Introduction to Data Warehousing

Ms Swapnil Shrivastava
swapnil@konark.ncst.ernet.in

Necessity is the mother of invention

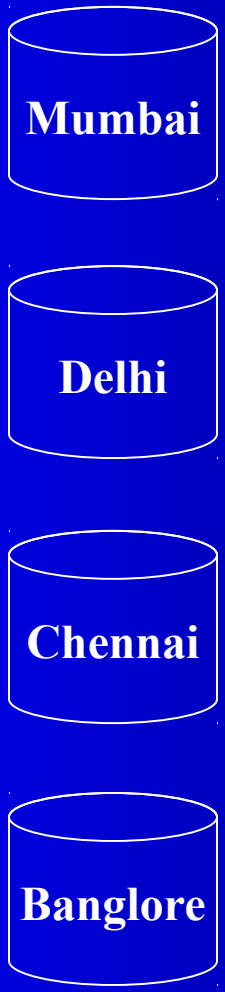
Why Data Warehouse?

Scenario 1

ABC Pvt Ltd is a company with branches at Mumbai, Delhi, Chennai and Banglore. The Sales Manager wants quarterly sales report.

Each branch has a separate operational system.

Scenario 1 : ABC Pvt Ltd.



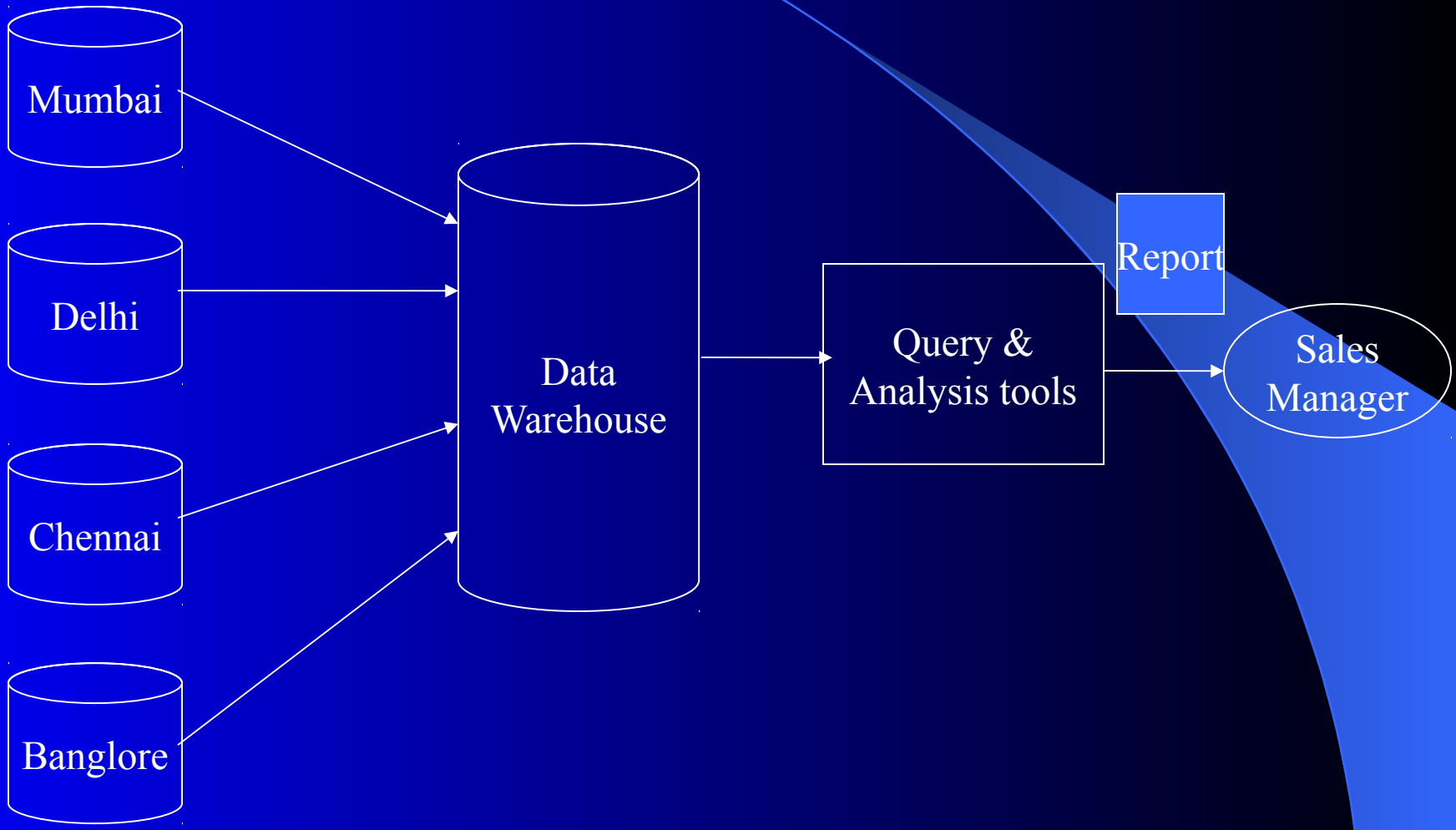
Sales per item type per branch for first quarter.

Sales Manager

Solution 1:ABC Pvt Ltd.

- Extract sales information from each database.
- Store the information in a common repository at a single site.

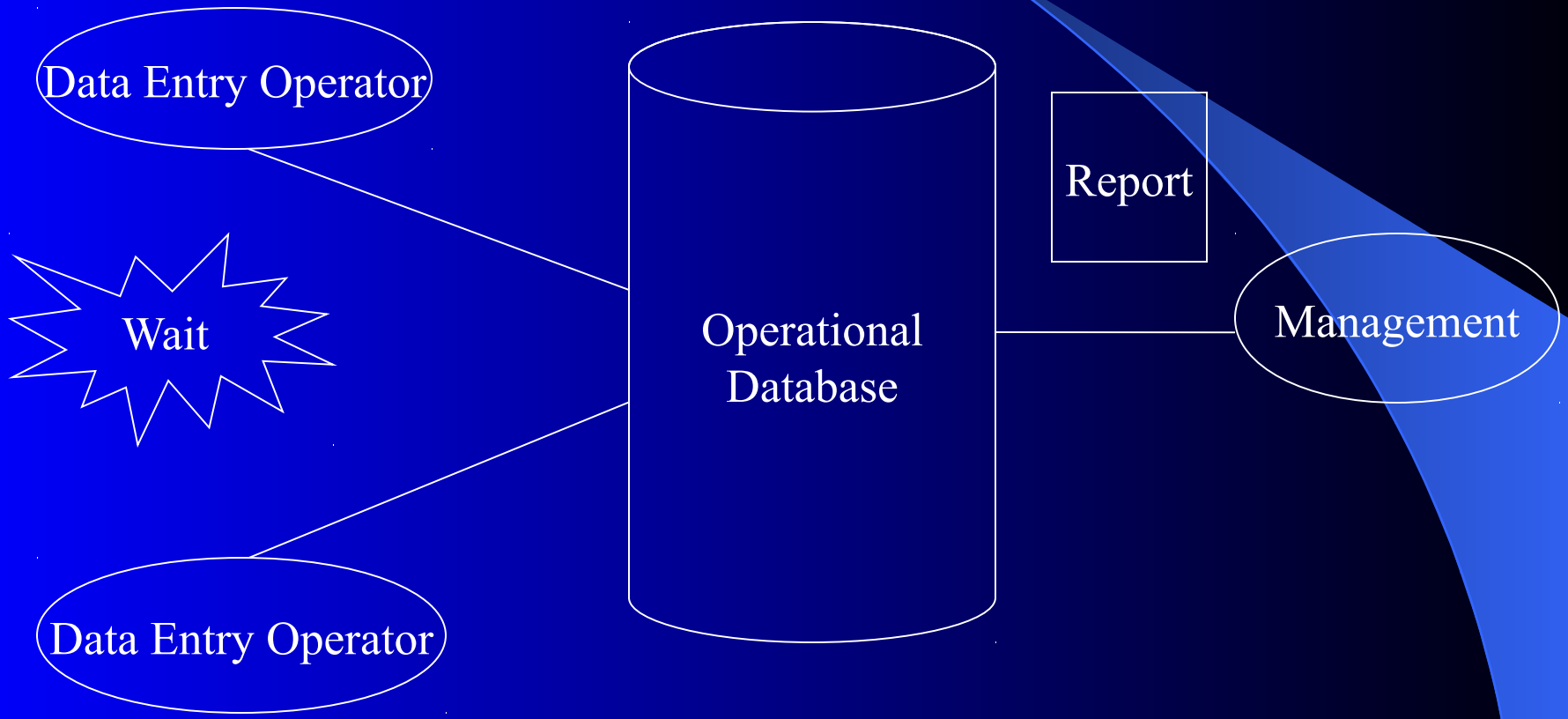
Solution 1:ABC Pvt Ltd.



Scenario 2

One Stop Shopping Super Market has huge operational database.Whenever Executives wants some report the OLTP system becomes slow and data entry operators have to wait for some time.

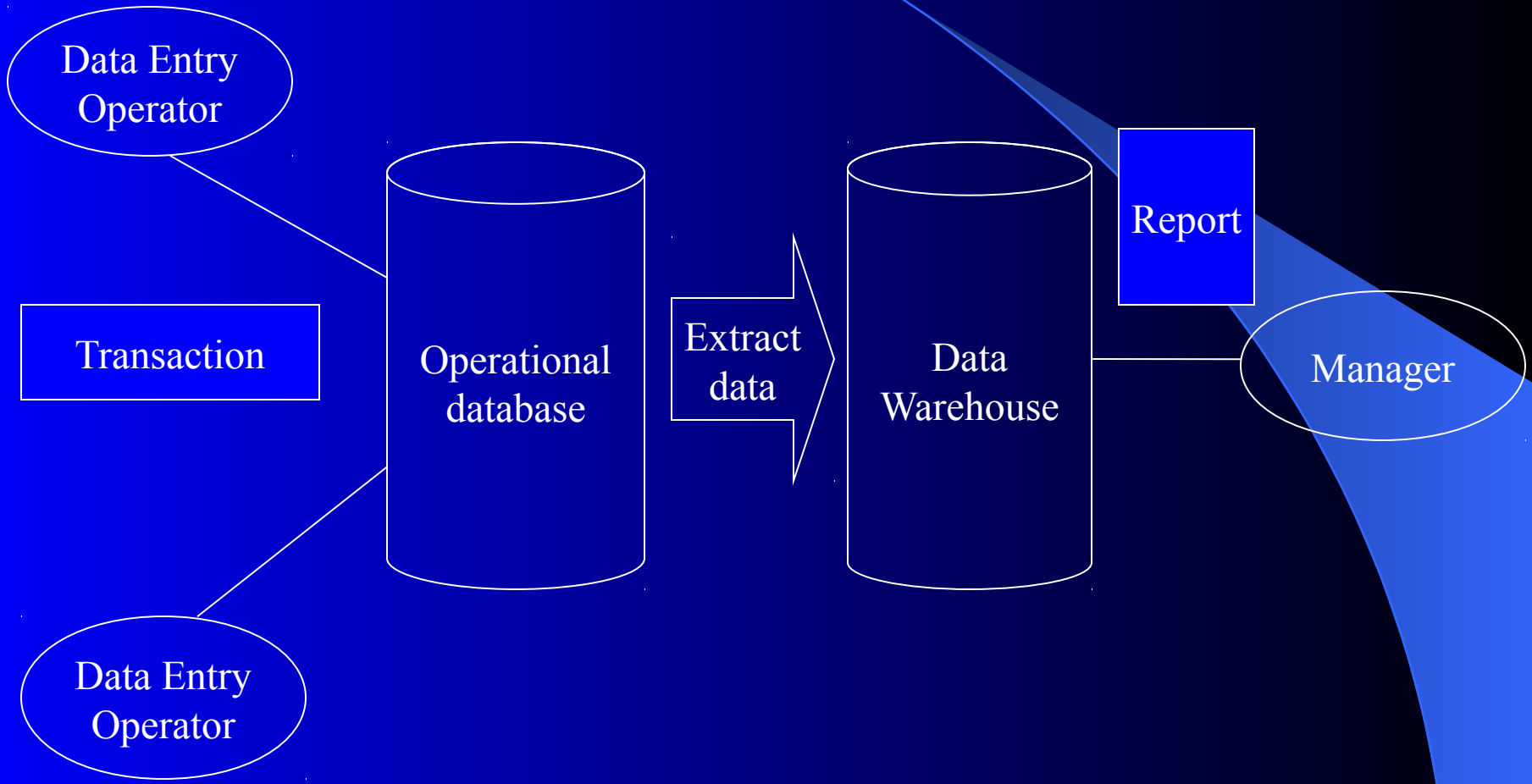
Scenario 2 : One Stop Shopping



Solution 2

- Extract data needed for analysis from operational database.
- Store it in warehouse.
- Refresh warehouse at regular interval so that it contains up to date information for analysis.
- Warehouse will contain data with historical perspective.

Solution 2


Data Entry Operator
Transaction
Data Entry Operator
Operational database
Extract data
Data Warehouse
Report
Manager

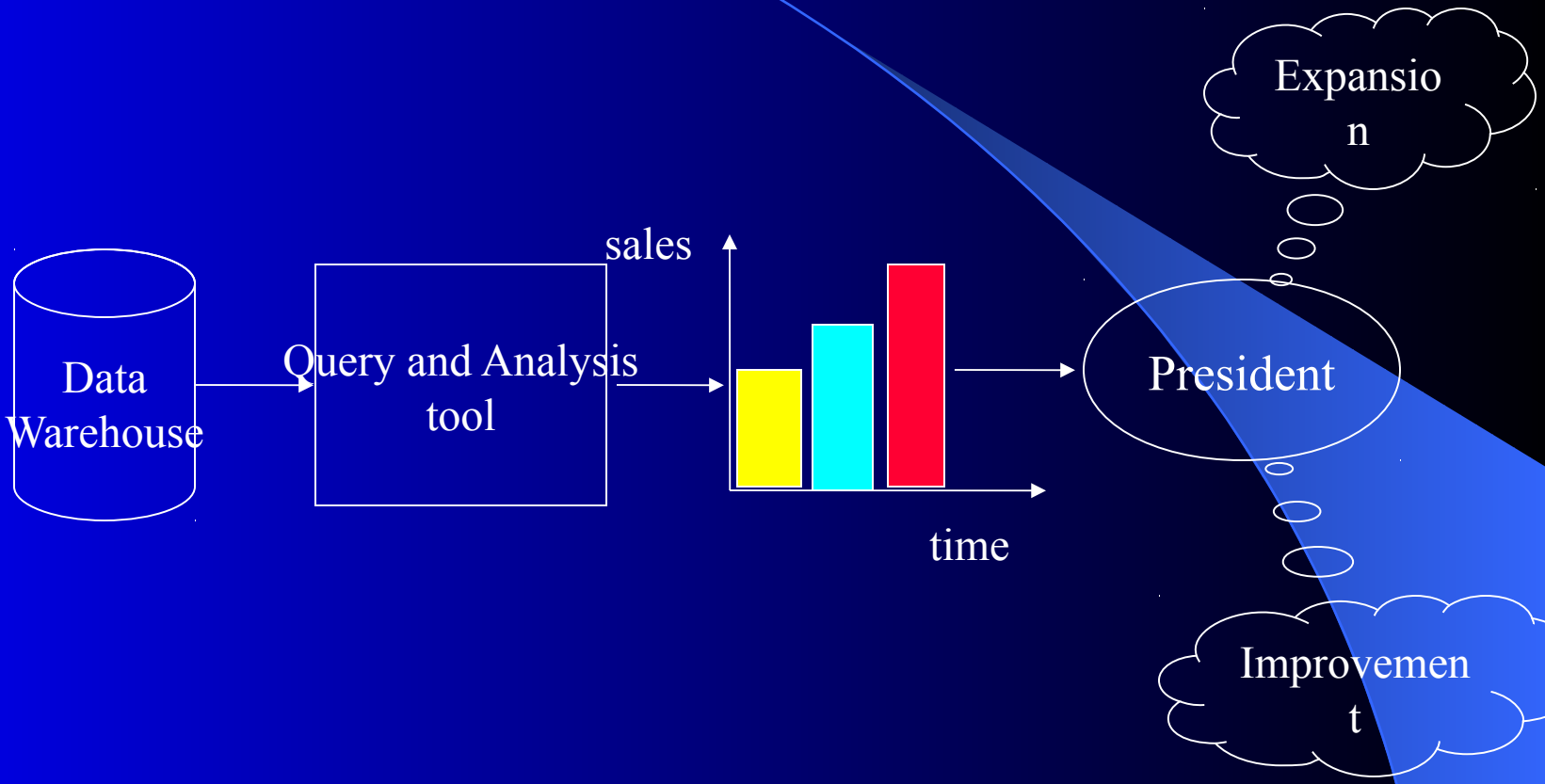
Scenario 3

Cakes & Cookies is a small,new company.President of the company wants his company should grow.He needs information so that he can make correct decisions.

Solution 3

- Improve the quality of data before loading it into the warehouse.
- Perform data cleaning and transformation before loading the data.
- Use query analysis tools to support adhoc queries.

Solution 3


Expansio
n
sales
Data
Warehouse
Query and Analysis
tool
time
President
Improvemen
t

What is Data Warehouse??

Inmons's definition

A data warehouse is

-subject-oriented,

-integrated,

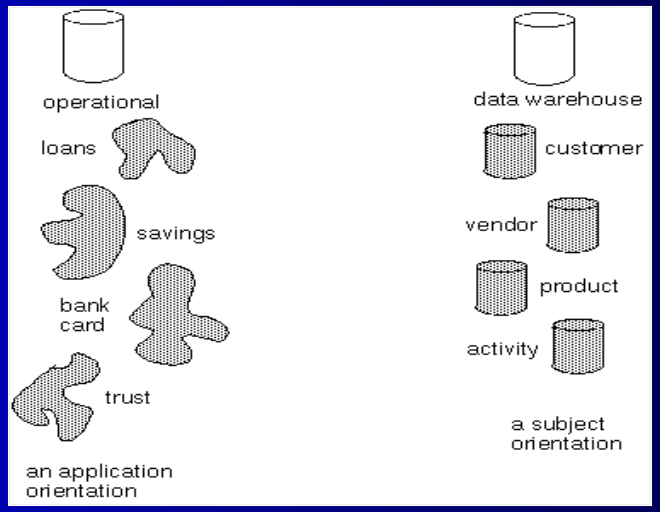
-time-variant,

-nonvolatile

collection of data in support of management's decision making process.

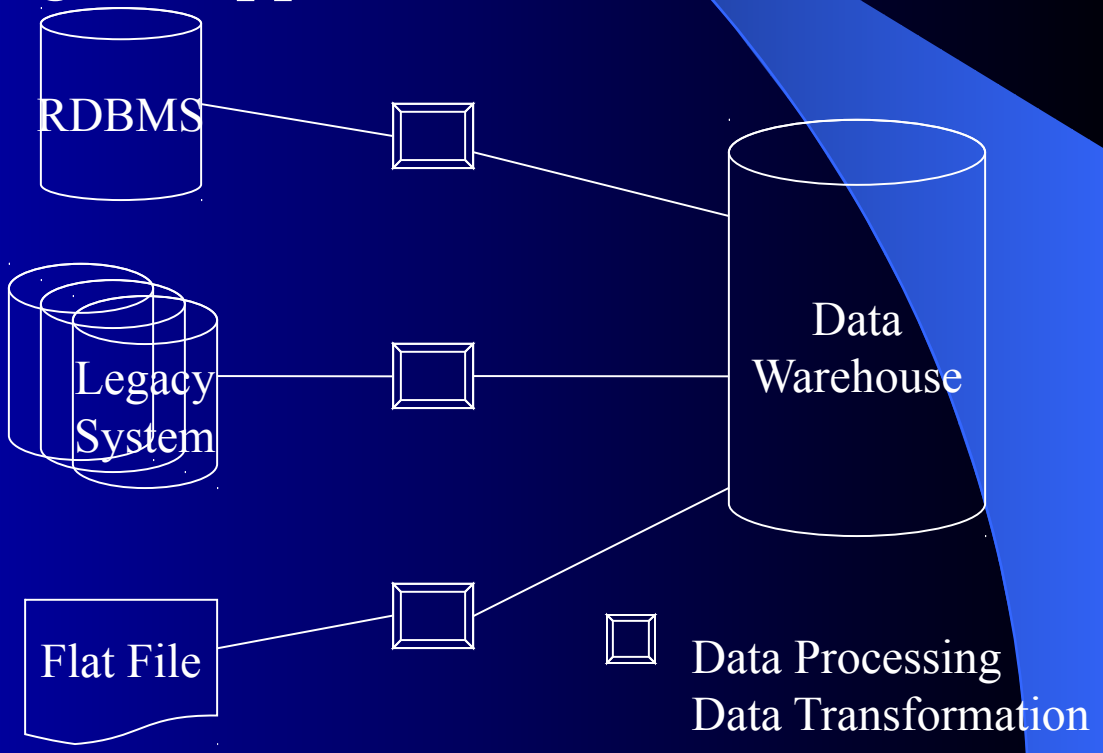
Subject-oriented

- Data warehouse is organized around subjects such as sales,product,customer.
- It focuses on modeling and analysis of data for decision makers.
- Excludes data not useful in decision support process.



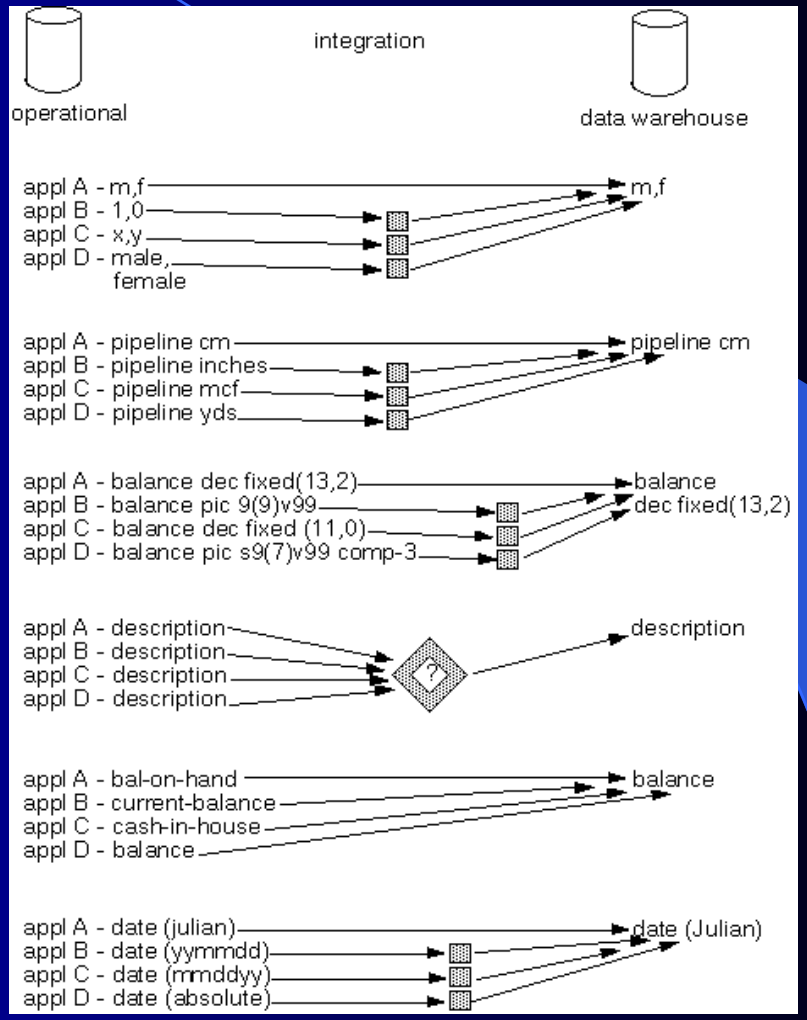
Integration

- Data Warehouse is constructed by integrating multiple heterogeneous sources.
- Data Preprocessing are applied to ensure consistency.



Integration

- In terms of data.
  - encoding structures.
  - Measurement of attributes.
  - physical attribute. of data
  - naming conventions.
  - Data type format

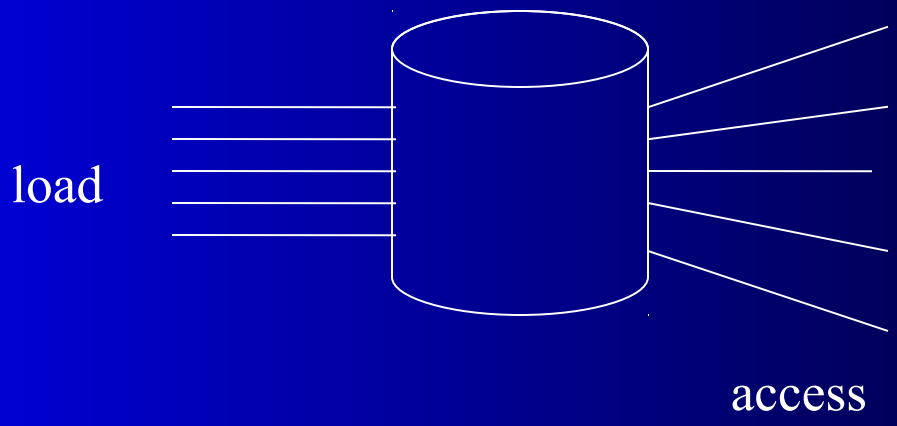


Time-variant

- Provides information from historical perspective e.g. past 5-10 years
- Every key structure contains either implicitly or explicitly an element of time

Nonvolatile

- Data once recorded cannot be updated.
- Data warehouse requires two operations in data accessing
  - Initial loading of data
  - Access of data



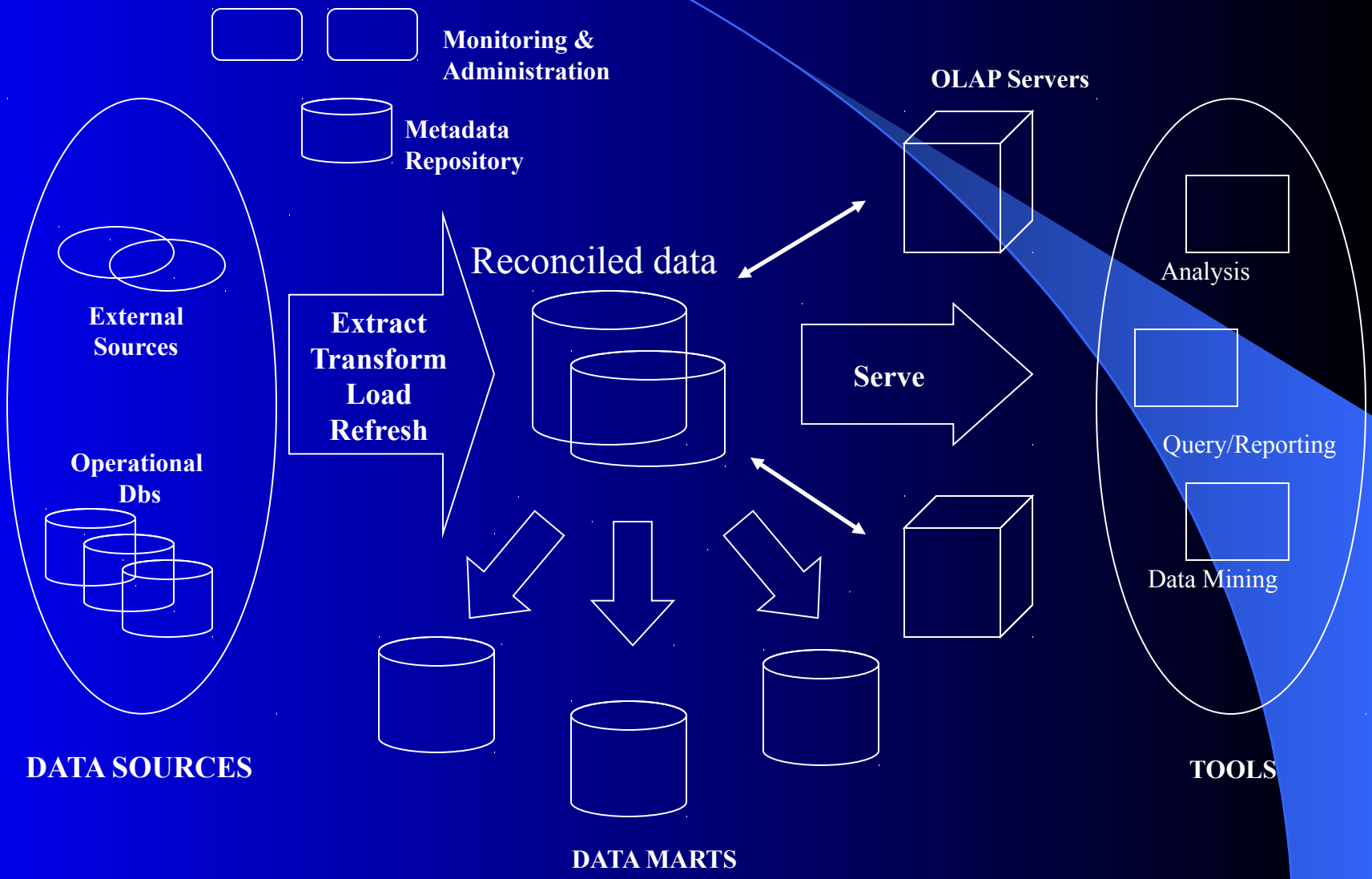
Operational v/s Information System

Features	Operational	Information
Characteristics	Operational processing	Informational processing
Orientation	Transaction	Analysis
User	Clerk,DBA,database professional	Knowledge workers
Function	Day to day operation	Decision support
Data	Current	Historical
View	Detailed,flat relational	Summarized, multidimensional
DB design	Application oriented	Subject oriented
Unit of work	Short ,simple transaction	Complex query
Access	Read/write	Mostly read

Operational v/s Information System

Features	Operational	Information
Focus	Data in	Information out
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100MB to GB	100 GB to TB
Priority	High performance,high availability	High flexibility,end-user autonomy
Metric	Transaction throughput	Query througput

Data Warehousing Architecture



Data Warehouse Architecture

- Data Warehouse server
  - almost always a relational DBMS,rarely flat files
- OLAP servers
  - to support and operate on multi-dimensional data structures
- Clients
  - Query and reporting tools
  - Analysis tools
  - Data mining tools

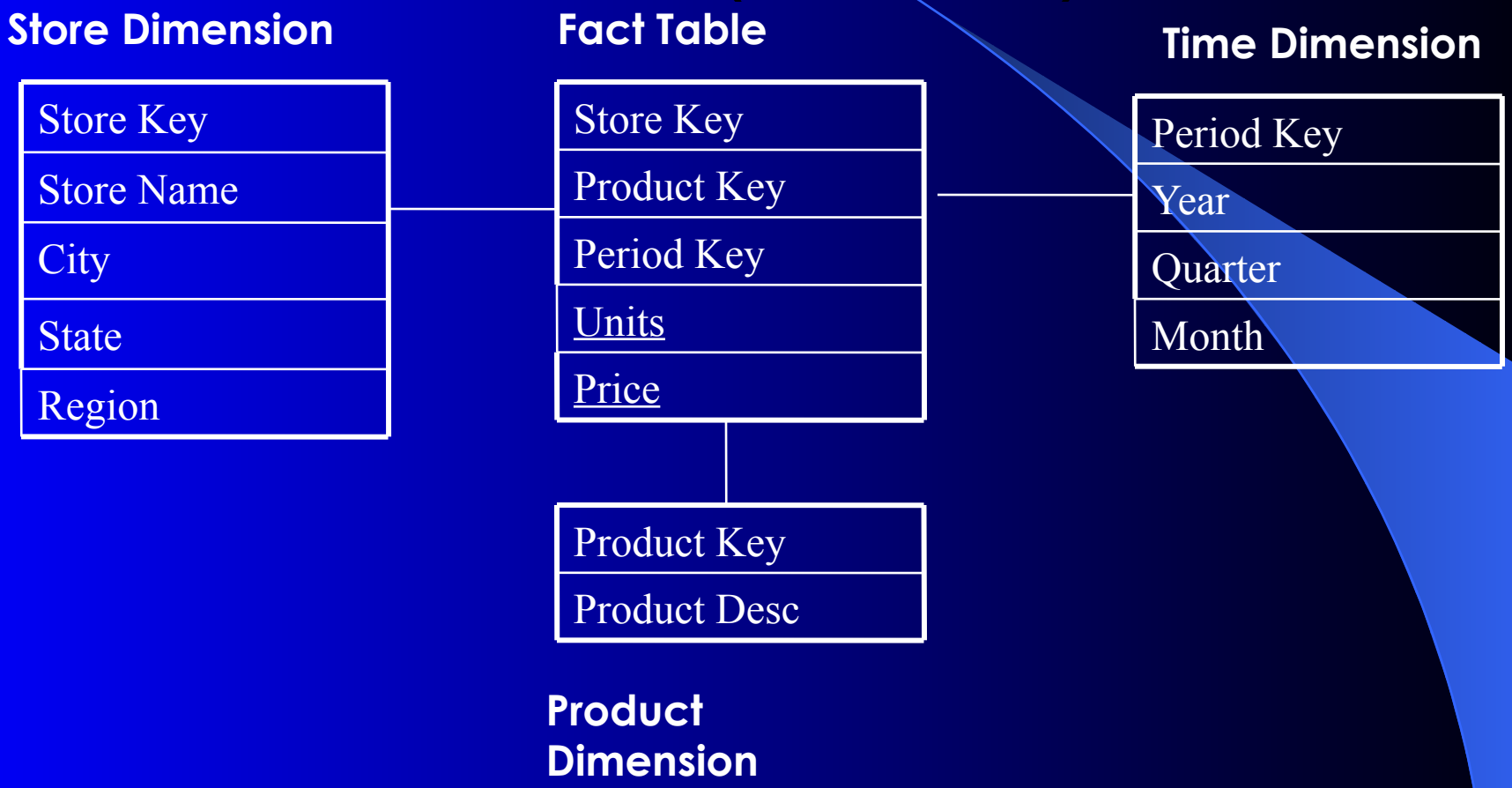
Data Warehouse Schema

- Star Schema
- Fact Constellation Schema
- Snowflake Schema

Star Schema

- A single,large and central fact table and one table for each dimension.
- Every fact points to one tuple in each of the dimensions and has additional attributes.
- Does not capture hierarchies directly.

Star Schema (contd..)

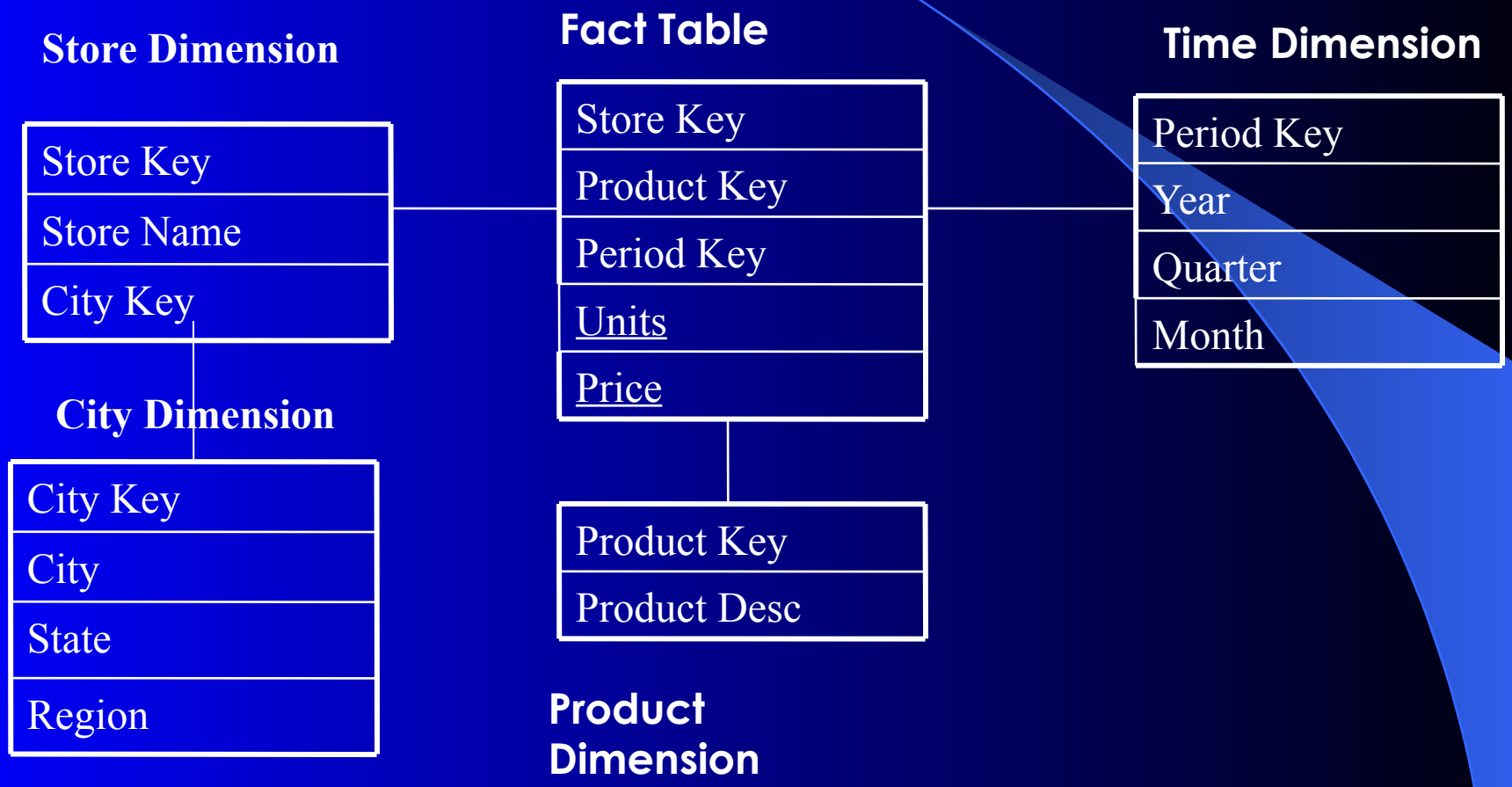


Benefits: Easy to understand, easy to define hierarchies, reduces no. of physical joins.

SnowFlake Schema

- Variant of star schema model.
- A single,large and central fact table and one or more tables for each dimension.
- Dimension tables are normalized i.e. split dimension table data into additional tables

SnowFlake Schema (contd..)

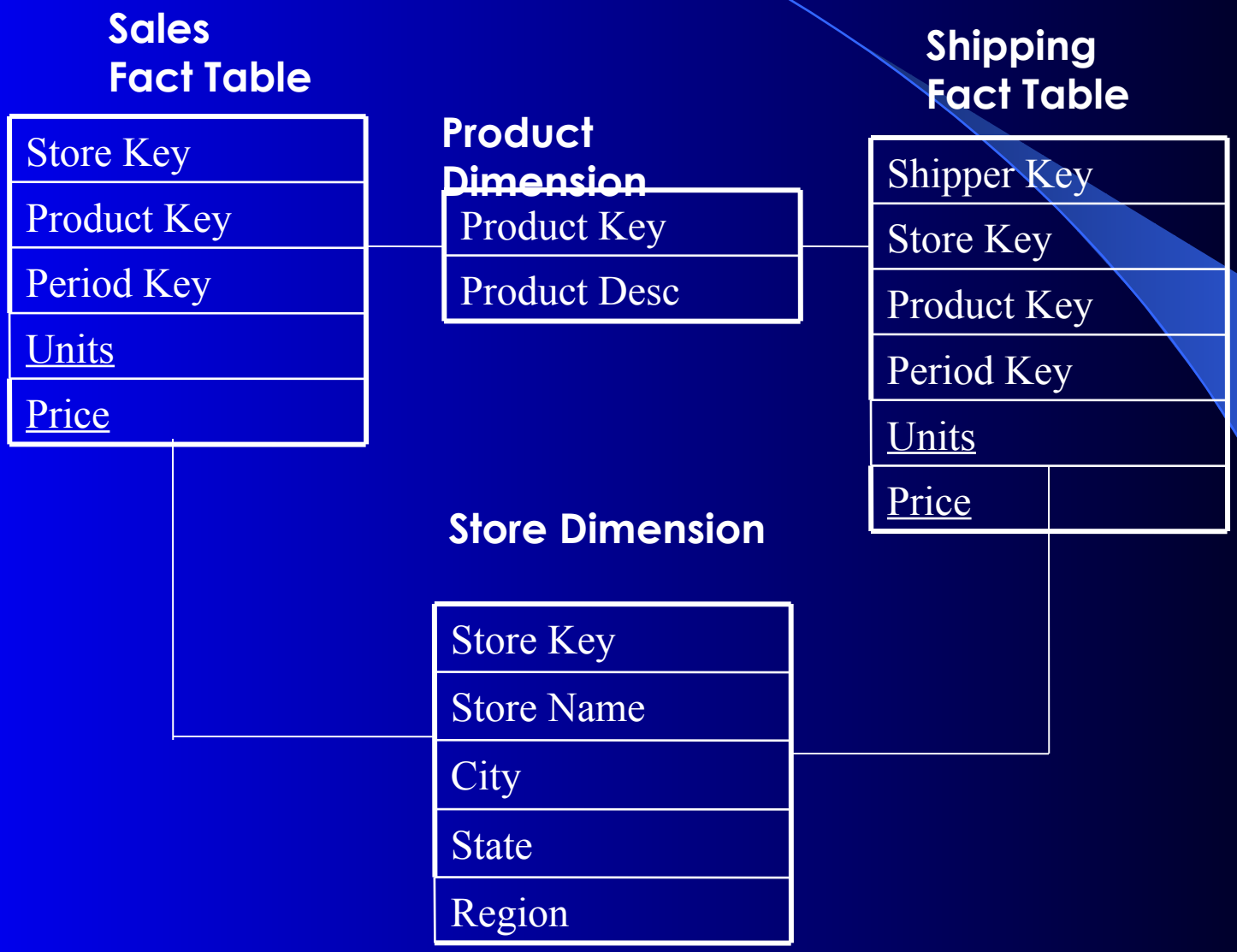


Drawbacks: Time consuming joins,report generation slow

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..)



Building Data Warehouse

- Data Selection
- Data Preprocessing
  - Fill missing values
  - Remove inconsistency
- Data Transformation & Integration
- Data Loading

Data in warehouse is stored in form of fact tables and dimension tables.

Case Study

- Afco Foods & Beverages is a new company which produces dairy,bread and meat products with production unit located at Baroda.
- There products are sold in North,North West and Western region of India.
- They have sales units at Mumbai, Pune , Ahemdabad ,Delhi and Baroda.
- The President of the company wants sales information.

Sales Information

Report: The number of units sold.

113

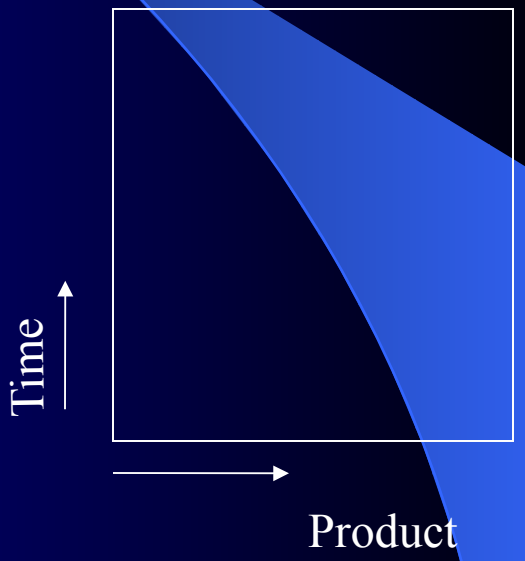
Report: The number of units sold over time

January	February	March	April
14	41	33	25

Sales Information

Report : The number of items sold for each product with time

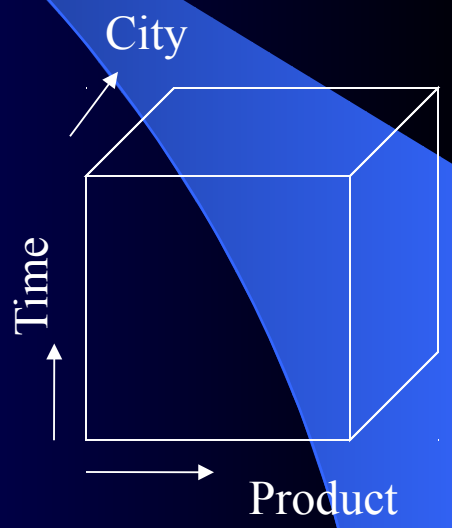
	Jan	Feb	Mar	Apr
Wheat Bread			6	17
Cheese	6	16	6	8
Swiss Rolls	8	25	21	



Sales Information

Report: The number of items sold in each City for each product with time

		Jan	Feb	Mar	Apr
Mumbai	Wheat Bread			3	10
	Cheese	3	16	6	
	Swiss Rolls	4	16	6	
Pune	Wheat Bread			3	7
	Cheese	3			8
	Swiss Rolls	4	9	15	



Sales Information

Report: The number of items sold and income in each region for each product with time.

		Jan		Feb		Mar		Apr	
		Rs	U	Rs	U	Rs	U	Rs	U
Mumbai	Wheat Bread					7.44	3	24.80	10
	Cheese	7.95	3	42.40	16	15.90	6		
	Swiss Rolls	7.32	4	29.98	16	10.98	6		
Pune	Wheat Bread					7.44	3	17.36	7
	Cheese	7.95	3					21.20	8
	Swiss Rolls	7.32	4	16.47	9	27.45	15		

Sales Measures & Dimensions

- Measure – Units sold, Amount.
- Dimensions – Product,Time,Region.

Sales Data Warehouse Model

Fact Table

City	Product	Month	Units	Rupees
Mumbai	Wheat Bread	January	3	7.95
Mumbai	Cheese	January	4	7.32
Pune	Wheat Bread	January	3	7.95
Pune	Cheese	January	4	7.32
Mumbai	Swiss Rolls	February	16	42.40

Sales Data Warehouse Model

City_ID	Prod_ID	Month	Units	Rupees
1	589	1/1/1998	3	7.95
1	1218	1/1/1998	4	7.32
2	589	1/1/1998	3	7.95
2	1218	1/1/1998	4	7.32
1	590	2/1/1998	16	42.40

Sales Data Warehouse Model

Product Dimension Tables

Prod_ID	Product_Name	Product_Category_ID
589	Wheat Bread	1
590	Swiss Rolls	1
288	Coconut Cookies	2

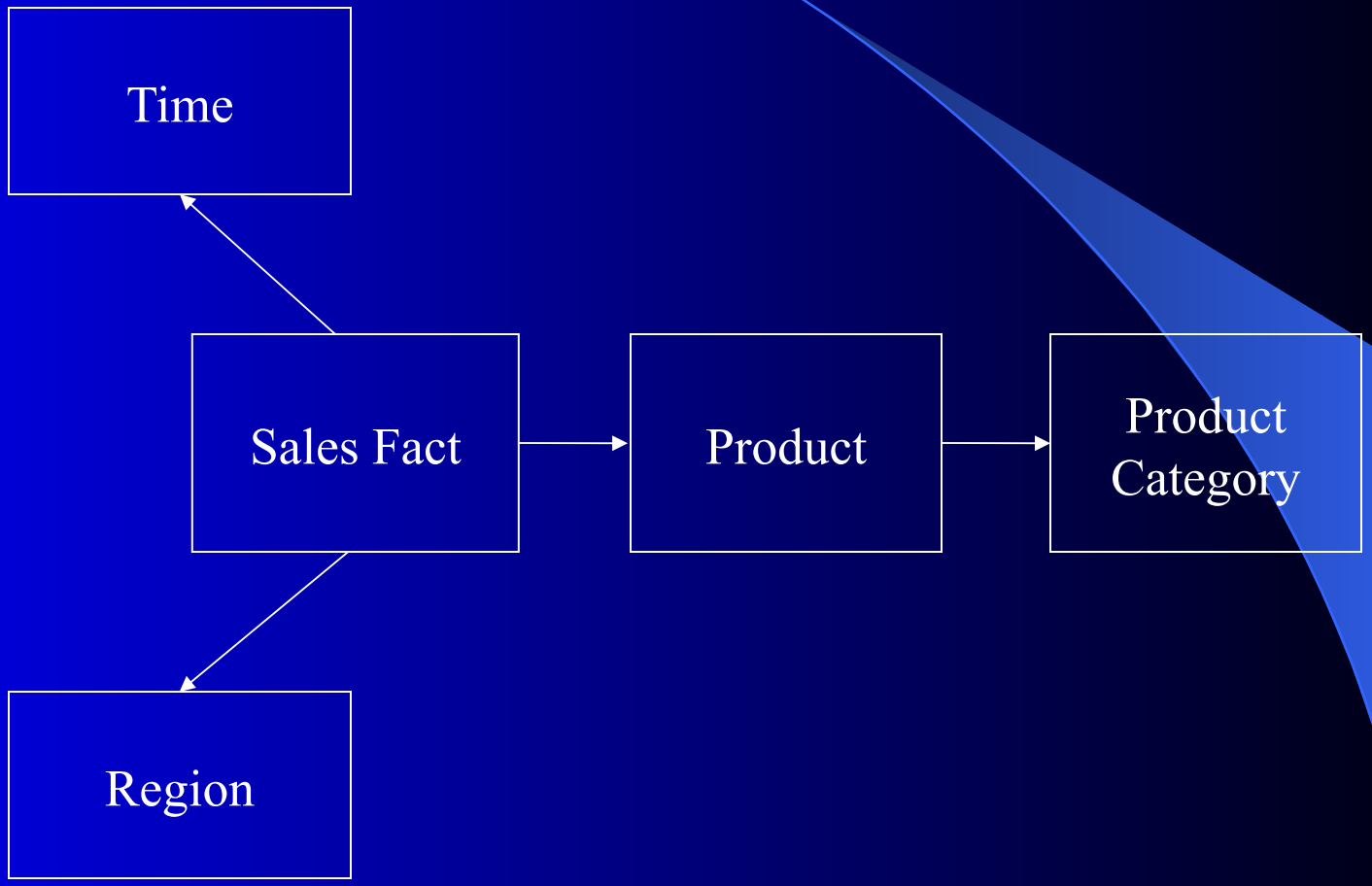
Product_Category_Id	Product_Category
1	Bread
2	Cookies

Sales Data Warehouse Model

Region Dimension Table

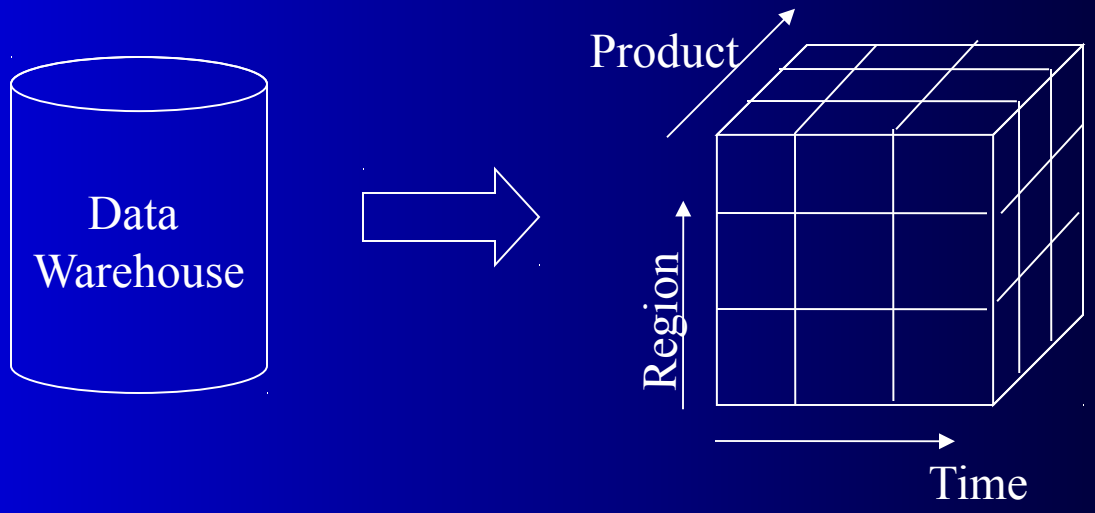
City_ID	City	Region	Country
1	Mumbai	West	India
2	Pune	NorthWest	India

Sales Data Warehouse Model



Online Analysis Processing(OLAP)

- It enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

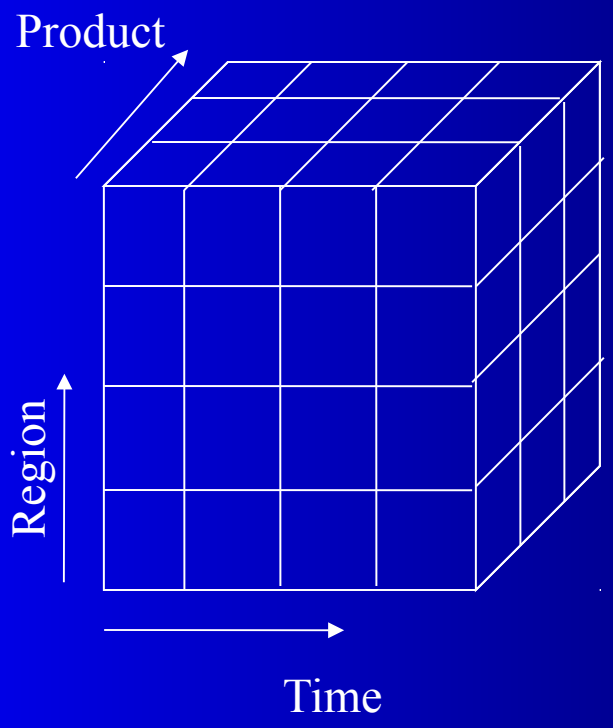


OLAP Cube

City	Product	Time	Units	Dollars
All	All	All	113	251.26
Mumbai	All	All	64	146.07
Mumbai	White Bread	All	38	98.49
Mumbai	Wheat Bread	All	13	32.24
Mumbai	Wheat Bread	Qtr1	3	7.44
Mumbai	Wheat Bread	March	3	7.44

OLAP Operations

Drill Down



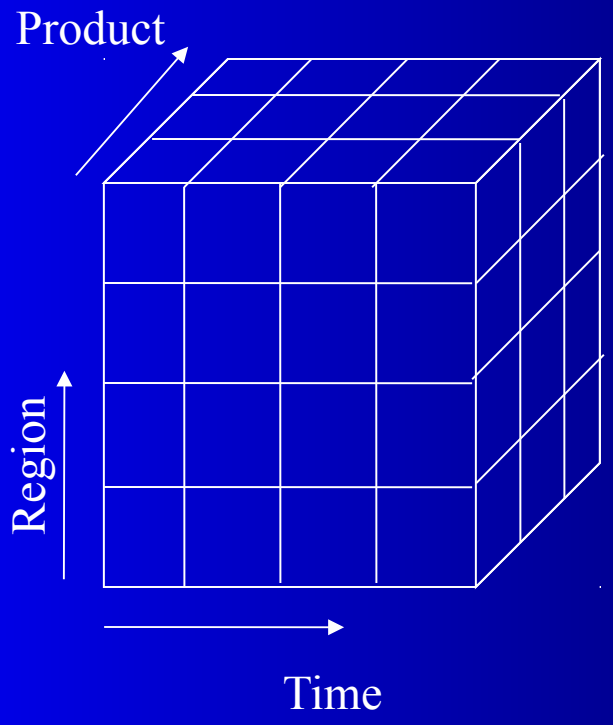
Category e.g Electrical Appliance

Sub Category e.g Kitchen

Product e.g Toaster

OLAP Operations

Drill Up



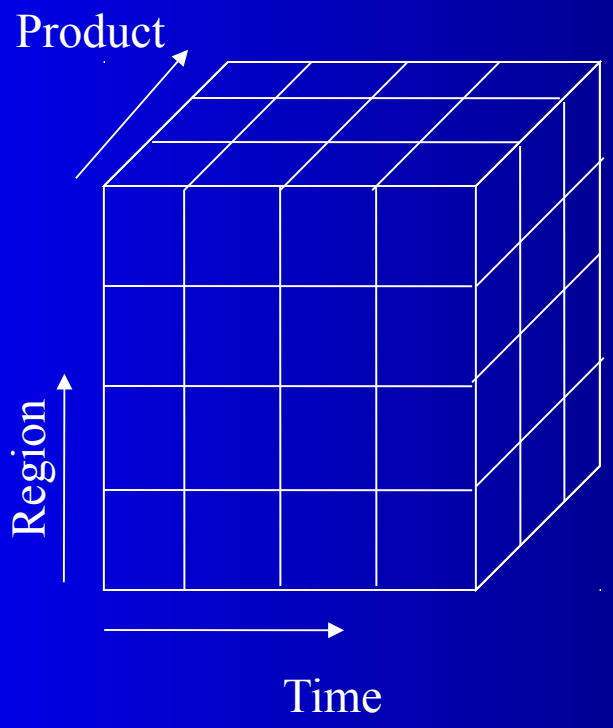
Category e.g Electrical Appliance

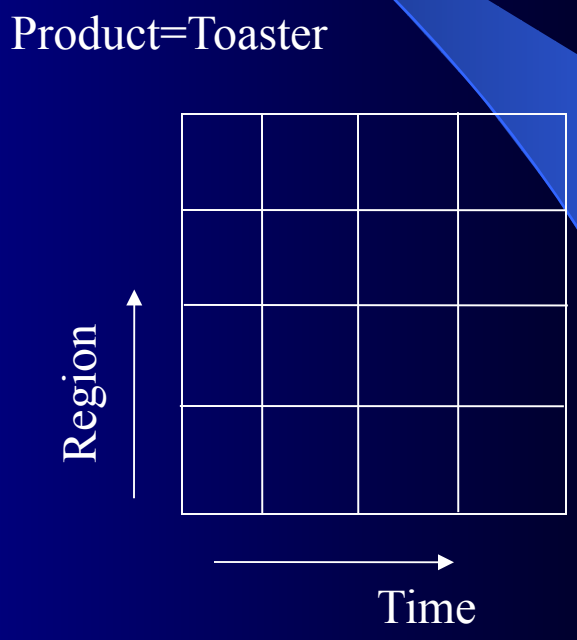
Sub Category e.g Kitchen

Product e.g Toaster

OLAP Operations

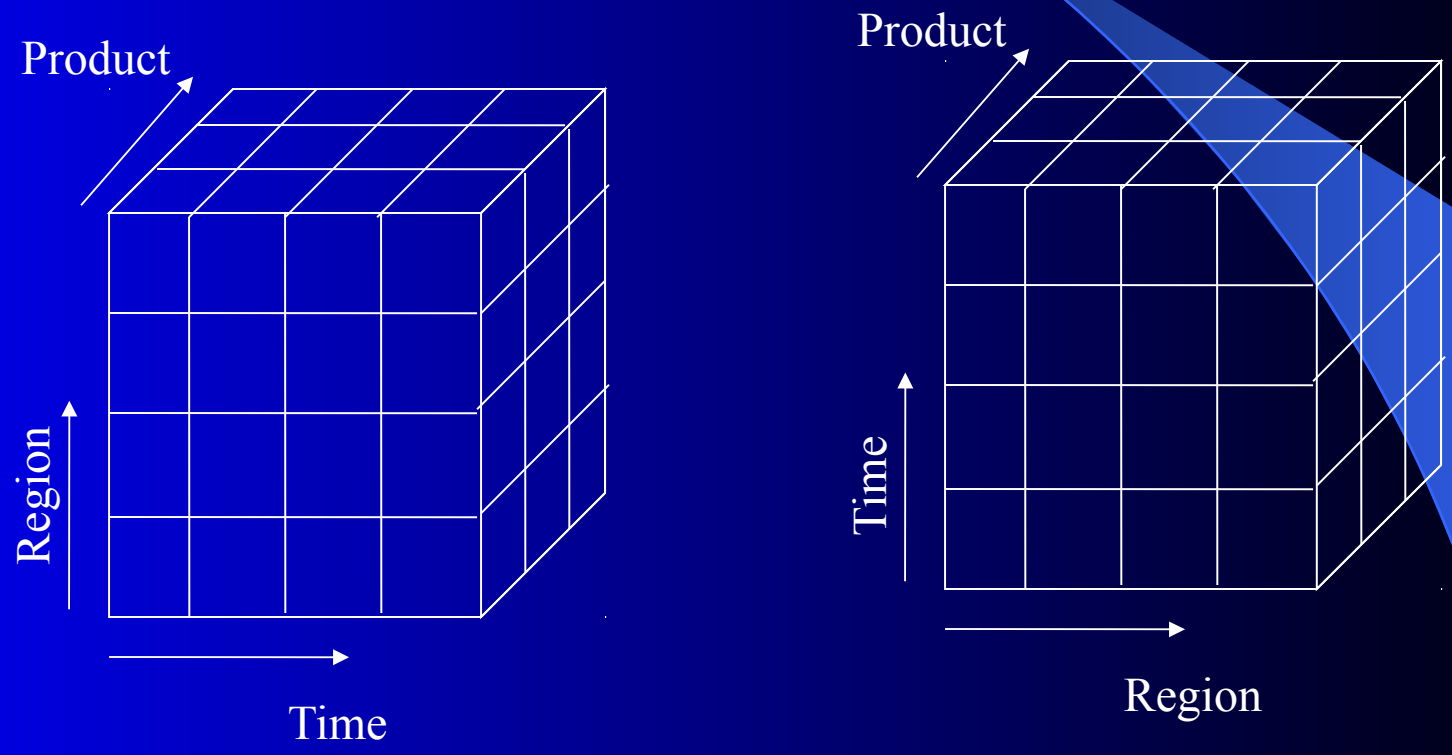
Slice and Dice





OLAP Operations

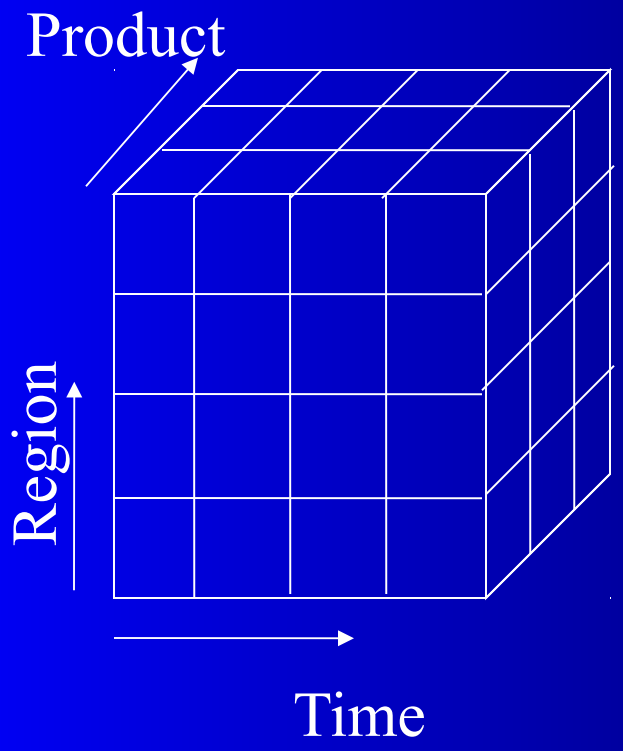
Pivot



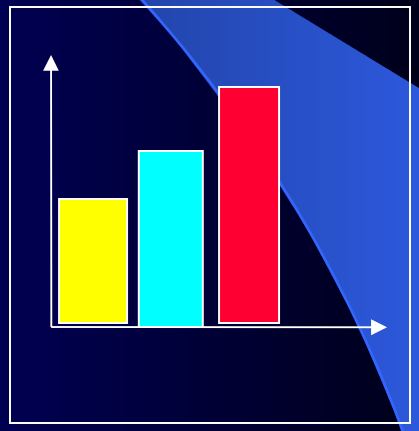
OLAP Server

- An OLAP Server is a high capacity,multi user data manipulation engine specifically designed to support and operate on multi-dimensional data structure.
- OLAP server available are
  - MOLAP server
  - ROLAP server
  - HOLAP server

Presentation



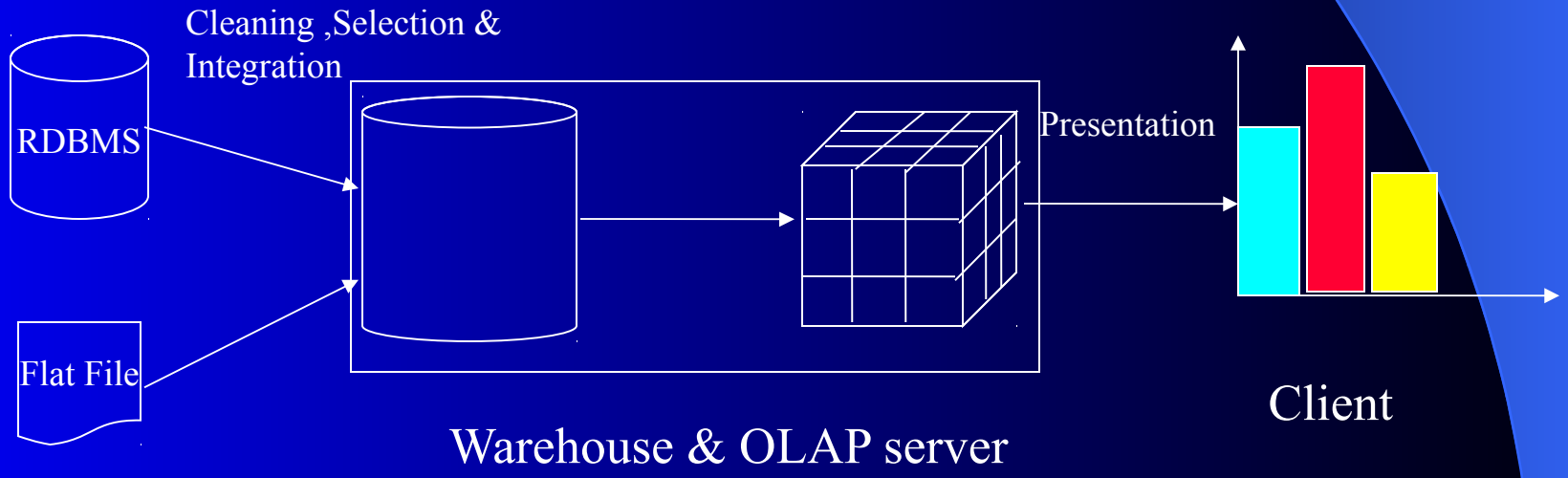
Reporting Tool



Report

Data Warehousing includes

- Build Data Warehouse
- Online analysis processing(OLAP).
- Presentation.



Need for Data Warehousing

- Industry has huge amount of operational data
- Knowledge worker wants to turn this data into useful information.
- This information is used by them to support strategic decision making .

Need for Data Warehousing (contd..)

- It is a platform for consolidated historical data for analysis.
- It stores data of good quality so that knowledge worker can make correct decisions.

Need for Data Warehousing (contd..)

- From business perspective

  -it is latest marketing weapon

  -helps to keep customers by learning more about their needs .

  -valuable tool in today's competitive fast evolving world.

Data Warehousing Tools

- Data Warehouse
  - SQL Server 2000 DTS
  - Oracle 8i Warehouse Builder
- OLAP tools
  - SQL Server Analysis Services
  - Oracle Express Server
- Reporting tools
  - MS Excel Pivot Chart
  - VB Applications

References

- Building Data Warehouse by Inmon
- Data Mining:Concepts and Techniques by Han,Kamber.
- www.dwinfocenter.org
- www.datawarehousingonline.com
- www.billinmon.com

Thank You