Data Warehousing Fundamentals

TE SEM V

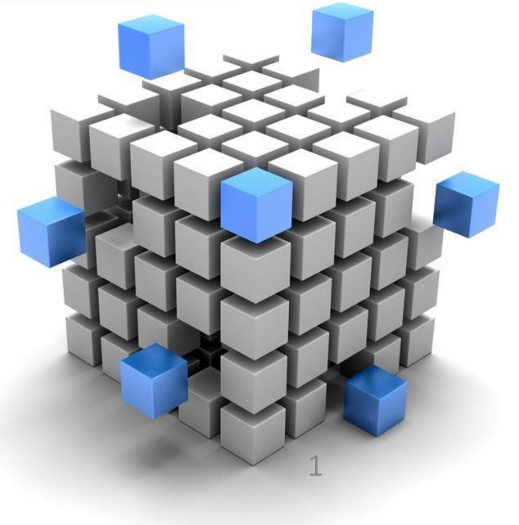
Module 1 : Data Warehousing Fundamentals

Prof. Archana Kotangale

Assistant Professor

Department of Computer Engineering

APSIT, Thane



Syllabus structure

06 Modules(Topics) :Total Lectures = 39

- **Module 1 :** Data Warehousing Fundamentals (8Hrs)
- **Module 2 :** Introduction to Data Mining, Data Exploration and Data Pre-processing (8Hrs)
- **Module 3 :** Classification (6Hrs)
- **Module 4 :** Clustering (6Hrs)
- **Module 5 :** Mining frequent patterns and associations (6Hrs)
- **Module 6 :** Web Mining (5Hrs)
- **Useful Links :**
- https://onlinecourses.nptel.ac.in/noc20_cs12/preview
- https://www.coursera.org/specializations/data-mining

Examination scheme and T/W marking scheme

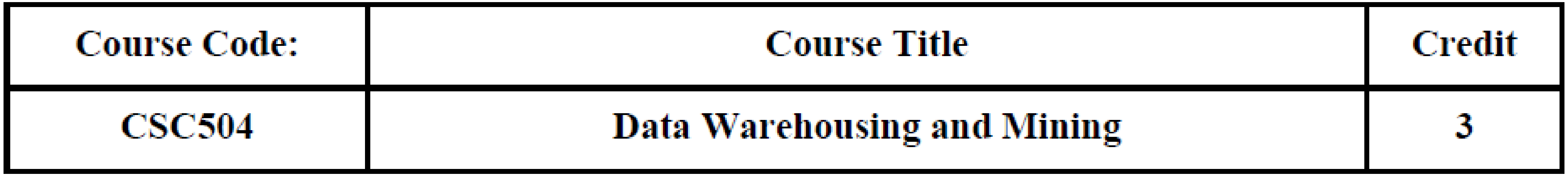
Total Marks=150

1. Internal assessment: Average Test Marks=20

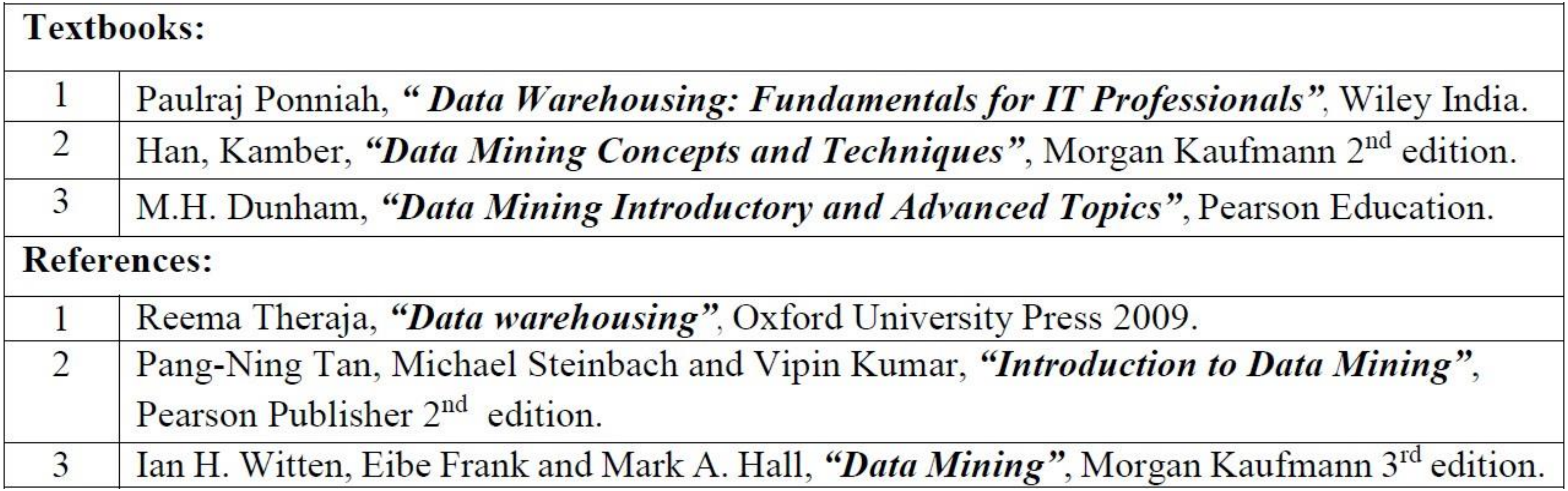
   [Test:1(20 Marks)+Test:2(20 Marks)]/2

2. End semester theory exam=80
3. External Oral and Practical exam=25
4. Term Work=25

Textbooks and Reference Books



Course Code:	Course Title	Credit
CSC504	Data Warehousing and Mining	3



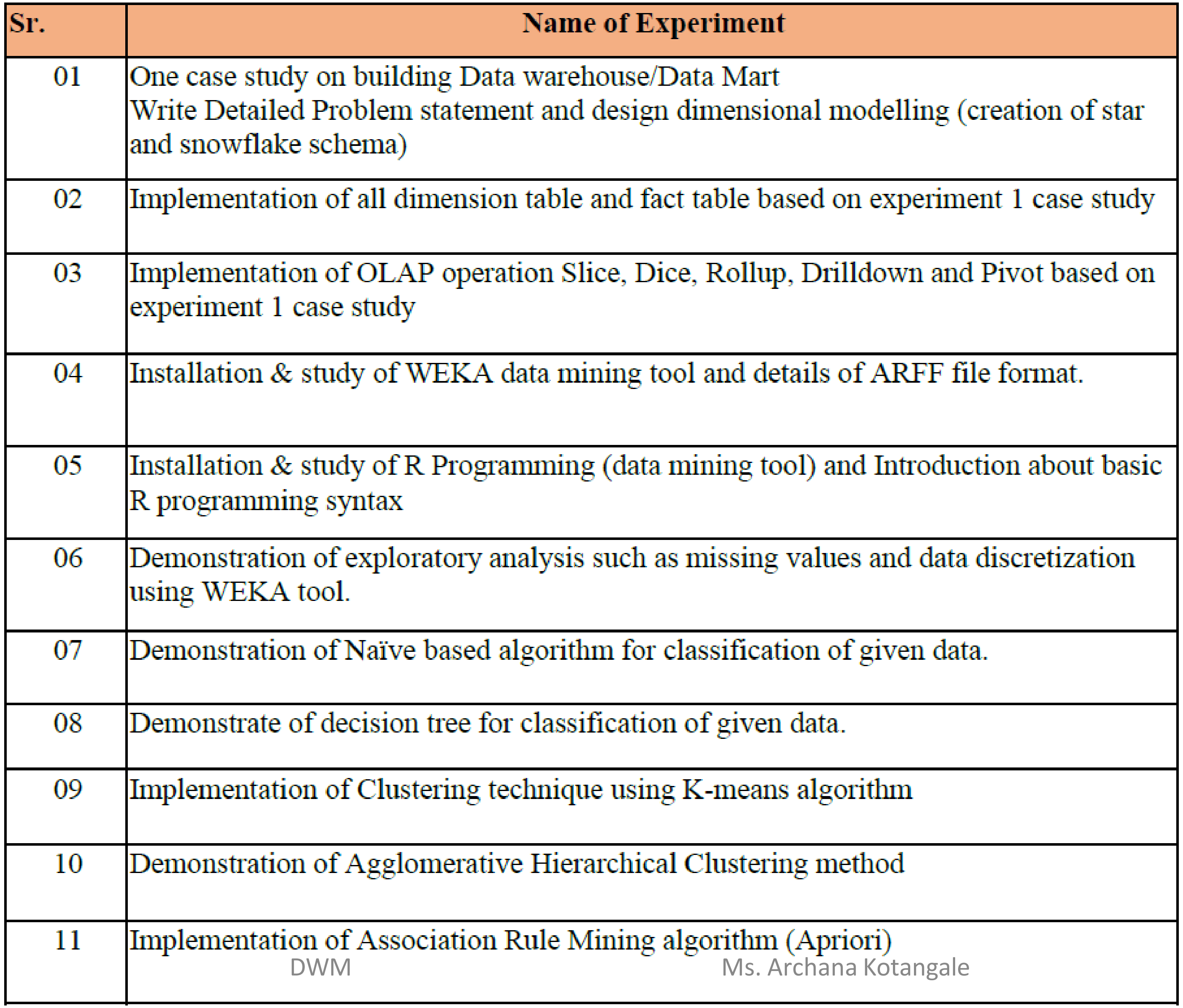
Textbooks:	
1	Paulraj Ponniah, "*Data Warehousing: Fundamentals for IT Professionals*", Wiley India.
2	Han, Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann 2nd edition.
3	M.H. Dunham, "*Data Mining Introductory and Advanced Topics*", Pearson Education.
References:	
1	Reema Theraja, "*Data warehousing*", Oxford University Press 2009.
2	Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "*Introduction to Data Mining*", Pearson Publisher 2nd edition.
3	Ian H. Witten, Eibe Frank and Mark A. Hall, "*Data Mining*", Morgan Kaufmann 3rd edition.

Pre- Requisite

Before proceeding with this course, you should have an understanding of basic database concepts such as

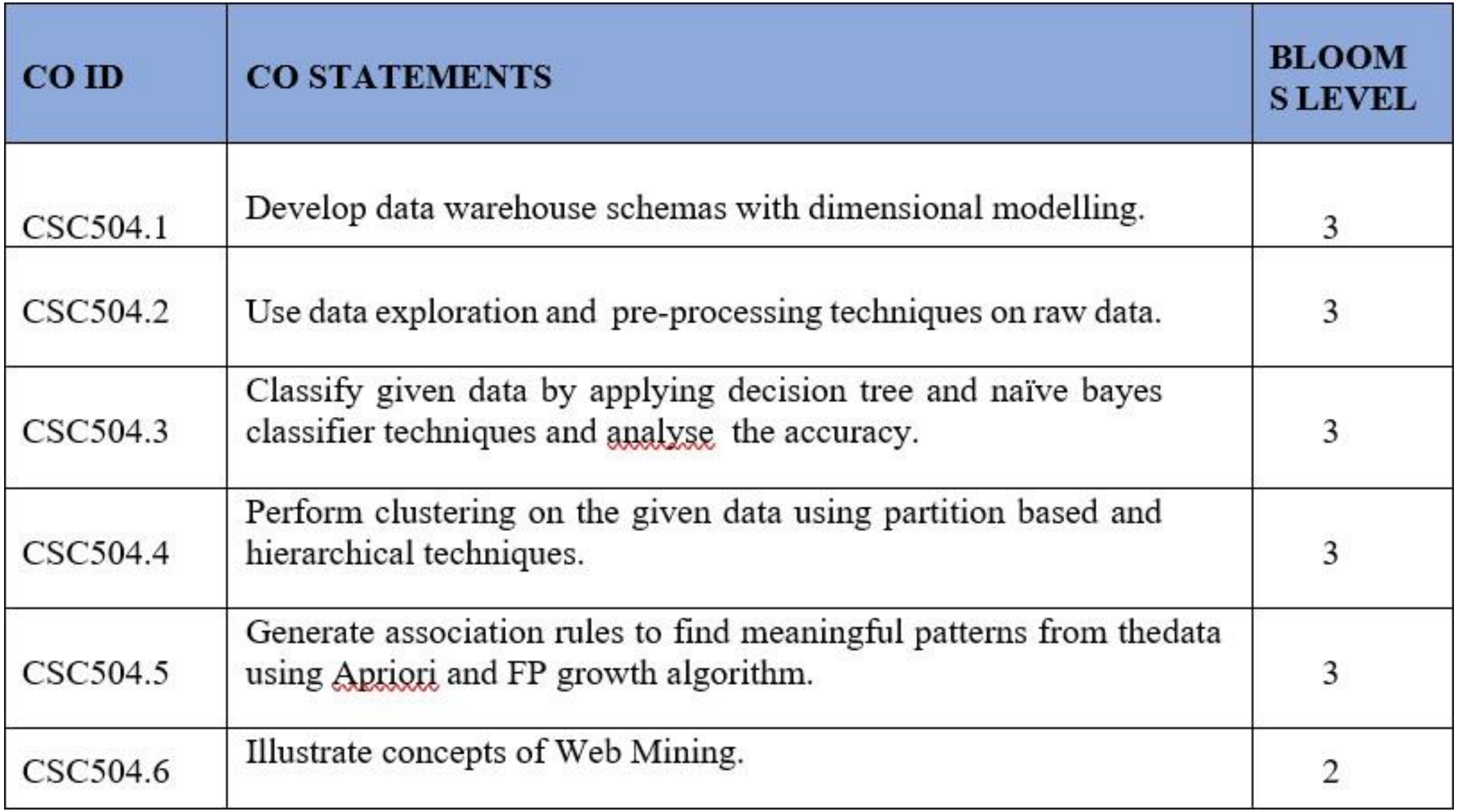
1. RDBMS
2. Schema
3. ER model
4. Normalization and denormalization
5. Keys
6. Structured query language, etc.

Practical List



Sr.	Name of Experiment
01	One case study on building Data warehouse/Data Mart Write Detailed Problem statement and design dimensional modelling (creation of star and snowflake schema)
02	Implementation of all dimension table and fact table based on experiment 1 case study
03	Implementation of OLAP operation Slice, Dice, Rollup, Drilldown and Pivot based on experiment 1 case study
04	Installation & study of WEKA data mining tool and details of ARFF file format.
05	Installation & study of R Programming (data mining tool) and Introduction about basic R programming syntax
06	Demonstration of exploratory analysis such as missing values and data discretization using WEKA tool.
07	Demonstration of Naïve based algorithm for classification of given data.
08	Demonstrate of decision tree for classification of given data.
09	Implementation of Clustering technique using K-means algorithm
10	Demonstration of Agglomerative Hierarchical Clustering method
11	Implementation of Association Rule Mining algorithm (Apriori)

Course Outcomes



CO ID	CO STATEMENTS	BLOOM S LEVEL
CSC504.1	Develop data warehouse schemas with dimensional modelling.	3
CSC504.2	Use data exploration and pre-processing techniques on raw data.	3
CSC504.3	Classify given data by applying decision tree and naïve bayes classifier techniques and analyse the accuracy.	3
CSC504.4	Perform clustering on the given data using partition based and hierarchical techniques.	3
CSC504.5	Generate association rules to find meaningful patterns from thedata using Apriori and FP growth algorithm.	3
CSC504.6	Illustrate concepts of Web Mining.	2

Objective

- Understand the desperate need for strategic information
- Recognize the information crisis at every enterprise
- Distinguish between operational and informational systems
- Why data warehousing is the viable solution

Data and Information

- **Data is a collection of raw facts while information puts those facts context**
- Data is raw and unorganized, information is organized
- Data points are individual and sometimes unrelated.
- Information maps out that data to provide a big-picture view of how it all fits together.

DATA
INFORMATION

- Data, on its own, is meaningless. When it’s analyzed and interpreted, it becomes meaningful information.
- Data does not depend on information; however, information depends on data.
- Data isn’t sufficient for decision-making, but you can make decisions based on information.

Data and Information

Example

- At a restaurant, a single customer’s bill amount is data. However, when the restaurant owners collect and interpret multiple bills over a range of time, they can produce valuable information, such as what menu items are most popular
- The number of likes on a social media post is a single element of data. When that’s combined with other social media engagement statistics, like followers, comments, and shares, a company can intuit which social media platforms perform the best and which platforms they should focus on to more effectively engage their audience.

General Meaning of warehouse and mining

1. Warehouse

A warehouse is a commercial building for storage of goods.

Warehouses are used by manufacturers, importers, exporters, wholesalers, transport businesses, customs, etc.

General Meaning of Data warehouse and Data mining

2. Mining

Mining is the extraction of valuable minerals or other geological materials from the earth.

Mining of stones, diamonds and metal has been a human activity since pre-historic times

Meaning of Data warehouse and Data mining (in context of computer science)

1. Data Warehouse:

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources for the strategic decisions and analysis of data.

Meaning of Data warehouse and Data mining
(in context of computer science)

2. Data Mining:

Data mining is the computing process of discovering patterns in large data sets to predict future trends.

e.g.

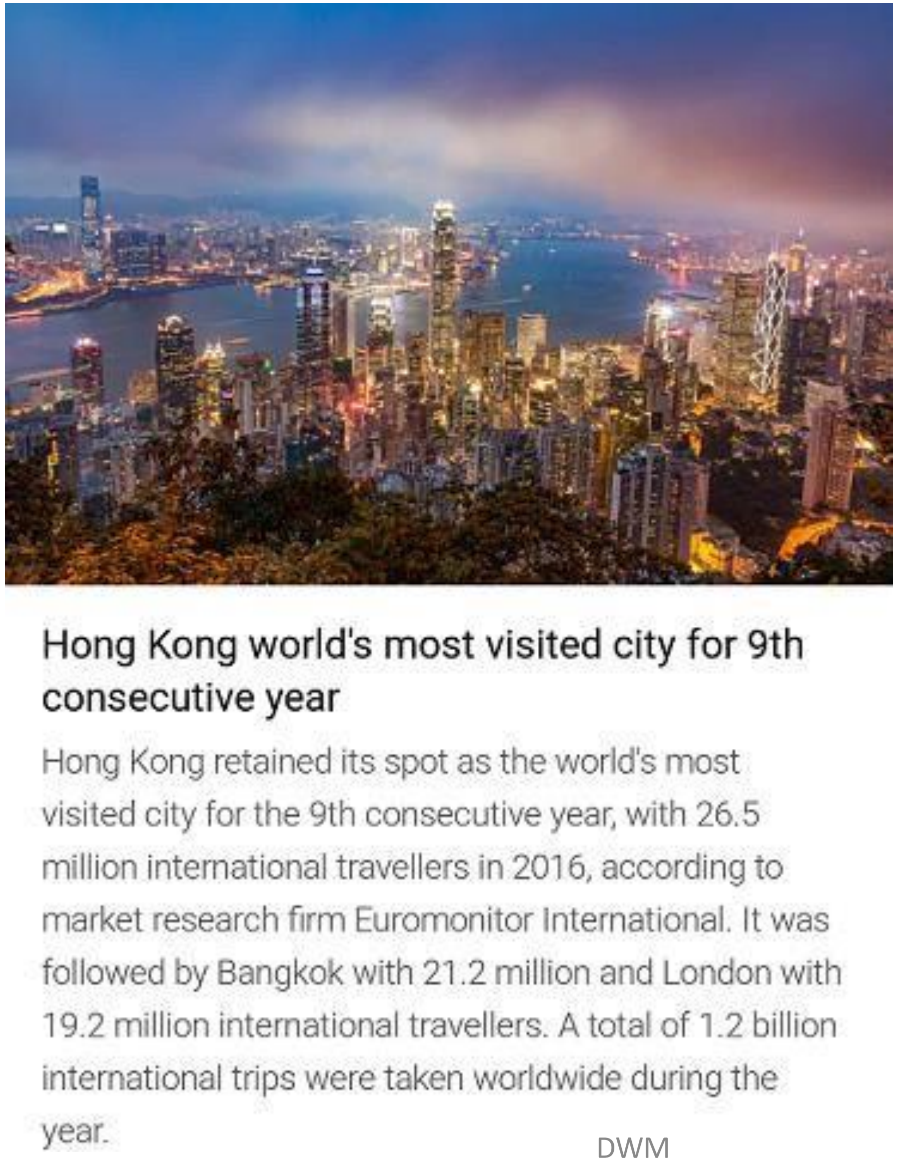
1.Wednesday bazaar (big bazaar offer)

2.KFC Wednesday offer

3.End of the season sale

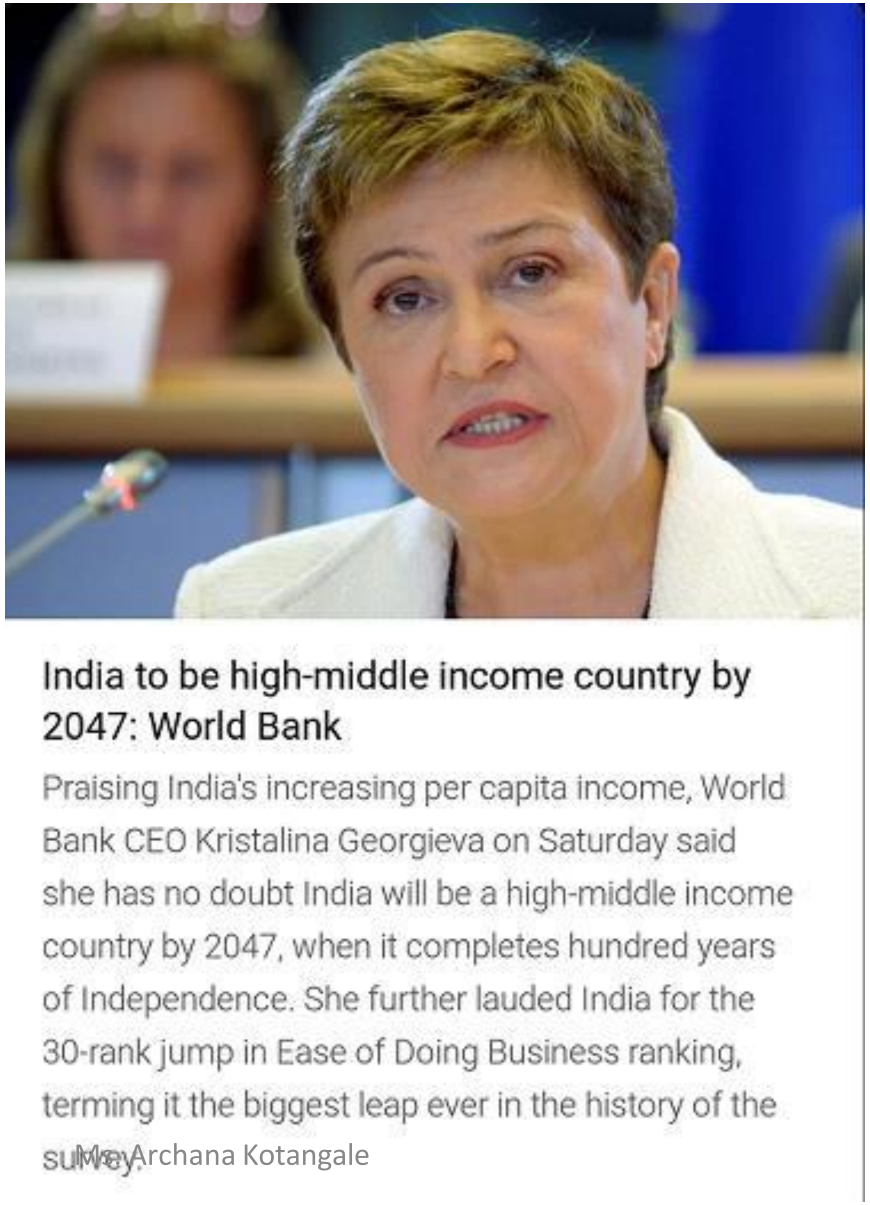
4.Buy one get one movie ticket offer on some credit card

Real time example of data warehousing and mining



Hong Kong world's most visited city for 9th consecutive year

Hong Kong retained its spot as the world's most visited city for the 9th consecutive year, with 26.5 million international travellers in 2016, according to market research firm Euromonitor International. It was followed by Bangkok with 21.2 million and London with 19.2 million international travellers. A total of 1.2 billion international trips were taken worldwide during the year.



India to be high-middle income country by 2047: World Bank

Praising India's increasing per capita income, World Bank CEO Kristalina Georgieva on Saturday said she has no doubt India will be a high-middle income country by 2047, when it completes hundred years of Independence. She further lauded India for the 30-rank jump in Ease of Doing Business ranking, terming it the biggest leap ever in the history of the survey.

How could they make such informed predictions??

Concept Background

Concept

- Some applications are very important to run the business
- These applications are order processing, maintain inventory, keep the accounting books, service the clients, receive payments, and process claims.
- In 1960s, companies are started building and using applications
- As company grows hundreds of computer applications are needed to support these various business process.
- These application store, gather and process the data needed to perform the daily routine business.

Need of data warehousing

Concept

- In 1990s, as business grew more complex, corporations spread globally and competition became aggressive
- Business executive becomes desperate for information to stay competitive and commercially successful.
- Operational systems was providing information

  to run day to day operations

  but not sufficient of **strategic decision**.

- Business executive needed the type of information with proper content and format that could help them make such strategic decision.
- Data warehousing is a new paradigm specially intended to provide vital strategic information

Need of data warehousing

Advantages achieved by organization using the data warehousing

- Retail
  - Customer Loyalty
  - Market Planning
- Financial
  - Risk Management
  - Fraud Detection
- Airlines
  - Route Profitability
  - Yield Management

- Manufacturing
  - Cost Reduction
  - Logistic Management
- Utilities
  - Asset Management
  - Resource Management
- Government
  - Manpower Planning
  - Cost Control

Escalating Need For Strategic Information

Strategic Decision

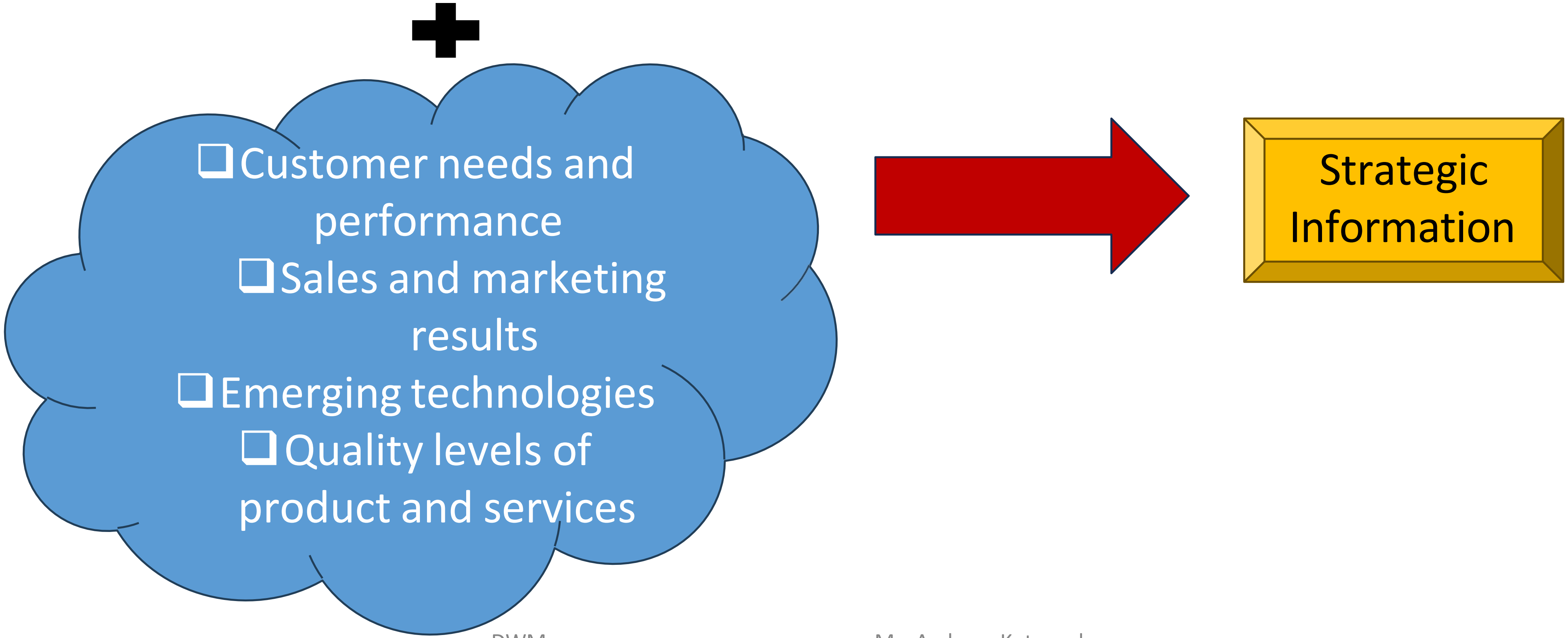
- Business executive and managers need information to formulate the business strategies, establish goals , set objectives and monitor results
- Some examples of business objectives

- Retain the present customer base
- Increase the customer base by 15% over the next 5 years
- Gain market share by 10% in the next 3 years
- Improve product quality levels in the top five product groups
- Enhance customer service level in shipments
- Bring three new products to market in 2 years
- Increase sales by 15% in the specific region

Escalating Need For Strategic Information

Strategic Information

In-depth knowledge of company's operations and all performance measures



+

- □ Customer needs and performance
- □ Sales and marketing results
- □ Emerging technologies
- □ Quality levels of product and services

Strategic Information

Escalating Need For Strategic Information

Characteristic Of Strategic Information

▪ Integrated	➢ Must have an overall enterprise wide view
▪ Data integrity	➢ Data in all the tables must be accurate and must conform to business rules
▪ Accessibility	➢ Easily accessible by the users with respective access paths
▪ Credible	➢ Every business factor must have one and only one value
▪ Timely	➢ Information must be available within stipulated time

Need of Data Warehousing

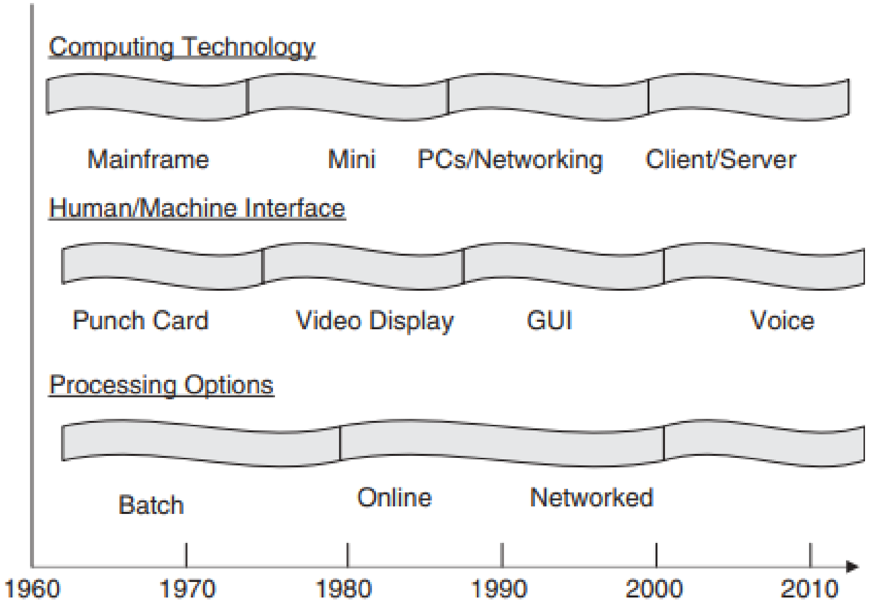
Data, Data Everywhere Yet The Information Crisis

- I can’t find the data I need
  - data is scattered over the network
  - many versions, subtle differences
- I can’t get the data I need
  - need an expert to get the data
- I can’t understand the data I found
  - available data poorly documented
- I can’t use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

Need of Data Warehousing

Supportive Technology Trends

- Computing Technology : become faster, cheaper and widely available, reduction in storage cost
- Human Machine Interface : more and more interactive interface software
- Processing Options : high processing speed can be possible



Need of Data Warehousing

History of decision support system

- Discuss a case study
- Ad hoc reports
- Special extract Program
- Small Application
- Information Centers
- Executive Information System

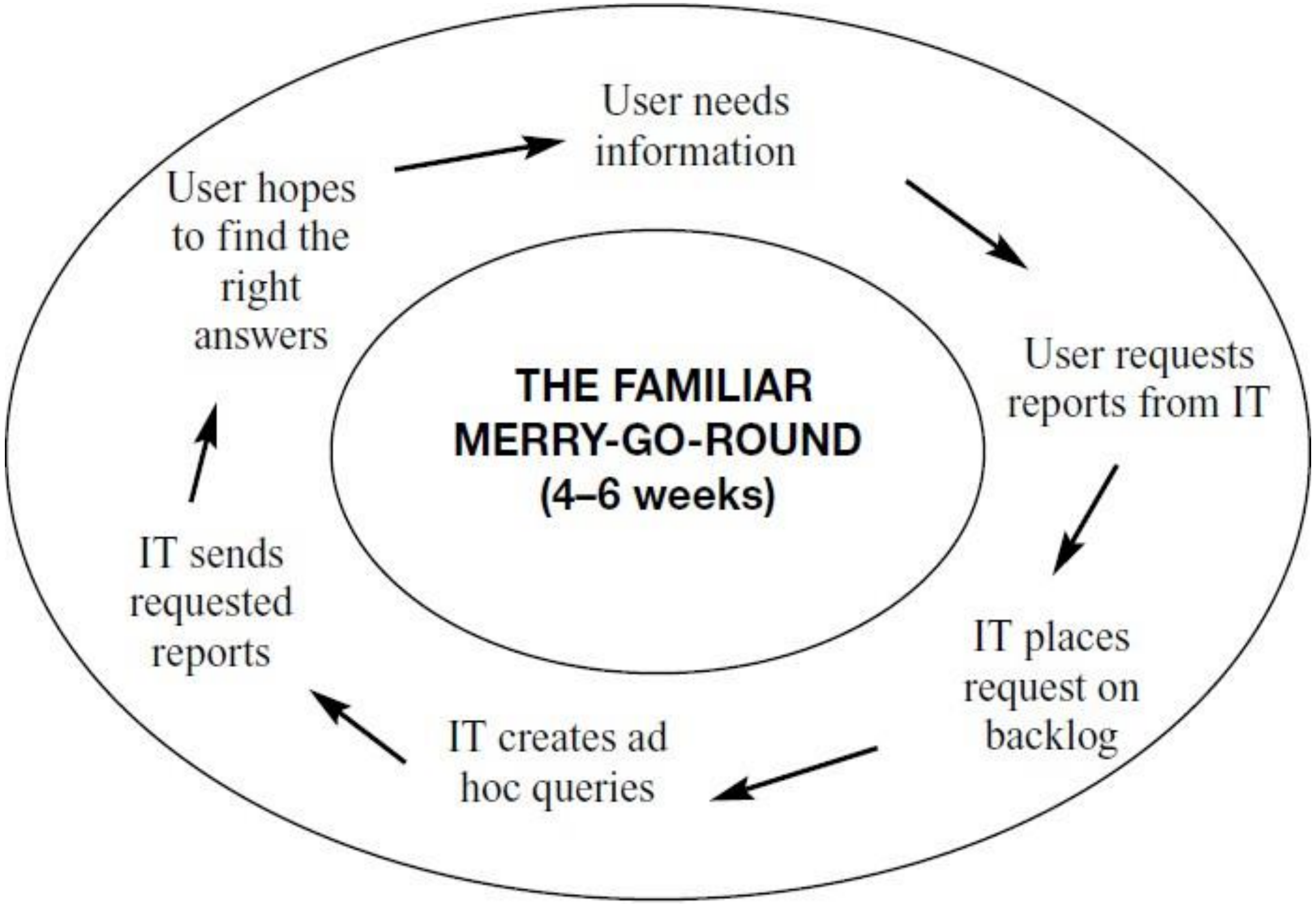

User needs information
User requests reports from IT
IT places request on backlog
IT creates ad hoc queries
IT sends requested reports
User hopes to find the right answers
THE FAMILIAR MERRY-GO-ROUND (4–6 weeks)

Figure 1-4 Inadequate attempts by IT to provide strategic information.

Operational versus Decision Support System

Operational System

- These are the systems that are used to run the day to day core business of the company.
- Operational systems are online transactional processing(OLTP) system
- Some examples as,
  - ✓Take an order
  - ✓Process a claim
  - ✓Make a shipment
  - ✓Generate an invoice
  - ✓Receive cash
  - ✓Online Payment
  - ✓Production management
  - ✓Reserves movie/airline/bus seat etc

RUN The Business

Ooperational verses Decision Support System

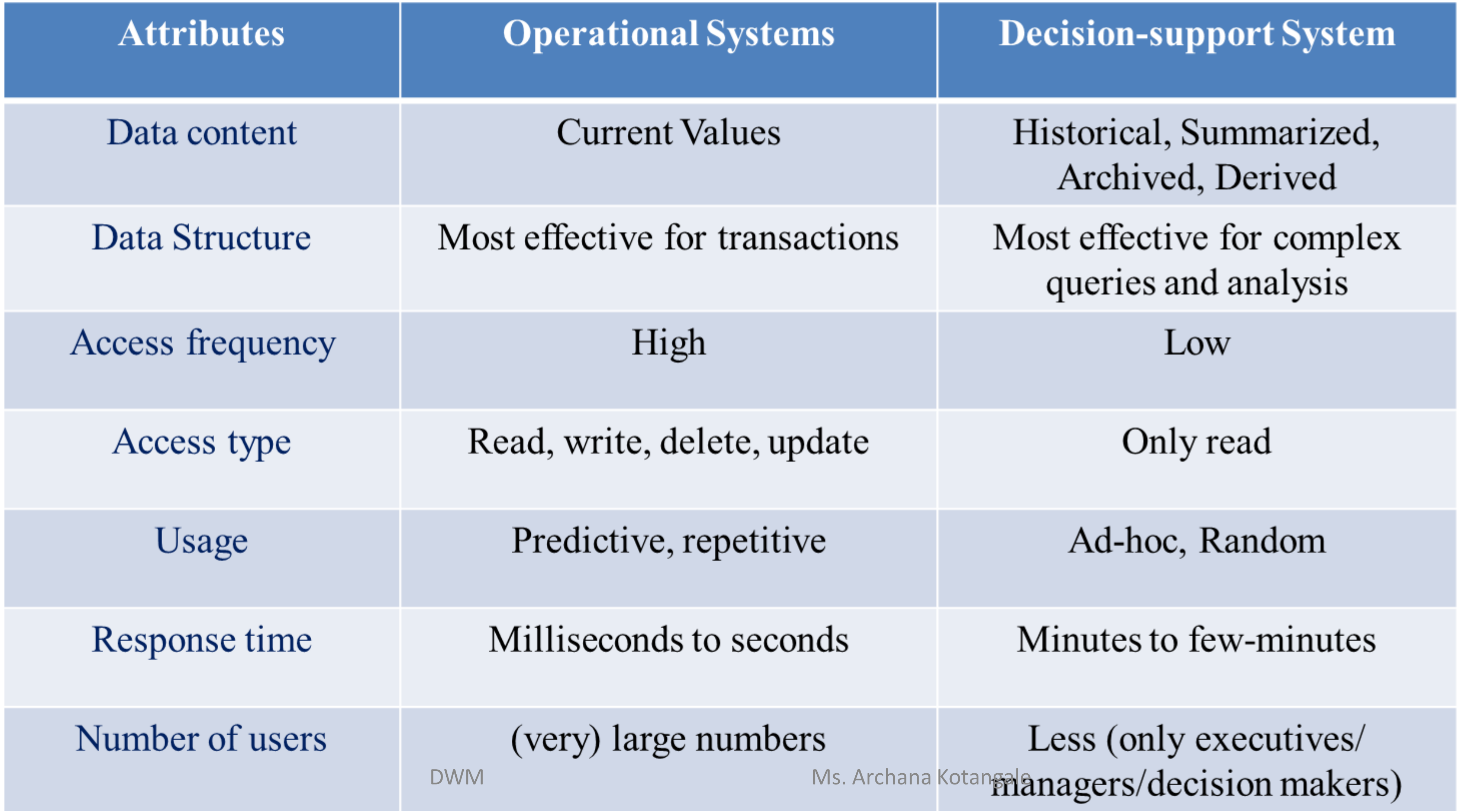
Decision Support System

- Based on the strategic information, decisions has to be taken
- How the business run and then make the strategic decisions to improve the business
  - ✓ Show me the top selling products
  - ✓ Show me the problem region
  - ✓ Tell me why loss happen in this region
  - ✓ Let me see the broader data
  - ✓ Show me the total profit
  - ✓ Alert me when sell goes below target etc

Business Decision

Ooperational verses Decision Support System

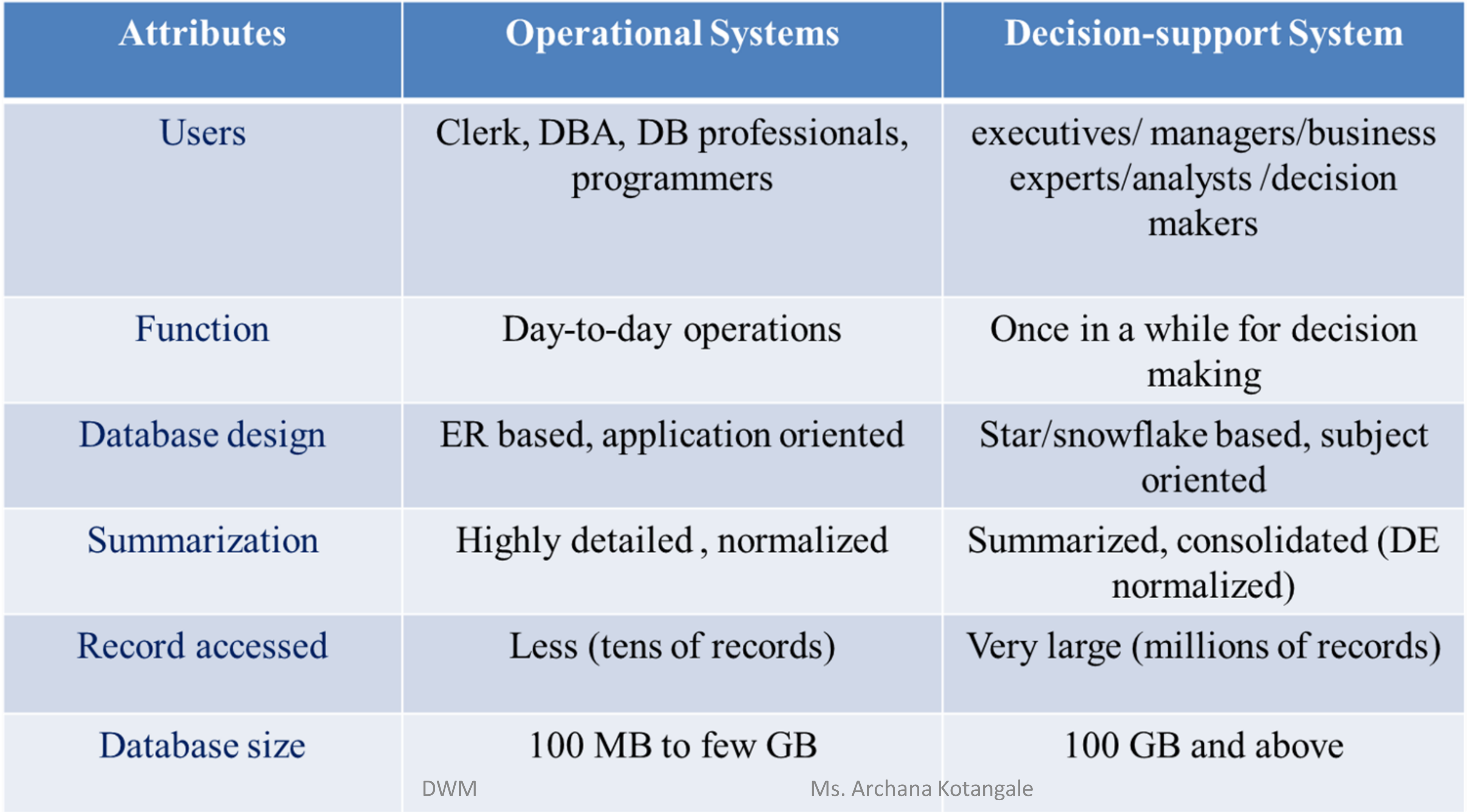
How are they different



Attributes	Operational Systems	Decision-support System
Data content	Current Values	Historical, Summarized, Archived, Derived
Data Structure	Most effective for transactions	Most effective for complex queries and analysis
Access frequency	High	Low
Access type	Read, write, delete, update	Only read
Usage	Predictive, repetitive	Ad-hoc, Random
Response time	Milliseconds to seconds	Minutes to few-minutes
Number of users	(very) large numbers	Less (only executives/ managers/decision makers)

Ooperational verses Decision Support System

How are they different



Attributes	Operational Systems	Decision-support System
Users	Clerk, DBA, DB professionals, programmers	executives/ managers/business experts/analysts /decision makers
Function	Day-to-day operations	Once in a while for decision making
Database design	ER based, application oriented	Star/snowflake based, subject oriented
Summarization	Highly detailed , normalized	Summarized, consolidated (DE normalized)
Record accessed	Less (tens of records)	Very large (millions of records)
Database size	100 MB to few GB	100 GB and above

Ooperational verses Decision Support System

How are they different

Attributes	Operational Systems	Decision-support System
Priority	High-performance and high-availability	High flexibility and end user autonomy
Indexes	Few	Many
Joins	Many	Few

Data warehouse Is The Solution

A features of new type of system Environment

- Database designed for analytical task
- Data from multiple applications
- Easy to use and conducive to long interactive sessions by users
- Read-intensive data usage
- Direct interaction with the system by the users without IT assistance
- Content updated periodically and stable
- Content to include current and historical data
- Ability for users to run queries and get results online
- Ability for users to initiate reports

Data warehouse Is The Solution

Strategic information from the data warehouse

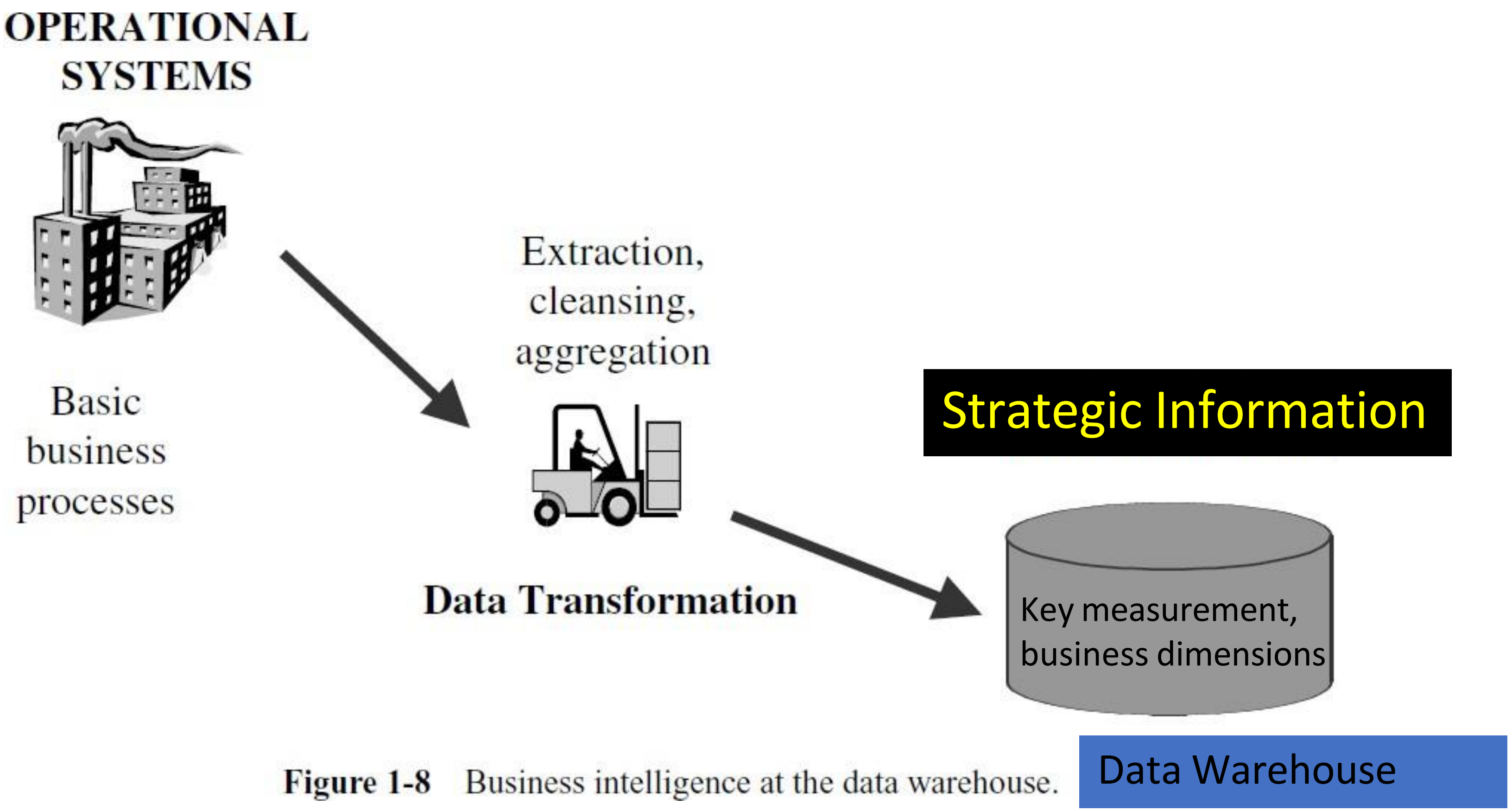


Figure 1-8 Business intelligence at the data warehouse.

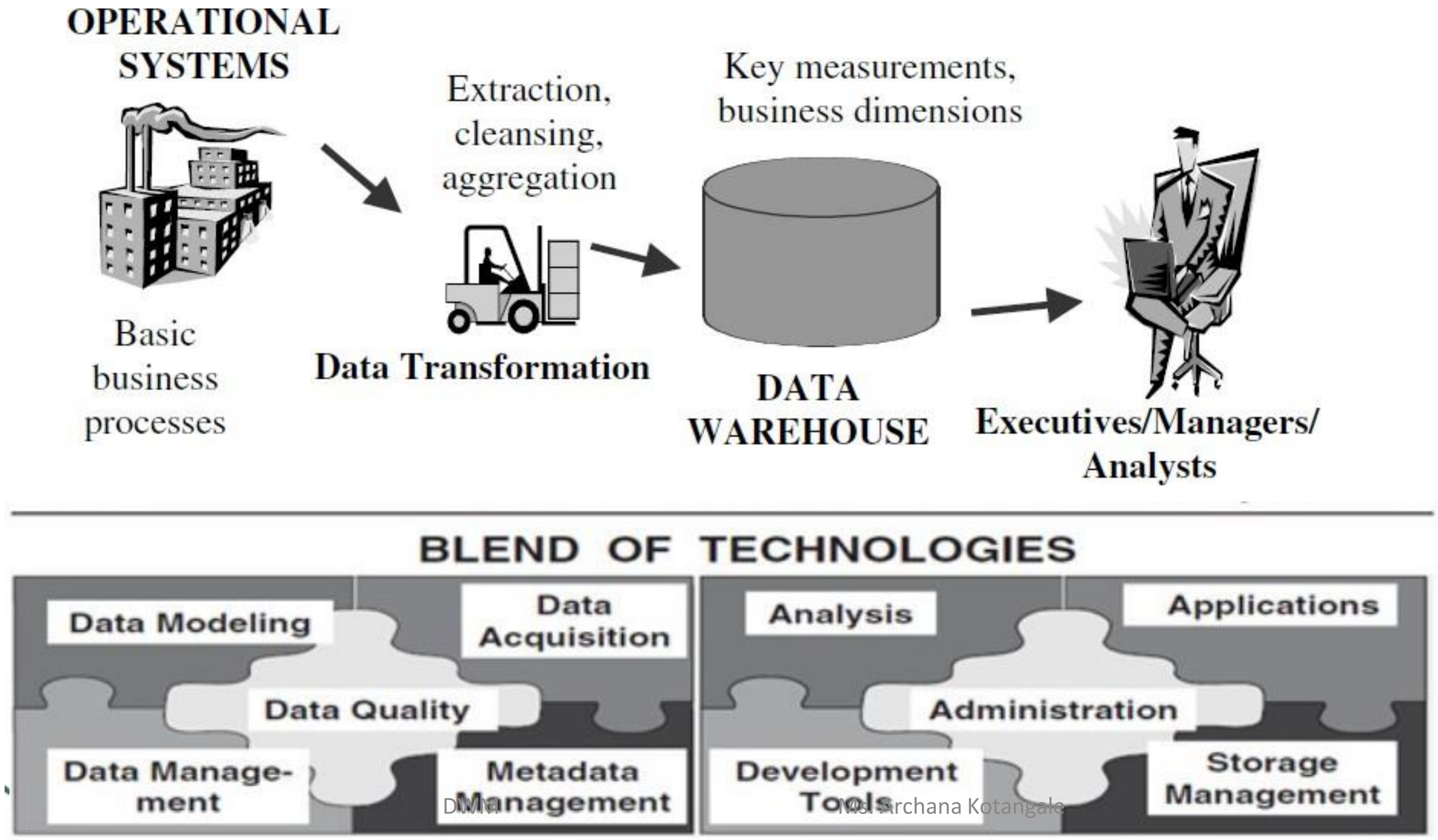
Data warehouse Defined

Data warehouse is informational environment

- Provide Provides an integrated and total view of the enterprise
- Makes the enterprise’s current and historical information easily available for decision making
- Makes decision-support transactions possible without hindering operational systems
- Renders the organization’s information consistent
- Presents a flexible and interactive source of strategic information

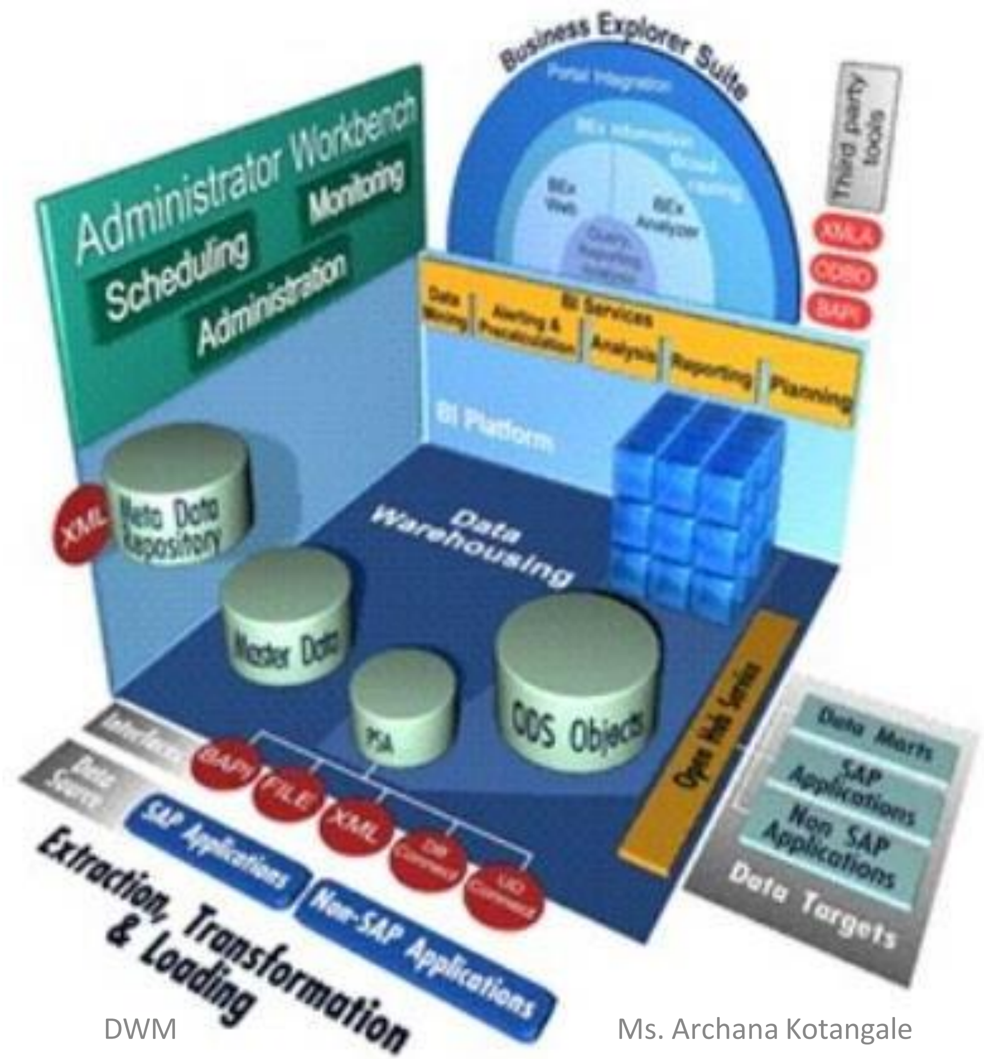
Data warehouse is environment not a product

Data warehouse is blend of many technologies



Data warehouse is environment not a product

Data warehouse is environment



Data warehouse application

- Consumer Goods
- Finance and Banking
- Government and Education
- Health Care
- Insurance
- Manufacturing and Distribution
- Automobile
- Clothing
- Pharmaceutical
- Marketing
- Sports
- Transportation
- Telecom…………… and many more

What a Data warehouse can do?

- Immediate information delivery
- Integration of data from within and outside the organization
- Provides an insight into the future
- Enables users to look at the same data in different ways
- Provides freedom from the dependency

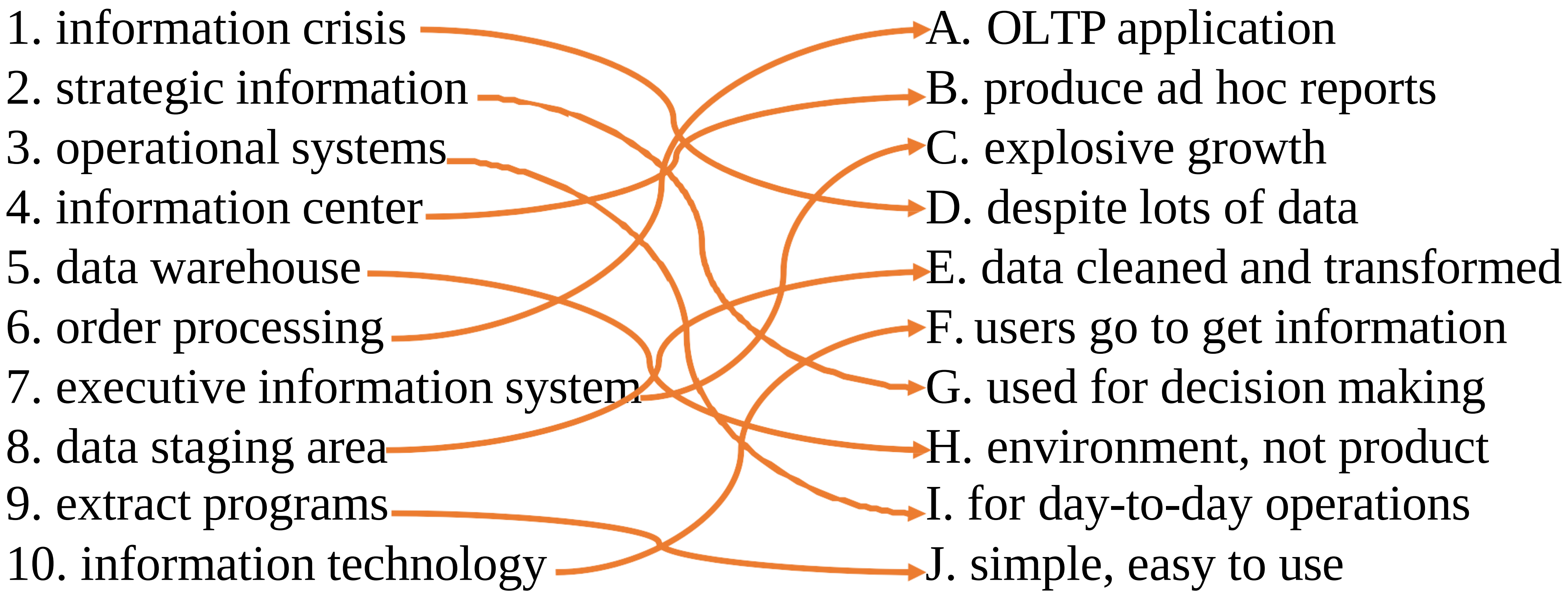
What a Data warehouse can not do?

- A data warehouse is not a magical tool; it does have some limitations
- It acts as an information repository that collects and reports data that already exists. It can not create additional data on its own
- E.g. if a manager wants to analyze the sales of a product based on customers income level, and if income of a customer is not captured by the source system then the DW will not be able to help(until and unless a mechanism is derived to gather the income data)

What a Data warehouse can not do?

- Apart from this if an organization has dirty data in the source system, the DW will not be able to correct results until and unless the data is cleaned
- DW is an environment not a product
- DW is a blend of many technologies like data modelling, data acquisition, data quality, data management, metadata management, analysis, applications, administration, development tools and storage management

Exercise



1. information crisis
2. strategic information
3. operational systems
4. information center
5. data warehouse
6. order processing
7. executive information system
8. data staging area
9. extract programs
10. information technology

A. OLTP application
B. produce ad hoc reports
C. explosive growth
D. despite lots of data
E. data cleaned and transformed
F. users go to get information
G. used for decision making
H. environment, not product
I. for day-to-day operations
J. simple, easy to use