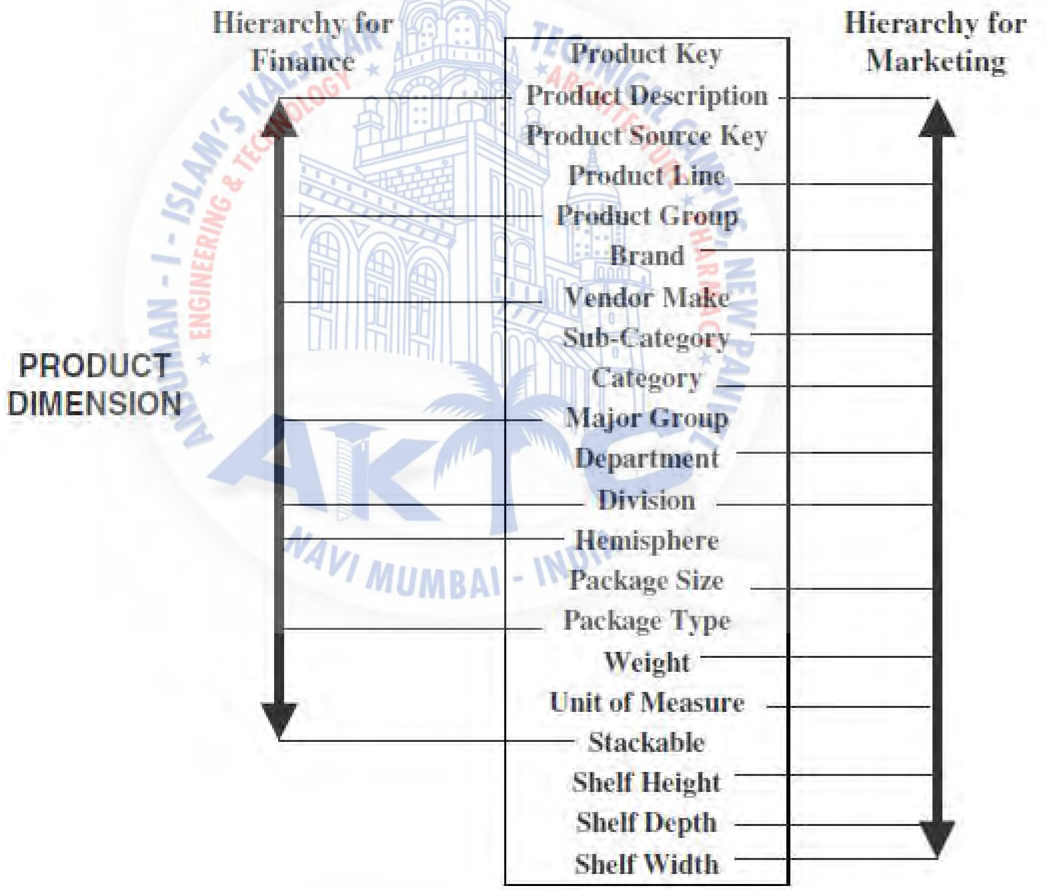
Territory Name: Chicago

	BEFORE	AFTER
Salesperson Key	12345	12345
Salesperson Name:	Robert Smith	Robert Smith
Old Territory Name:		New England
Current Territory Name:	New England	Chicago
Effective Date:	January 1, 1998	December 1, 2000
Region Name:	North	North

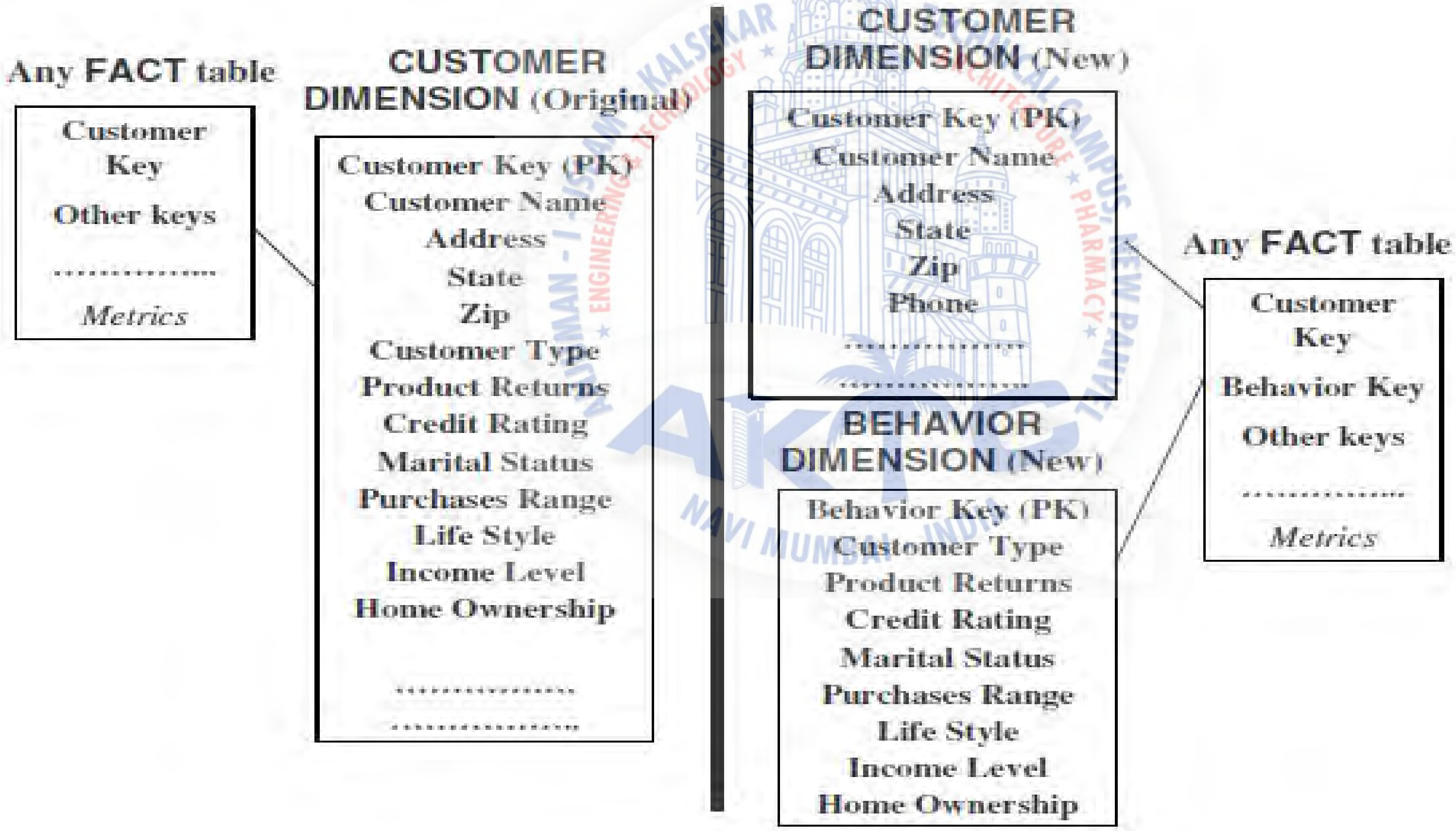
Large dimensions

- Very deep(large number of rows)
- Very wide(large number of attributes)
- Have multiple hierarchies
- Rapidly changing dimensions
- Junk dimensions

Multiple hierarchies



Rapidly changing dimensions



Any FACT table

Customer Key
Other keys
....
Metrics

CUSTOMER DIMENSION (Original)

Customer Key (PK)
Customer Name
Address
State
Zip
Customer Type
Product Returns
Credit Rating
Marital Status
Purchases Range
Life Style
Income Level
Home Ownership
....
....

CUSTOMER DIMENSION (New)

Customer Key (PK)
Customer Name
Address
State
Zip
Phone
....
....

BEHAVIOR DIMENSION (New)

Behavior Key (PK)
Customer Type
Product Returns
Credit Rating
Marital Status
Purchases Range
Life Style
Income Level
Home Ownership

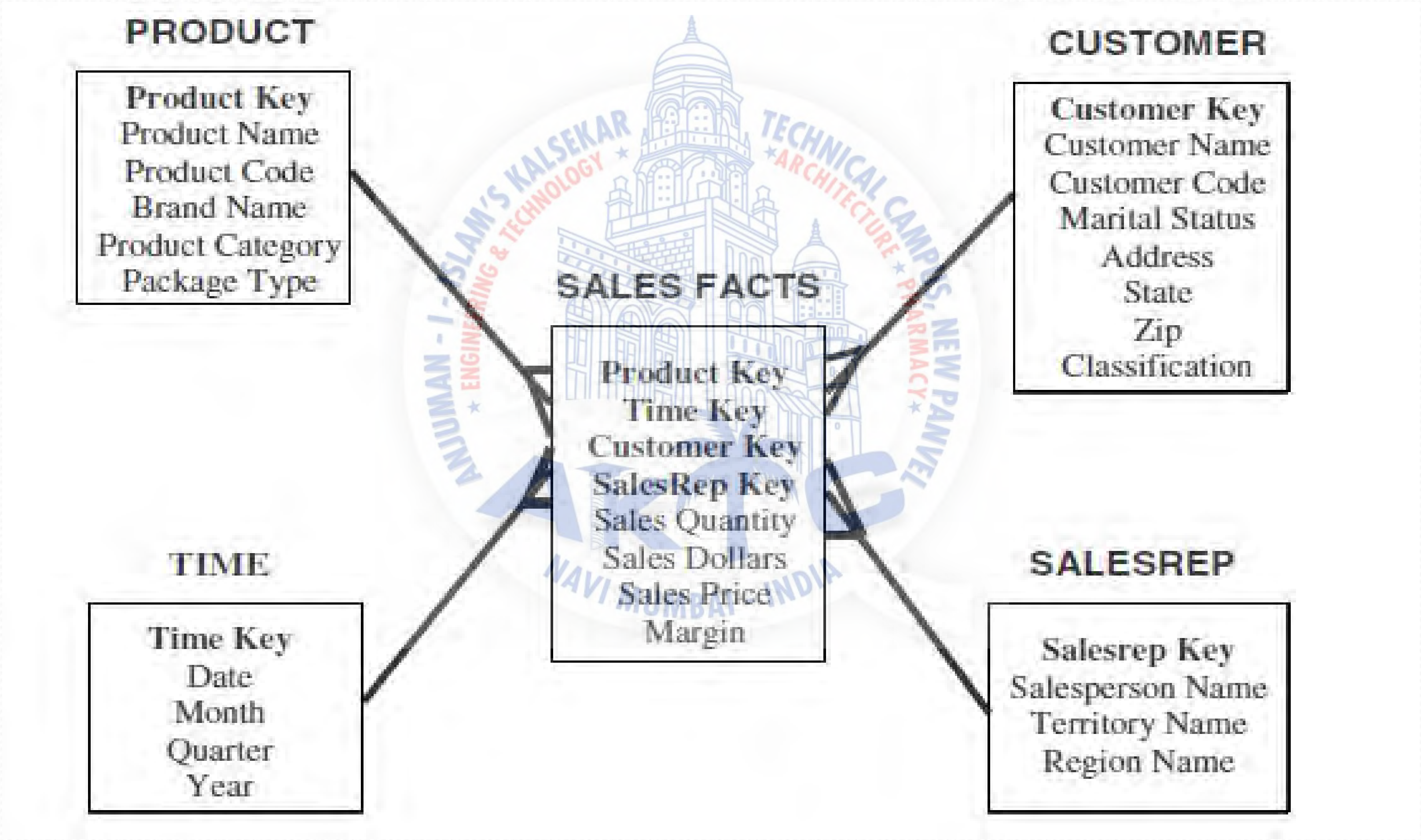
Any FACT table

Customer Key
Behavior Key
Other keys
....
Metrics

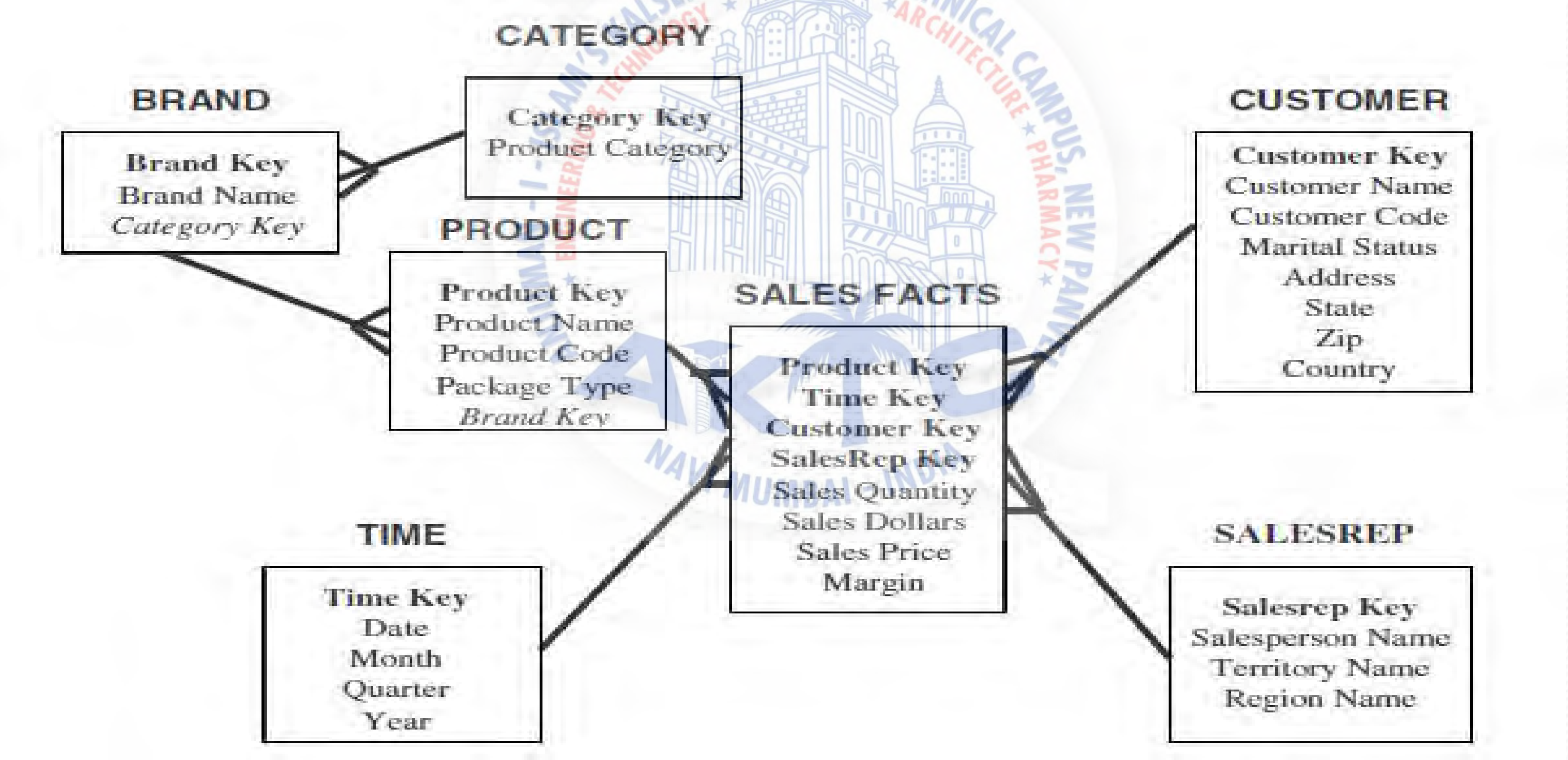
Snowflake schema

- A variation of the star schema, in which all or some of the dimension tables may be normalized.
- Eliminates redundancy
- Generally used when a dimension table is wide.
- Saves space
- Complex querying is required.

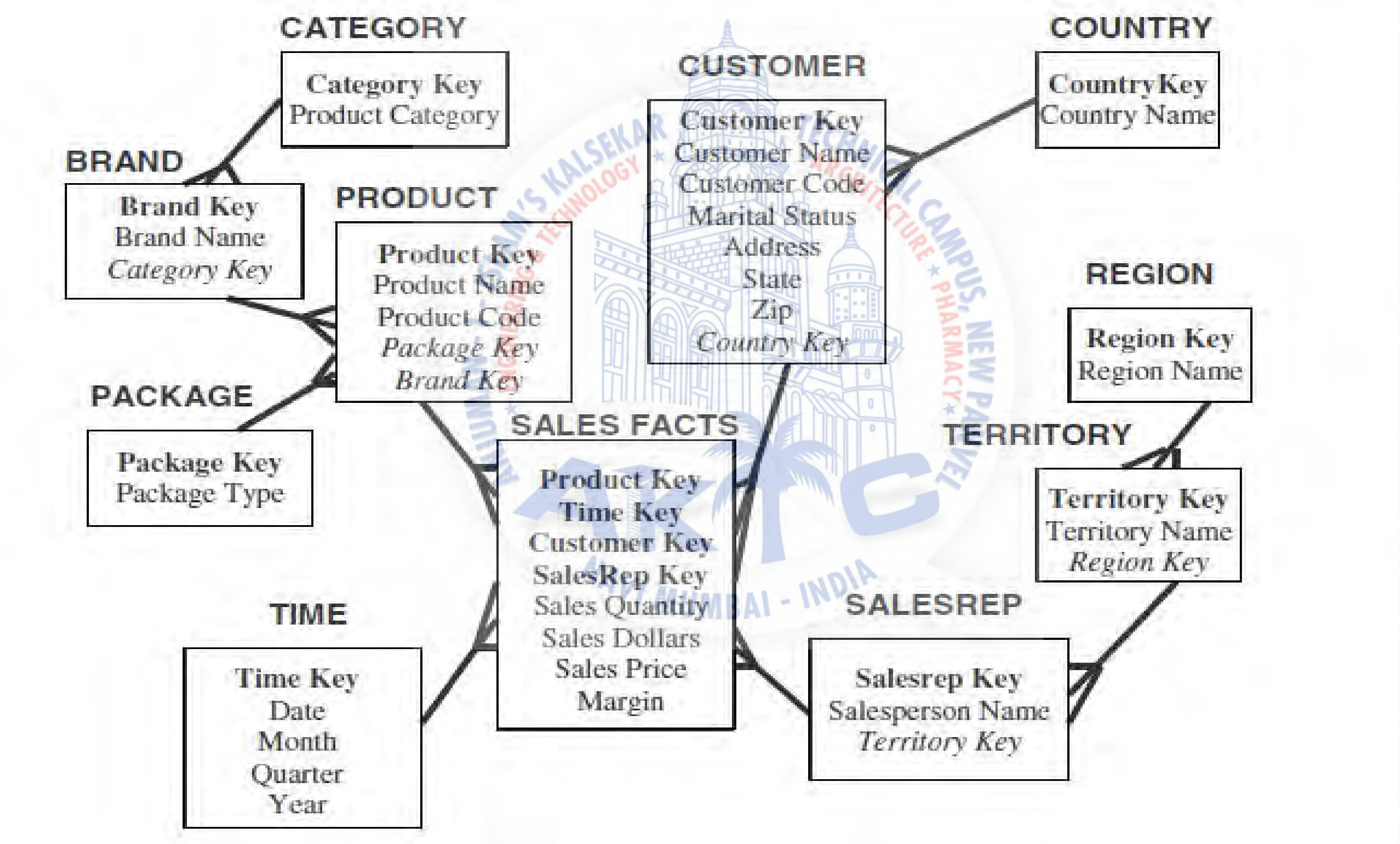
Star schema for sales



Normalized product dimension



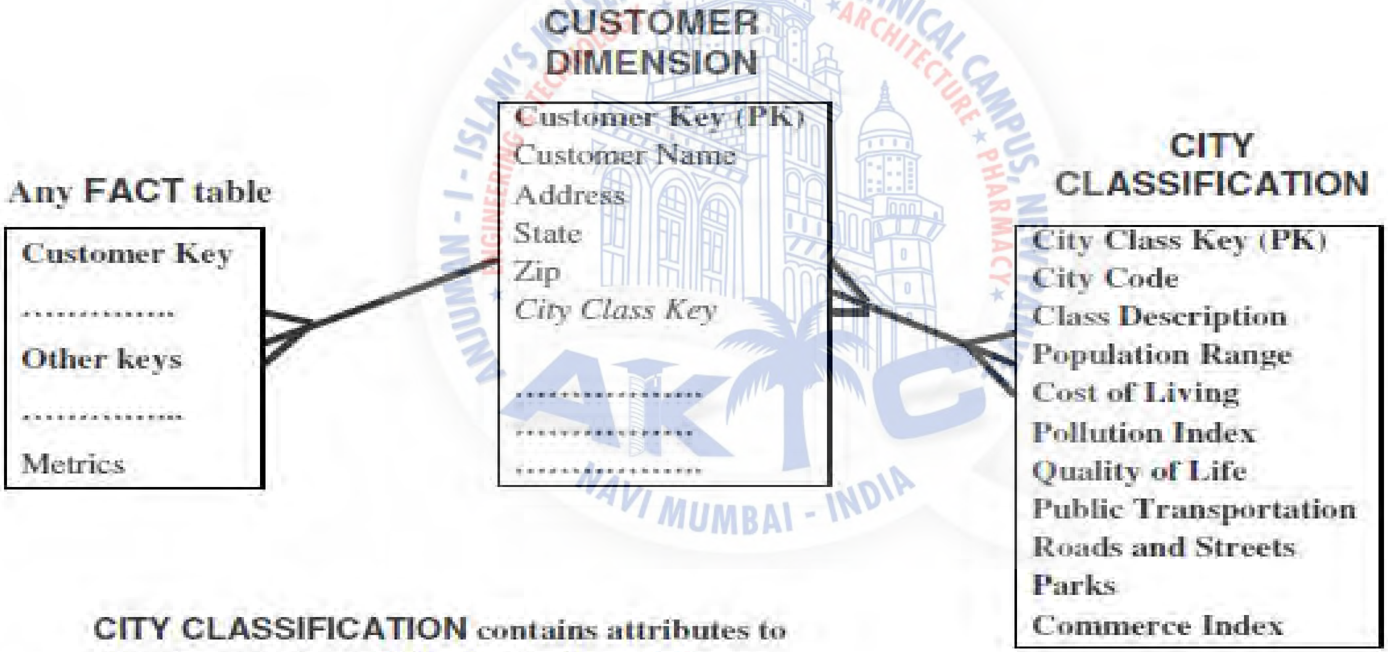
Sales snowflake schema



Advantages and disadvantages

- Advantages
  - Small savings in storage space
  - Normalized structures are easier to update and maintain
- Disadvantages
  - Schema is less intuitive
  - Browsing becomes difficult
  - Degraded query performance because of additional joins

When to snowflake



Any FACT table

- Customer Key
- ....
- Other keys
- ....
- Metrics

CUSTOMER DIMENSION

- Customer Key (PK)
- Customer Name
- Address
- State
- Zip
- *City Class Key*
- ....
- ....
- ....

CITY CLASSIFICATION

- City Class Key (PK)
- City Code
- Class Description
- Population Range
- Cost of Living
- Pollution Index
- Quality of Life
- Public Transportation
- Roads and Streets
- Parks
- Commerce Index

CITY CLASSIFICATION contains attributes to classify each city within a limited set of classes. These attributes are separated from the CUSTOMER DIMENSION to form a separate sub-dimension as CITY CLASSIFICATION.

Aggregate fact tables

- Contain pre/ calculated summaries derived from the most granular (detailed) fact table.
- Created as a specific summarization across any number of dimensions.
- Reduces runtime processing.

Why need aggregate fact tables?

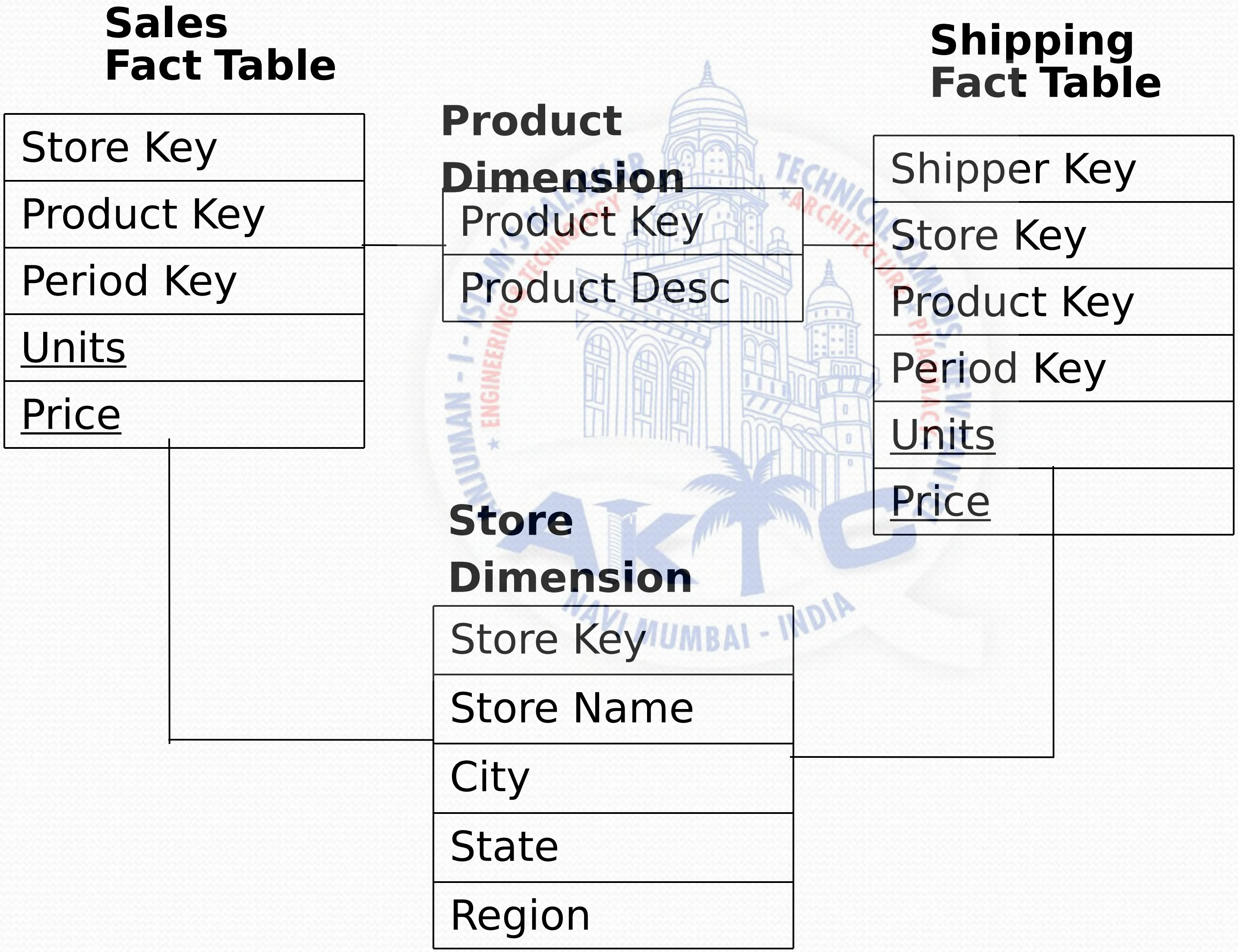
- Large size of the fact table
- To speed up query extraction

- Limitations
  - Must be re/ aggregated each time there is a change in the source data
  - Do not support exploratory analysis
  - Limited interactive use.

Fact Constellation

- Multiple fact tables share dimension tables.
- This schema is viewed as collection of stars hence called galaxy schema or fact constellation.
- Sophisticated application requires such schema.

Fact Constellation (contd..)



Sales Fact Table

Store Key
Product Key
Period Key
Units
Price

Product Dimension

Product Key
Product Desc

Shipping Fact Table

Shipper Key
Store Key
Product Key
Period Key
Units
Price

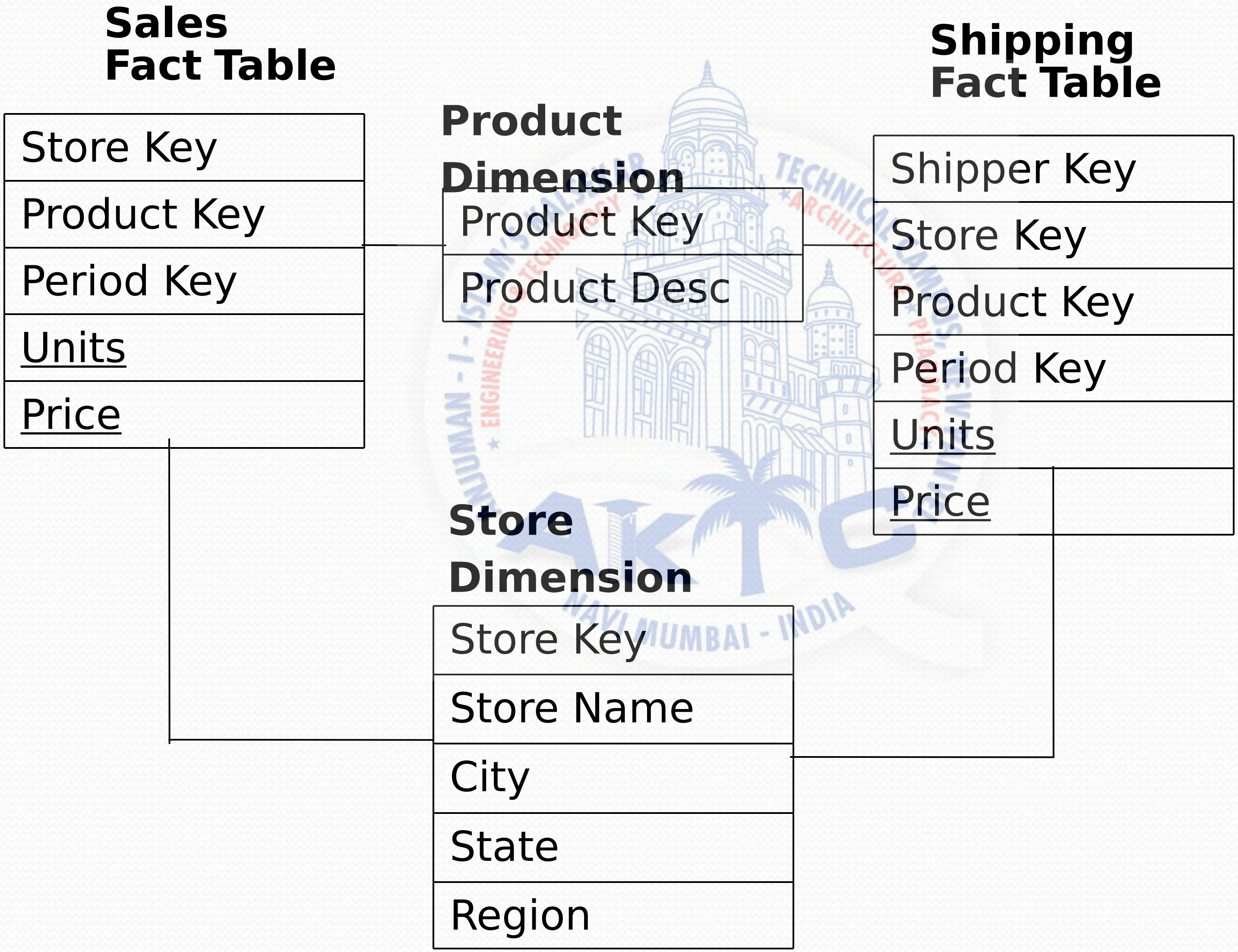
Store Dimension

Store Key
Store Name
City
State
Region

Fact Constellation

- Multiple fact tables share dimension tables.
- This schema is viewed as collection of stars hence called galaxy schema or fact constellation.
- Sophisticated application requires such schema.

Fact Constellation (contd..)



Sales Fact Table

Store Key
Product Key
Period Key
Units
Price

Product Dimension

Product Key
Product Desc

Shipping Fact Table

Shipper Key
Store Key
Product Key
Period Key
Units
Price

Store Dimension

Store Key
Store Name
City
State
Region