AS PER RANI CHANNAMMA UNIVERSITY (NON CBCS), BELAGAVI – MAY 2021

BCA VI SEM

BUSINESS INTELLIGENCE

UNIT – III
BI DEFINITIONS & CONCEPTS

PRESENTATION BY :
SANTOSH S.UMADI,
DEPARTMENT OF COMPUTER SCIENCE & BCA,
BLDEA's COMMERCE, BHS ARTS & TGP SCIENCE COLLEGE,
JAMKHANDI – 587301
Email id: santoshumadi@gmail.com
Blog: santoshumadi.blogspot.com

SYLLABUS

RANI CHANNAMMA UNIVERSITY, BELAGAVI

17BCAECOT63: Business Intelligence

Teaching Hours: 4 Hrs/week **Marks: Main Exam: 50**
IA: 20

Unit I: Business View of Information Technology applications: Business Enterprise Organization , Its functions, and core business process, baldrige business excellence frame work (Optional reading) Key purpose of using IT in business, The connected world : Characteristics of Internet _Ready IT Applications, Enterprise applications(ERP/CRM) and bespoke IT applications, information users and their requirements, Types of digital data , structured data , unstructured data, Semi-structured data , Difference between semi structured and structured data. 10 Hrs

Unit II: Introduction to OLTP and OLAP : OLTP(online transaction processing) OLAP(online Analytical Processing) Different OLAP Architectures , OLTP and OLAP, Data models for OLTP and OLAP, Role of OLAP tools in the BI Architecture , should OLAP be performed directly on operational data bases. Business intelligence: Using analytical information of decision support, Information sources before dawn of BI , BI defined , evolution of BI and role of DSS , EIS, MIS and digital dash boards, Need for BI at virtually all levels , BI for past , present and future, The BI value Chain , Introduction to Business analytics. 08 Hrs

Unit III:BI definitions and concepts : BI component Framework , BI Users, Business Intelligence Applications, BI roles and responsibilities, Basics of data integration , Need for data Warehouse ,Definition of data Warehouse, ODS, Ralph Kimball's Approach vs Inmon's Approach , Goals of data warehouse, Constituents of data Warehouse , Data integration, Data integration technologies , Data Quality , Data Profiling, A case Study from the Healthcare Domain. 10 Hrs

SYLLABUS

Unit IV:Types of Data Model: Data Modelling techniques, Fact table, Dimension table, Typical dimensional Models, Dimensional Modelling Life cycle, Understanding Measures and performance measurement System terminology , navigating a Business Enterprise. 10 Hrs

Unit V:Basics of Enterprise Reporting: Reporting perspectives common to all levels of Enterprise, Report Standardization and Presentation practices, Enterprise Reporting characteristics in OLAP World , Balanced score card , Dash boards. 10 Hrs

Text Books:

1. R.N.Prasad, Seema Acharya , Fundamentals of Business analytics, First Edition , 2011, Wiley-India

Reference Books:

1. GaliShmueli,. Nitin R Patel , peter C . Bruce, " Data mining for Business Intelligence" Wiley-India, 2011.
2. Ralph Kimball ,Margy Ross, "Practical tools for Data Warehosuing and Business Intelligence" , second Edition Wiley-India 2011.

BI COMPONENT FRAMEWORK

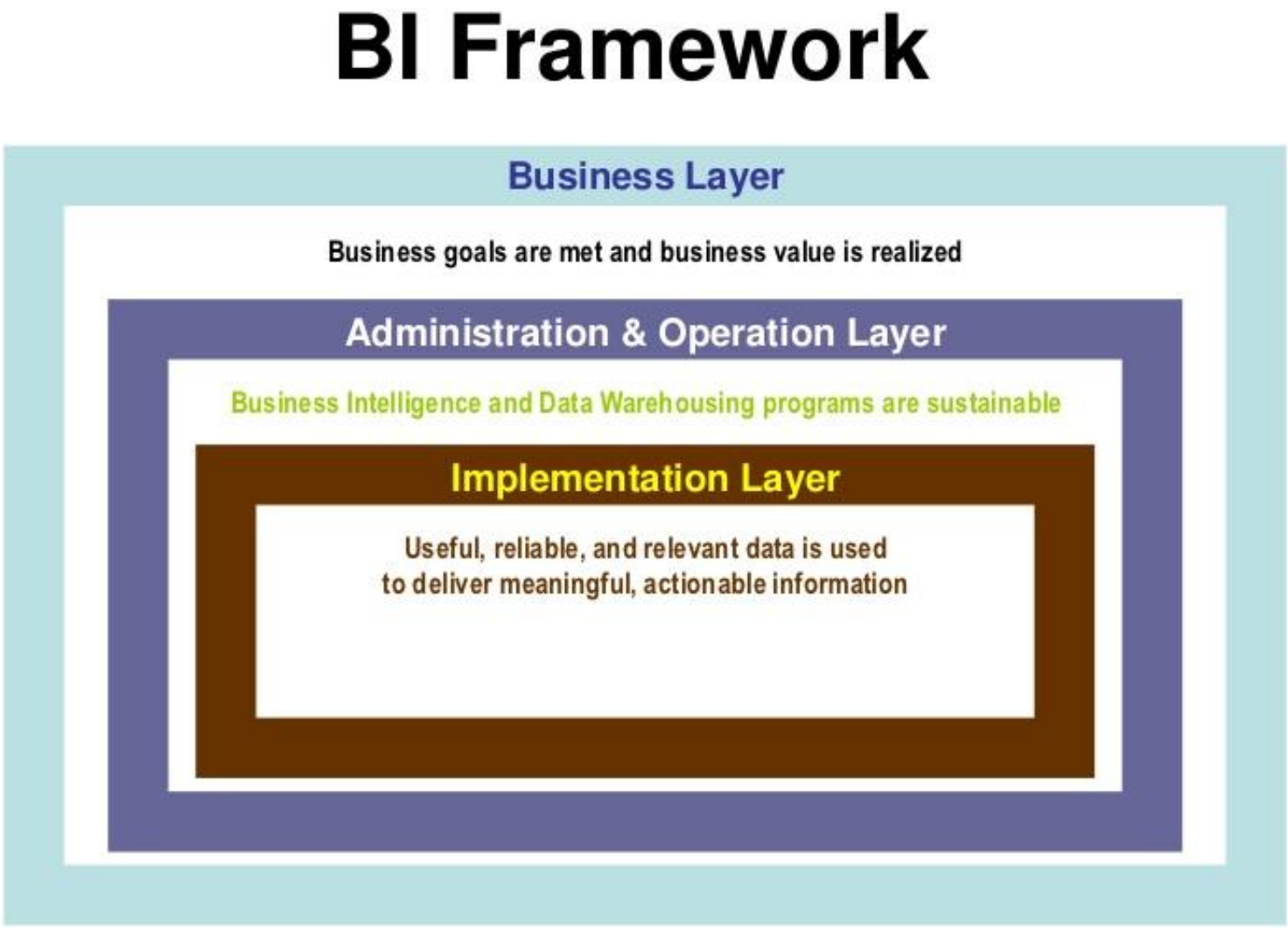
- In today's environment, the organizations which are successful are those with good architectures
- The perfect architecture supports functional, technical and data needs of the enterprise.
- The perfect architecture help the organization become better equipped to respond to the business queries posted by the users

• Business layer
• Administrator and Operation layer
• Implementation layer

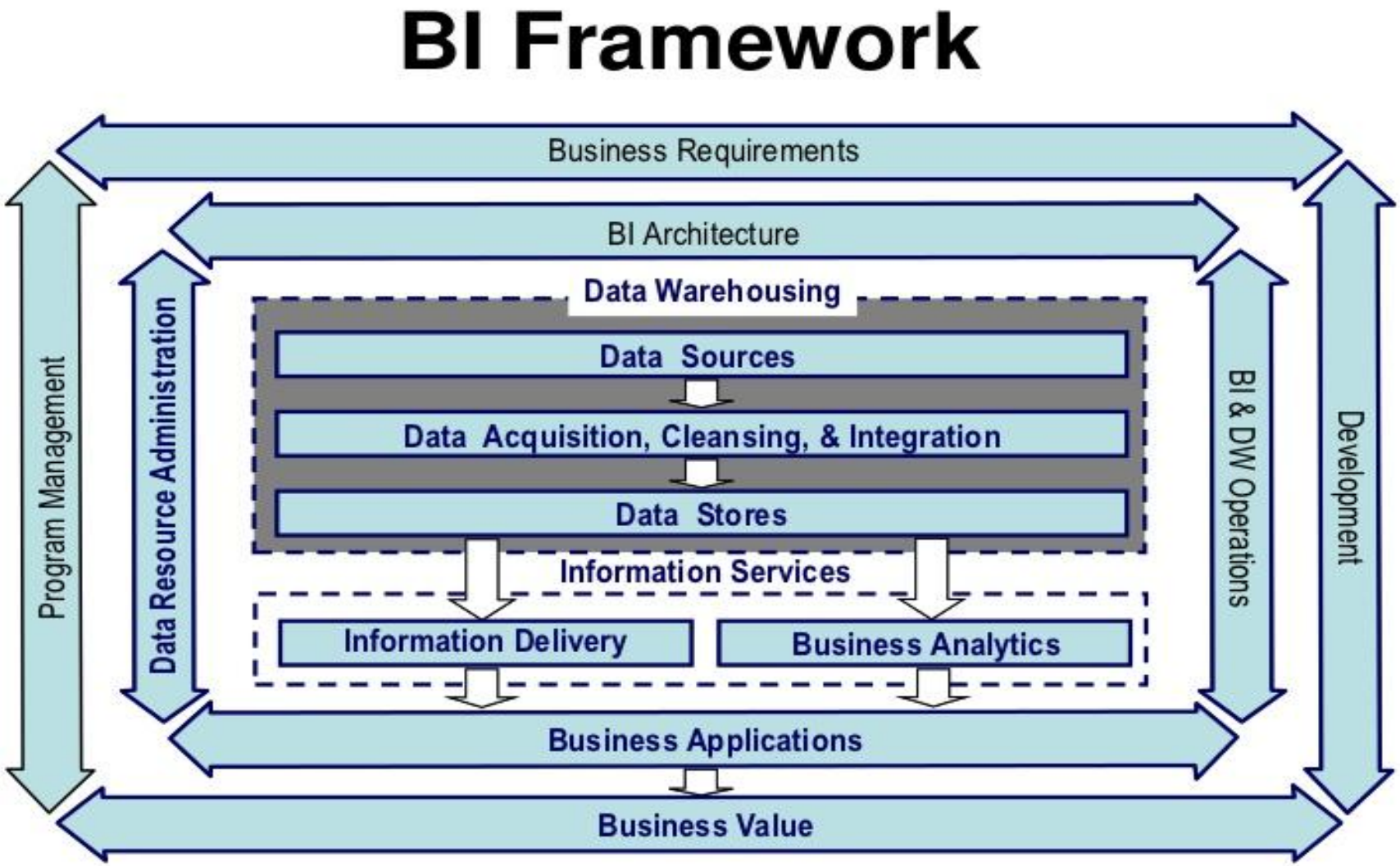
BI COMPONENT FRAMEWORK

- The business layer includes those components needed for BI to fit seamlessly into business organizations, processes and activities
- The administration and operation layer provides connections between business components and technical components
- The implementation layer comprises all technical components needed to capture data, turn data into information, and deliver that information to the business

BI COMPONENT FRAMEWORK

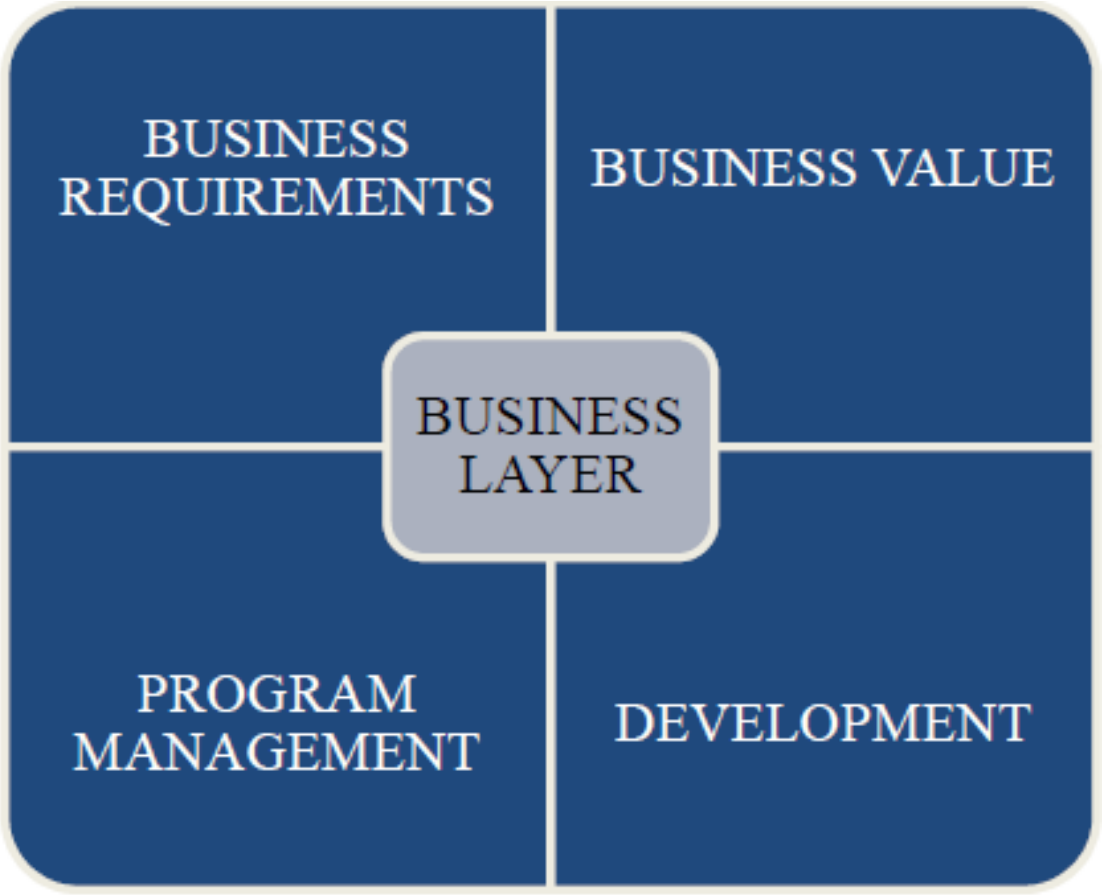


BI COMPONENT FRAMEWORK



BI COMPONENT FRAMEWORK

BUSINESS LAYER



BI COMPONENT FRAMEWORK

BUSINESS LAYER

BUSINESS REQUIREMENTS

- **Business drivers:** They help to drive the business which initiate the need to act.

  Ex: changing workforce, changing technology, changing labor laws

- **Business goals:** These are the targets to be achieved in response to business drivers.

  Ex: increased productivity, improved market share, improved profit margins, improved customer satisfaction, cost reduction

BI COMPONENT FRAMEWORK

BUSINESS LAYER

BUSINESS REQUIREMENTS

- **Business strategies:** These are the planned course of action that will help achieve the set goals

  Ex: customer retention programs, employee retention programs, competitive pricing, global delivery model.. etc

BI COMPONENT FRAMEWORK

BUSINESS LAYER

BUSINESS VALUE

- Business value of BI lies in its use within management processes that impact operational processes that drive revenue or reduce costs, and/or in its use within those operational processes themselves
- **ROI: Return on Investment:** It is a financial metric that is widely used to measure the probability of gaining a return from an investment

BI COMPONENT FRAMEWORK

BUSINESS LAYER

BUSINESS VALUE

- **ROA: Return on Asset:** It is an indicator of how profitable a company is relative to its total assets. ROA gives a manager, investor, or analyst an idea as to how efficient a company's management is at using its assets to generate earnings
- **TCO: Total Cost of Ownership:** TCO is the purchase price of an asset or product, plus the costs of operation, ie. what the product is and what its value is over time

BI COMPONENT FRAMEWORK

BUSINESS LAYER

PROGRAM MANAGEMENT

- This component of the business layer ensures that people, projects, and priorities work in a manner in which individual processes are compatible with each other so as to ensure seamless integration and smooth functioning of the entire program

Business priorities	Mission & goals	Strategies & risks
Multiple projects	Dependencies	Cost and value
Business rules	Infrastructure	

BI COMPONENT FRAMEWORK

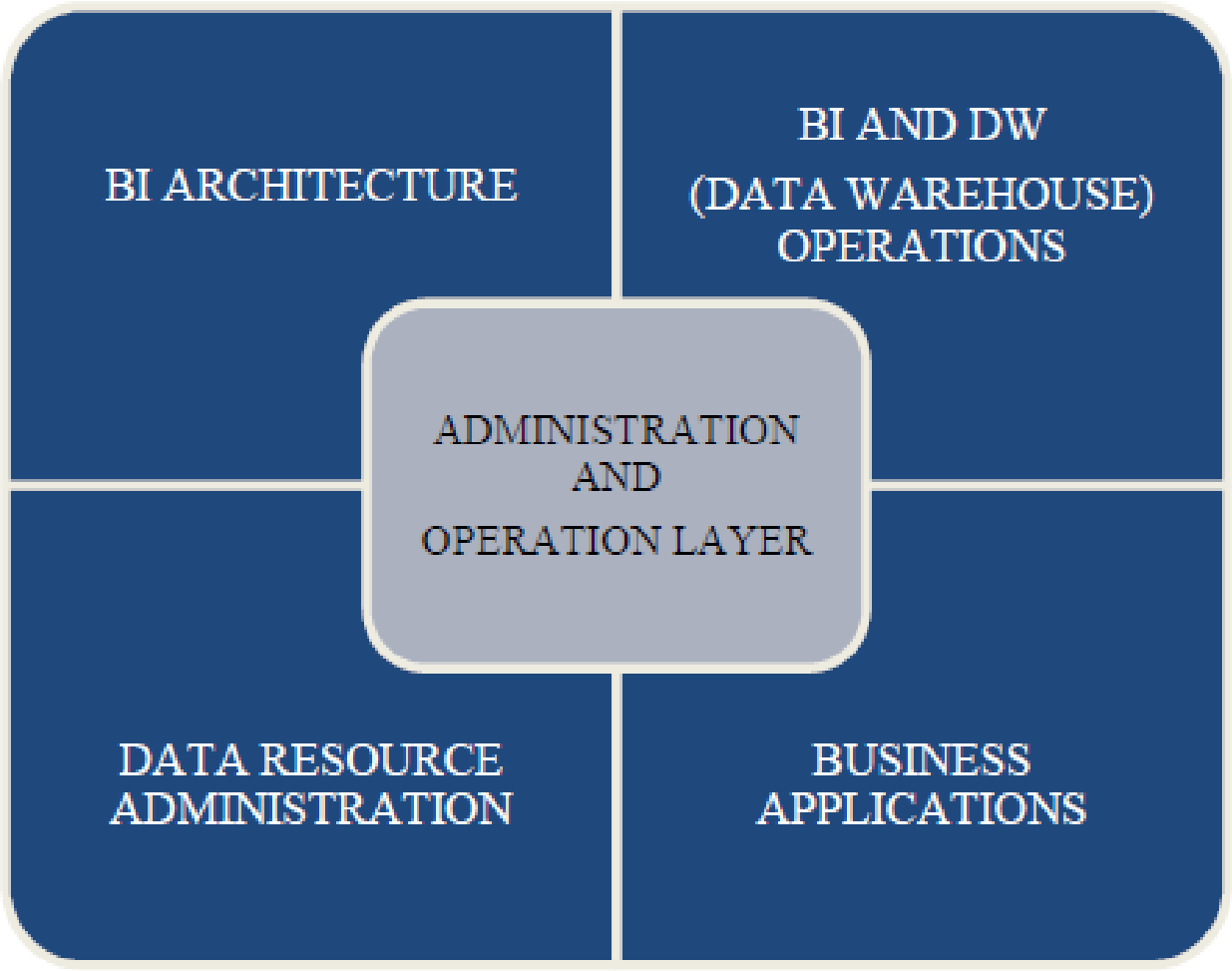
BUSINESS LAYER

DEVELOPMENT

- Database/data-warehouse development (consisting of ETL, data profiling, data cleansing and database tools)
- Data integration system development (consists of data integration tools and data quality tools)
- Business analytics development (about processes and various technologies used)

BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER



BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER

BI ARCHITECTURE

DATA	• Should follow design standards • Must have a logically apt data model • Metadata should be of high standard
INTEGRATION	• Performed according to business semantics and rules • During integration, certain processing standards have to be followed • Data must be consistent
INFORMATION	• Information derived from data that has been integrated should be usable, findable and as per the requirements
TECHNOLOGY	• Technology used for deriving information must be accessible • Also, it should have a good user-interface • Should support analysis, decision support, data and storage management
ORGANIZATION	• Consists of different roles and responsibilities, like management, development, support and usage roles

BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER

BI AND DW OPERATIONS

- Data warehouse administration requires the usage of various tools to monitor the performance and usage of the warehouse, and perform administrative tasks on it, some of tools are
- Backup & restore
- Security
- Configuration management
- Database management

BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER

DATA RESOURCE ADMINISTRATION

- **Data governance:** It is a technique for controlling data quality, which is used to assess, improve, manage and maintain information. It helps to define standards that are required to maintain data quality
- **Metadata management:** It is the data about the data. It includes

  *Business metadata *Process metadata

  *Technical metadata *Application metadata

BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER

DATA RESOURCE ADMINISTRATION

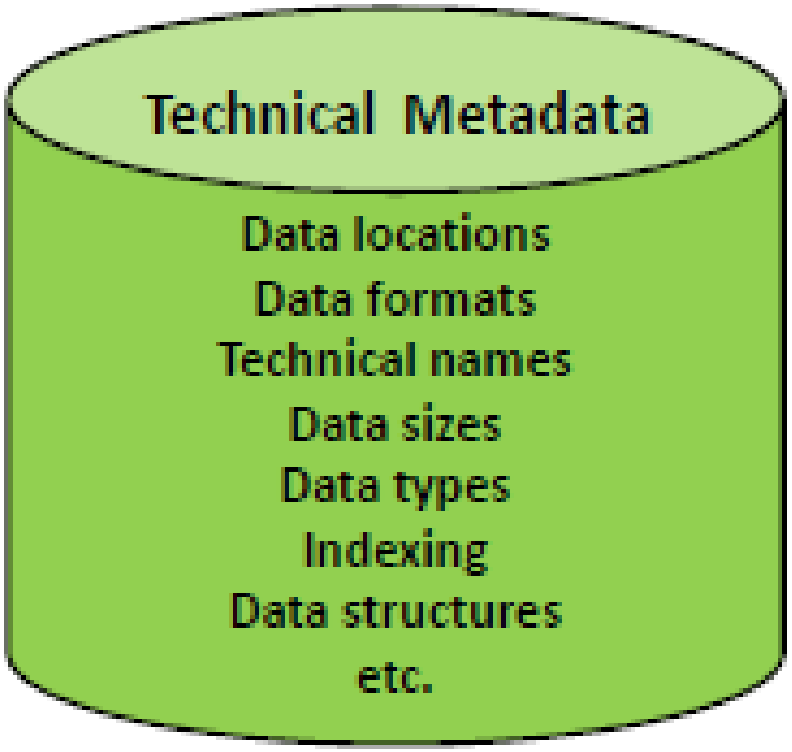
- **Metadata management:** Consider CD/DVD of music. There is the date of recording, the name of the artist, the genre of music, the songs in the album, copyright information, etc. All this information constitutes the metadata for the CD/DVD of music. In the context of a camera, the data is the photographic image. The metadata then is the date and time when the was taken. In simple words, metadata is data about data

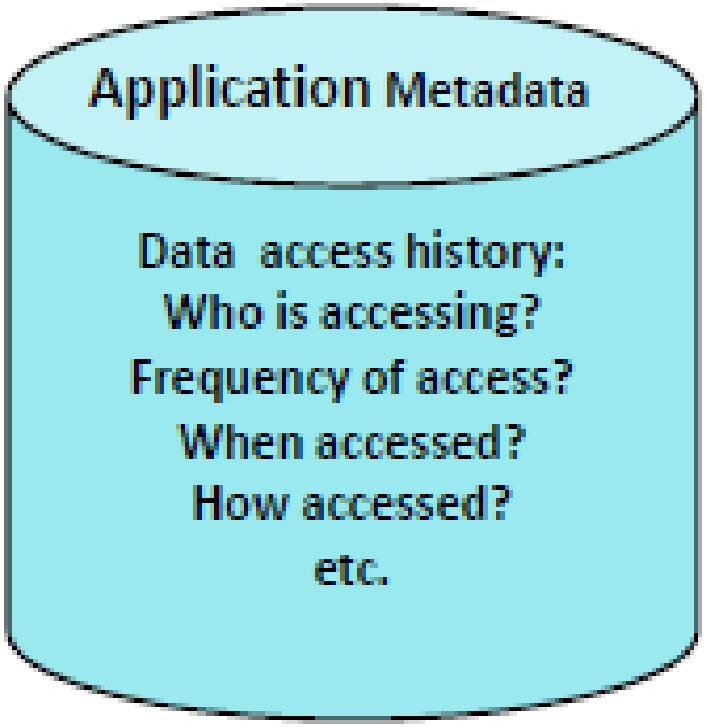
BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER

DATA RESOURCE ADMINISTRATION

- **Metadata management:**





BI COMPONENT FRAMEWORK

ADMINISTRATION & OPERATION LAYER

BUSINESS APPLICATIONS

- The application of technology to produce value for the business refers to the generation of information or intelligence from data assets like data warehouses
- Using BI tools, we can generate strategic, financial, customer, or risk intelligence
- This information can be obtained through various BI applications, such as DSS, EIS, OLAP, data mining and discovery, etc

BI COMPONENT FRAMEWORK

IMPLEMENTATION LAYER

- **The implementation layer of the BI component framework consists of technical components that are required for data capture, transformation and cleaning, data into information, and finally delivering that information to leverage business goals and produce value for the organization**

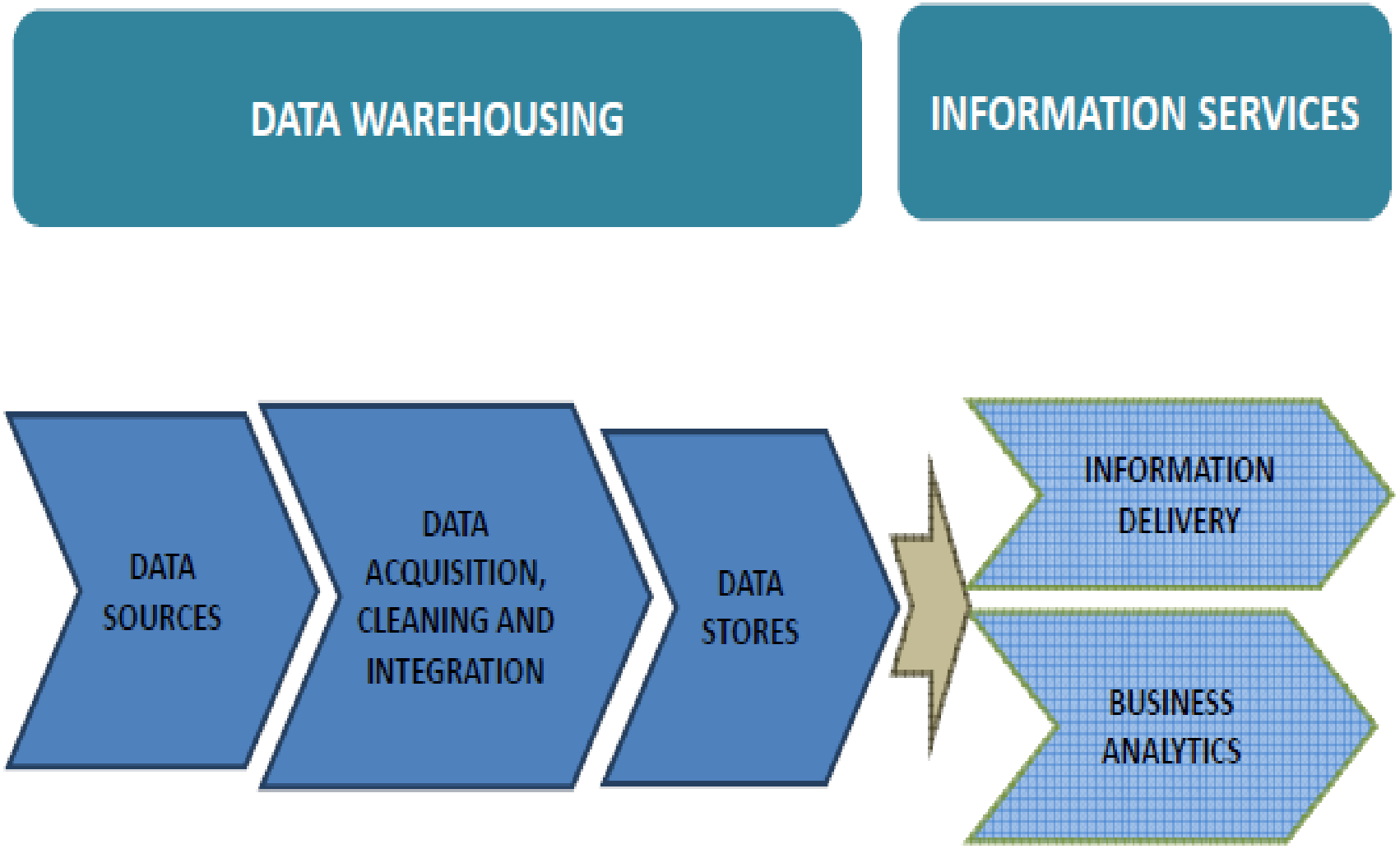
BI COMPONENT FRAMEWORK

IMPLEMENTATION LAYER

- **Data warehousing:**
  - Data Sources
  - Data Acquisition, Cleaning & Integration
  - Data Stores
- **Information Services:**
  - Information Delivery
  - Business Analytics

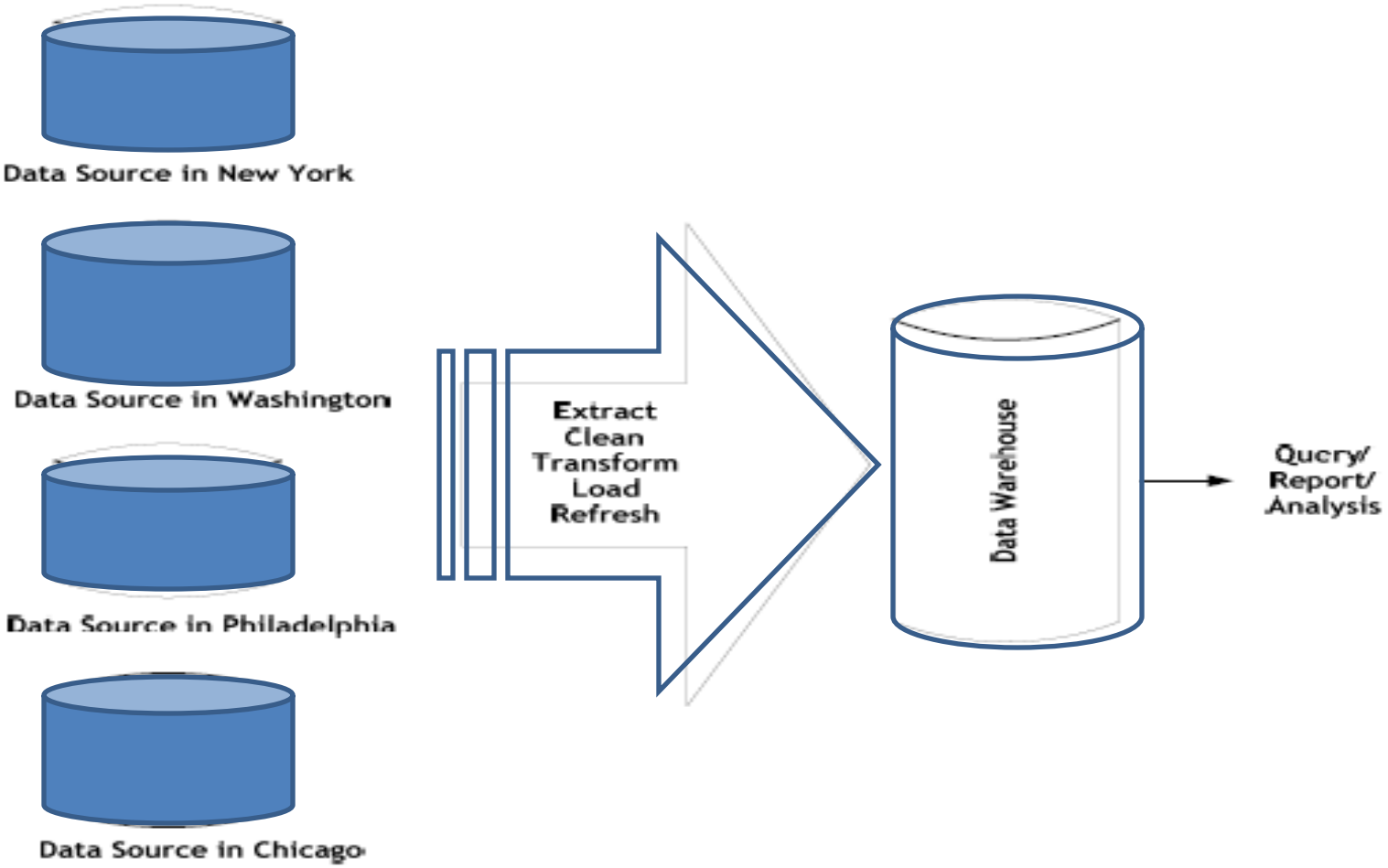
BI COMPONENT FRAMEWORK

IMPLEMENTATION LAYER

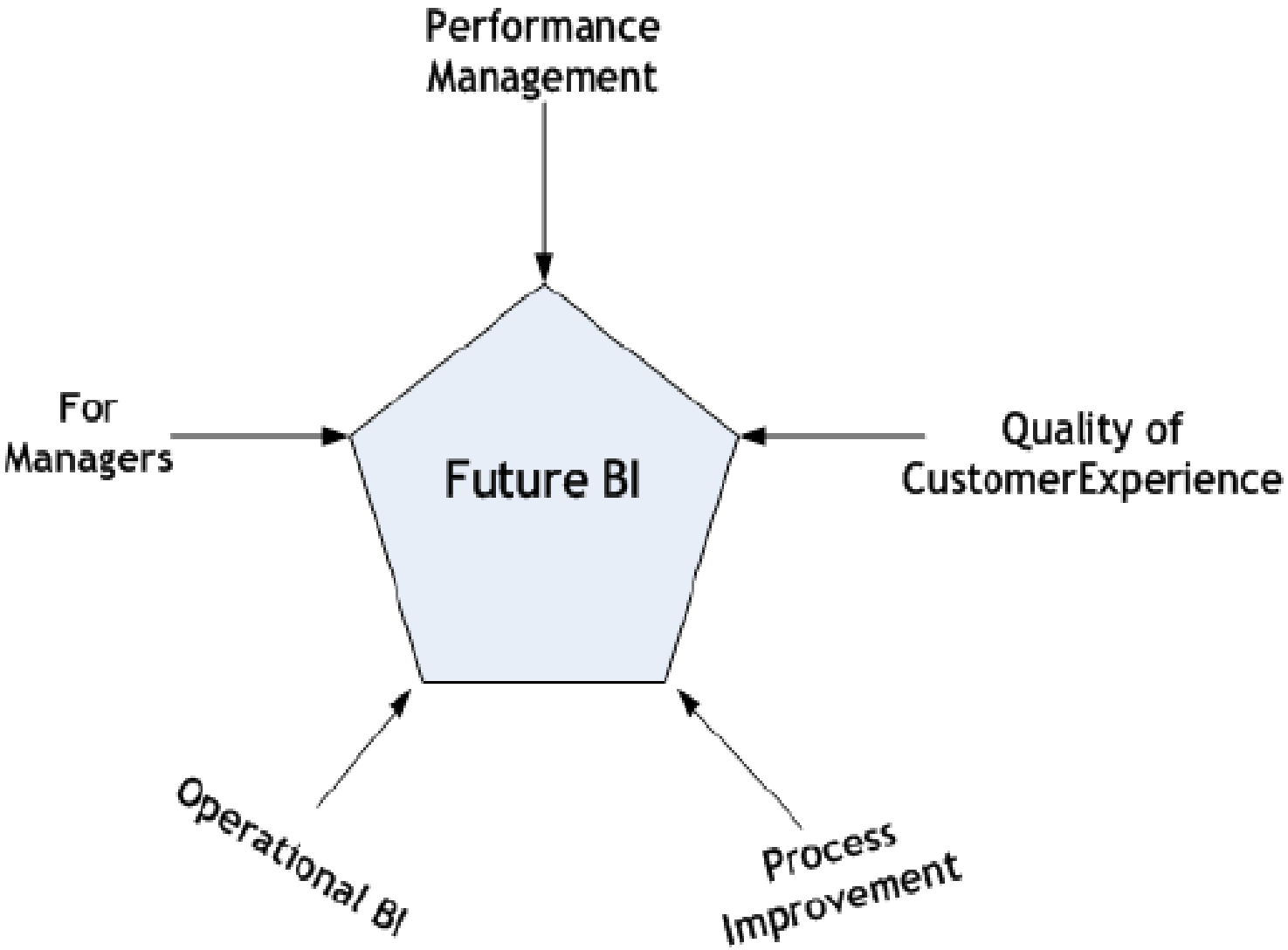


BI COMPONENT FRAMEWORK

IMPLEMENTATION LAYER



WHO IS BI FOR?


Performance Management
For Managers
Future BI
Quality of CustomerExperience
Operational BI
Process Improvement

BI FOR MANAGEMENT

- BI is very powerful tool for managers and executives
- With BI, one can have right information at the right time in the right format
- Managers need not require to wait for the reports, every data is available for them with a click of a button or on a dashboard
- BI helps report:

How sales are in the various regions?
Whether costs are exceeding budgets?
Whether customers are satisfied with the service?
What items customers buy the more?
What is it that your company is best at?

OPERATIONAL BI

- BI doesn't rely on historical data, it is helpful to perform daily operations
- Operational BI will have to interact with a transaction based system that is updated in real time several times in a day
- BI also provides reports on day to day operations
- Ex: When a customer places an order, before accepting the order, the customer service representative might want to check whether sufficient inventory is available,. For this he will look at BI reports in inventory system

BI FOR PROCESS IMPROVEMENT

- BI also leads to enhancement in the performance of enterprises
- With BI some of the processes were automated, which helps in process improvement
- Ex: Cash flow problem: In the earlier days, when goods delivered to the customers on time. The invoice was sent to the customer after about a week, this lead to cash flow problem to the company. BI had helped the company to monitor this process

BI FOR PERFORMANCE IMPROVEMENT

- For the companies, the measures for business performance are: revenue, profitability, etc
- BI helps companies in performance improvement
- Ex: BI helps companies identify customers who are regular buyers of the product, a BI systems sales history data can help determine the products that customers are likely to buy together
- A company can use BI tools to understand the customer demographics before deciding to launch a product

BI TO IMPROVE CUSTOMER EXPERIENCE

- With BI, it is possible to know customers well
- In case of online shopping, BI helps to know customers buying pattern, with this a company can attract customers with similar products
- With technology, customer is experiencing hassle free shopping
- BI help companies serve the customers better and win the customers satisfaction, loyalty and advocacy

BI USERS

- Casual users & Power users
- **Casual users:** These users are the consumers of information in pre existing reports
- Executives, managers, workers, customers are the casual users
- They might base their decisions on the acquired information
- They do not create reports
- They use the reports created by power users whenever the need arises

BI USERS

- Casual users & Power users
- **Power users:** These users are the producers of the information
- They produce information either for their own needs or to satisfy the information needs of others
- Developers, administrators, analysts, IT professionals belong to this category
- Power users take decisions on issues such as

What information should be placed on the report?
What is the best way to present the information?
Who should see what information?

BI USERS

Types of BI Users

Type of user	Casual users/ Information consumers	Power users/Information producers
Example of such users	Executives, managers, customers, suppliers, field/operation workers, etc.	SAS, SPSS developers, administrators, business analysts, analytical modelers, IT professionals, etc.
Usage	Information consumers	Information producers
Data Access	Tailor made to suit the needs of their respective role	Ad hoc/exploratory
Tools	Pre-defined reports/dashboards	Advanced Analytical/ Authoring tools
Sources	Data warehouse/Data Marts	Data Warehouse/Data Marts (both internal and external)

BI APPLICATIONS

- BI applications can be divided into
- Technological solutions:

*DSS *EIS *OLAP *Data mining

*Managed query and reporting

- Business solutions:

*Performance analysis *Customer analysis

*Marketplace analysis *Sales channel analysis

*Behavior analysis

BI APPLICATIONS

- **Technical solutions**
- **DSS: DSS stands for Decision Support System. It supports decision making at operational and tactical levels.**
- **EIS: Executive Information System supports decision making at the senior management level. It is a specialized form DSS. EIS provides graphical interface with strong reporting capabilities**
- **OLAP: Online Analytical Processing which stores the data in multidimensional form which helps in decision making**

BI APPLICATIONS

- **Technical solutions**
- **Managed query and reporting: This tool includes predefined standard reports, report designer which are essentially used by developers to create reports**
- **Data mining: Data mining is about unraveling hidden patterns, spotting trends**

BI APPLICATIONS

- **Business solutions**
- **Customer analytics: It plays a important role in predicting the customer's behaviour, his/her buying pattern. It helps capture data about customer's behaviour**
- **Marketplace analysis: It helps understand the marketplace better. It is about understanding the customers, competitors, the products, the changing market dynamics**
- **Performance analysis: This analysis facilitates optimum utilization of utilization of employees, finance & resources, etc.**

BI APPLICATIONS

- **Business solutions**
- **For ex: when you do the performance analysis of your employees, you will get to know about the employees that you want to retain, the employees whom you want to reward, etc**
- **Behavior analysis: This analysis predict trends such as purchasing patterns, online buying patterns, etc**
- **Sales channel analysis: This analysis will help decide the best channel for reaching out your product/services for use by the customers**

BI ROLES & RESPONSIBILITIES

- BI roles can be broadly classified into two categories
- Program roles & Project roles
- For a BI project to succeed, one requires senior responsible owner. It should be backed up by senior leadership of the organization, owner is responsible to allocate the funds
- The program team concentrates on implementing the strategy how the BI project will execute, also responsible for coordination and integration

BI ROLES & RESPONSIBILITIES

Program Roles	Project Roles
	Business Manager
BI Program Manager	BI Business Specialist
BI Data Architect	BI Project Manager
BI ETL Architect	Business Requirements Analyst
BI Technical Architect	Decision Support Analyst
Metadata Manager	BI Designer
BI Administrator	ETL Specialist
	Data Administrator

NEED FOR DATA WAREHOUSE

- The Institute of Information Technology (IIT) is an engineering institution that conducts engineering courses in Information Technology (IT), Computer Science (CS), System Engineering (SE), Information Science (IS), etc
- Each department (IT, CS, SE, IS, etc.) has an automated library that meticulously handles library transactions and has good learning content in the form of DVDs, magazines, journals, several online references, etc.

PICTURE THIS

NEED FOR DATA WAREHOUSE

- The only downside of the library data is that it is stored differently by different departments. One department stores it in MS Excel spreadsheets, another stores it in MS Access database, and yét another department maintains a .CSV (Comma Separated Values) file
- The IIT administration is in need of report that indicates the annual spending on library purchases

PICTURE THIS

NEED FOR DATA WAREHOUSE

- The report should further drill down to the spending by each department by category (books, CDs/DVDs, magazines, journals, etc.). However, preparing such a report is not easy because of different data formats used by different departments
- An expert on database technology was called upon to suggest a possible solution to the problem at hand. He feels it would be better to start archiving the data in a data warehouse

PICTURE THIS

NEED FOR DATA WAREHOUSE

- Data from several heterogeneous data sources (MS Excel spreadsheets, MS Access CSV file, etc.) can be extracted and brought together in a data warehouse
- Even when IIT expands into several branches in multiple cities, it still can have one ware-house to support the information needs of the institution
- Data anomalies can be corrected through an ETL package
- Missing or incomplete records can be detected and duly corrected

NEED FOR DATA WAREHOUSE

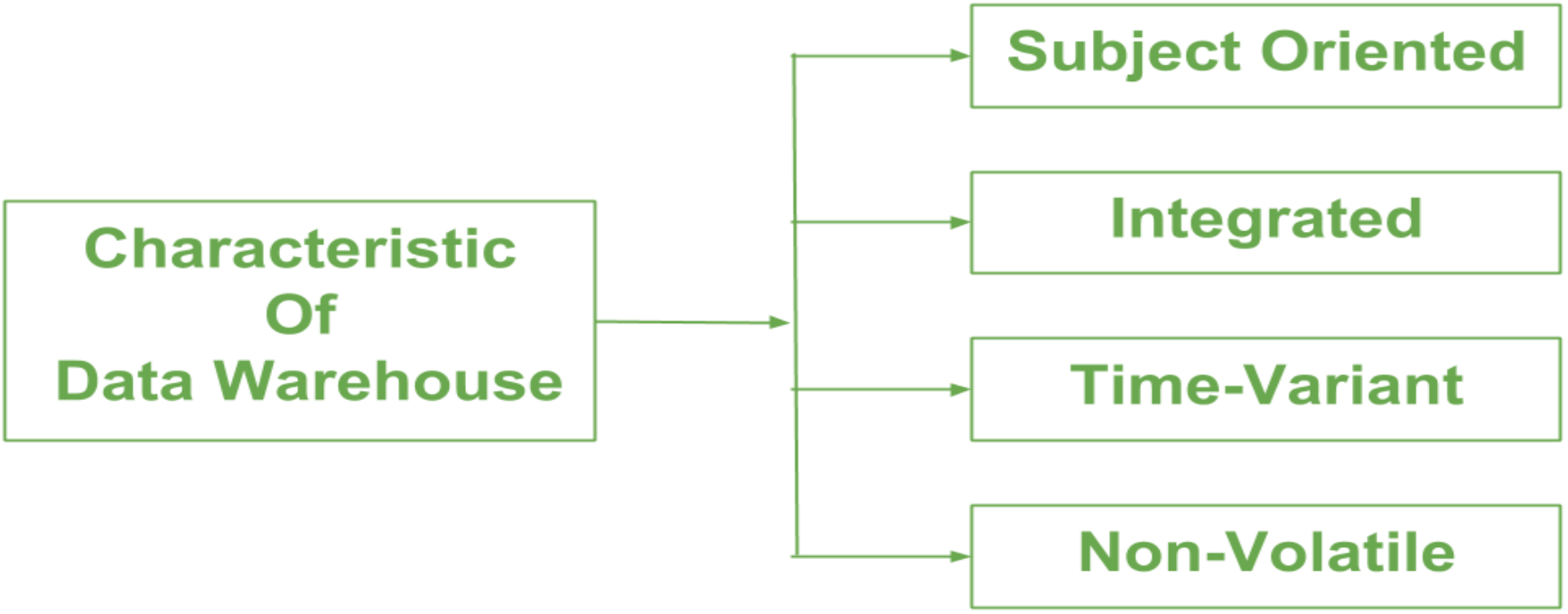
- Uniformity can be maintained over each attribute of a table
- Data can be conveniently retrieved for analysis and generating reports
- Fact-based decision making can be easily supported by a data warehouse
- Ad hoc queries can be easily supported

CAN HAVE SOLUTIONS FOR FOLLOWINGS WITH DATA WAREHOUSE

- Lack of information sharing
- Lack of information credibility
- Reports take a longer time to be prepared
- Little or no scope for ad hoc querying or queries that require historical data

DEFINING DATA WAREHOUSE

- According to William H. Inmon, "A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process



DEFINING DATA WAREHOUSE

SUBJECT-ORIENTED

- A data warehouse collects data of subjects such as “customers”, “suppliers”, “partners”, “sales”, “products”, etc. spread across the enterprise or organization
- A data mart on the other hand deals with the analysis of a particular subject such as “sales”

DEFINING DATA WAREHOUSE

INTEGRATED

- **A data warehouse serve to bring together the data from the multiple disparate (meaning differing in the format and content of data) sources after careful cleansing and transformation into a unified format to serve the information needs of the enterprise**

DEFINING DATA WAREHOUSE

TIME - VARIANT

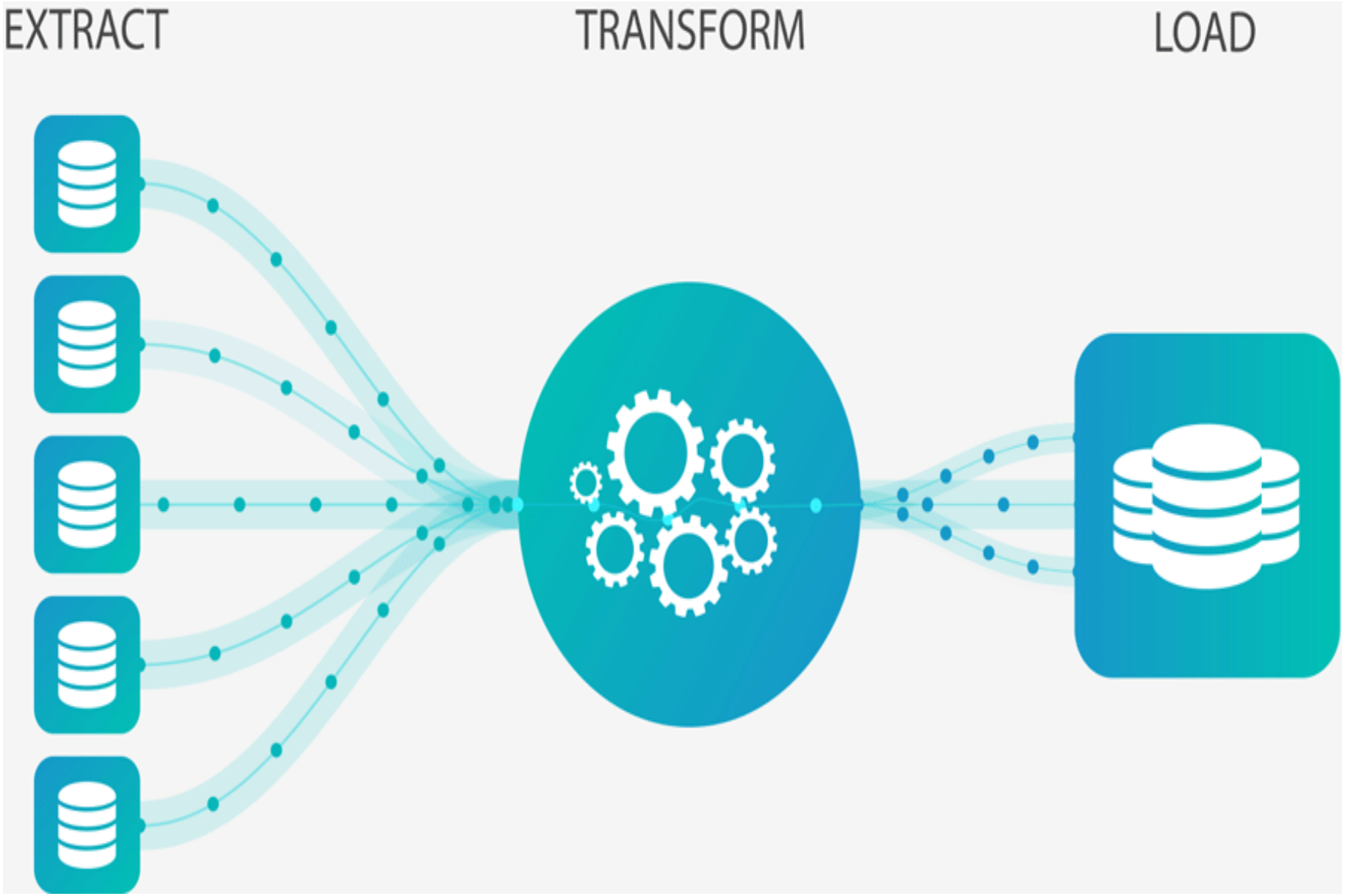
- A data warehouse keeps historical data
- From a data warehouse, one can retrieve data that is 3 months, 6 months, 12 months, or even older
- For example, a system may hold the most recent address of a customer, whereas a data warehouse addresses associated with a customer recorded, say, over the last five years

DEFINING DATA WAREHOUSE

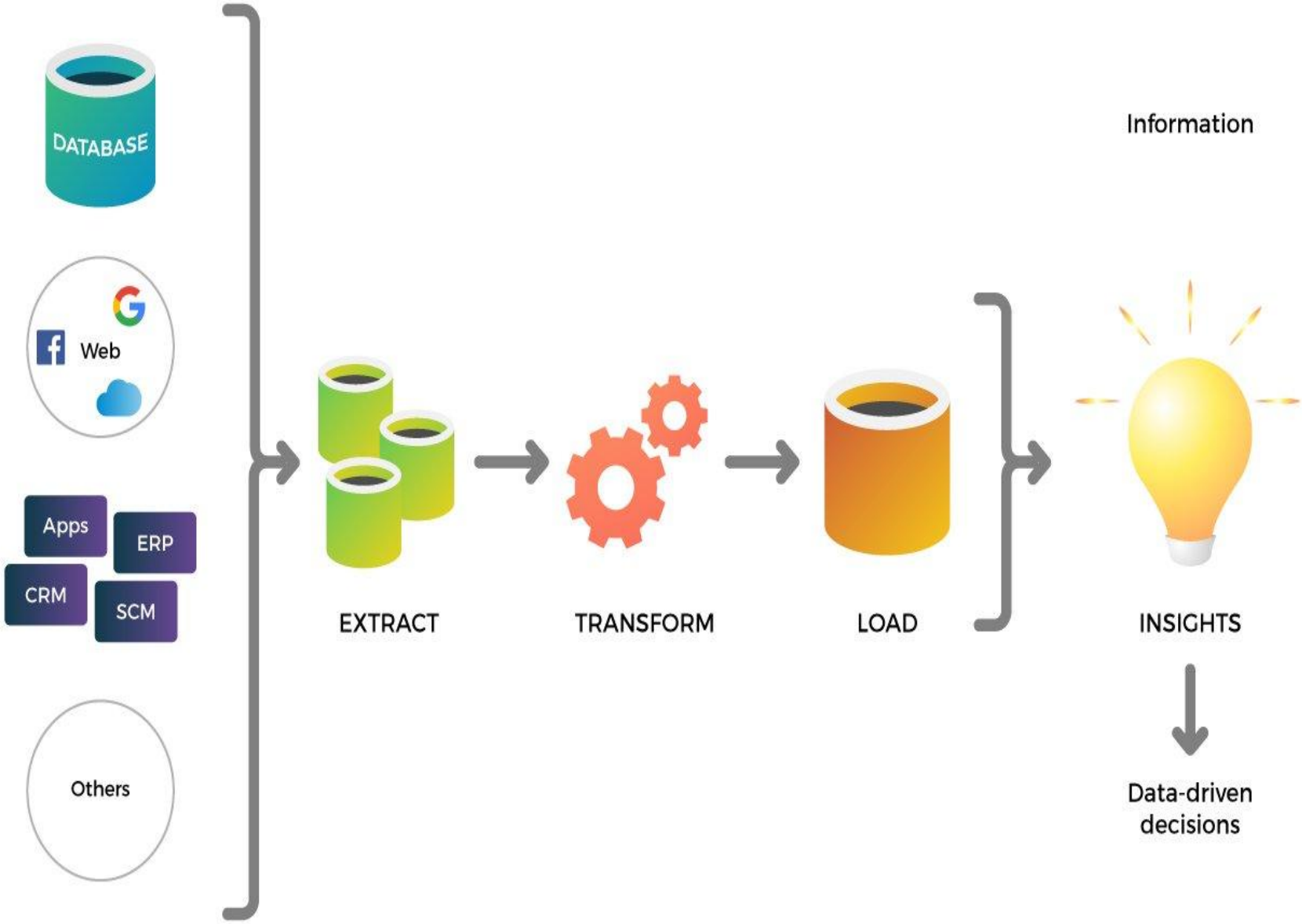
NON - VOLATILE

- We have learnt earlier that transaction processing, recovery, and concurrency control mechanisms are usually associated with OLTP systems
- A data warehouse is a separate physical store of data transformed from the application data found in the operational environment

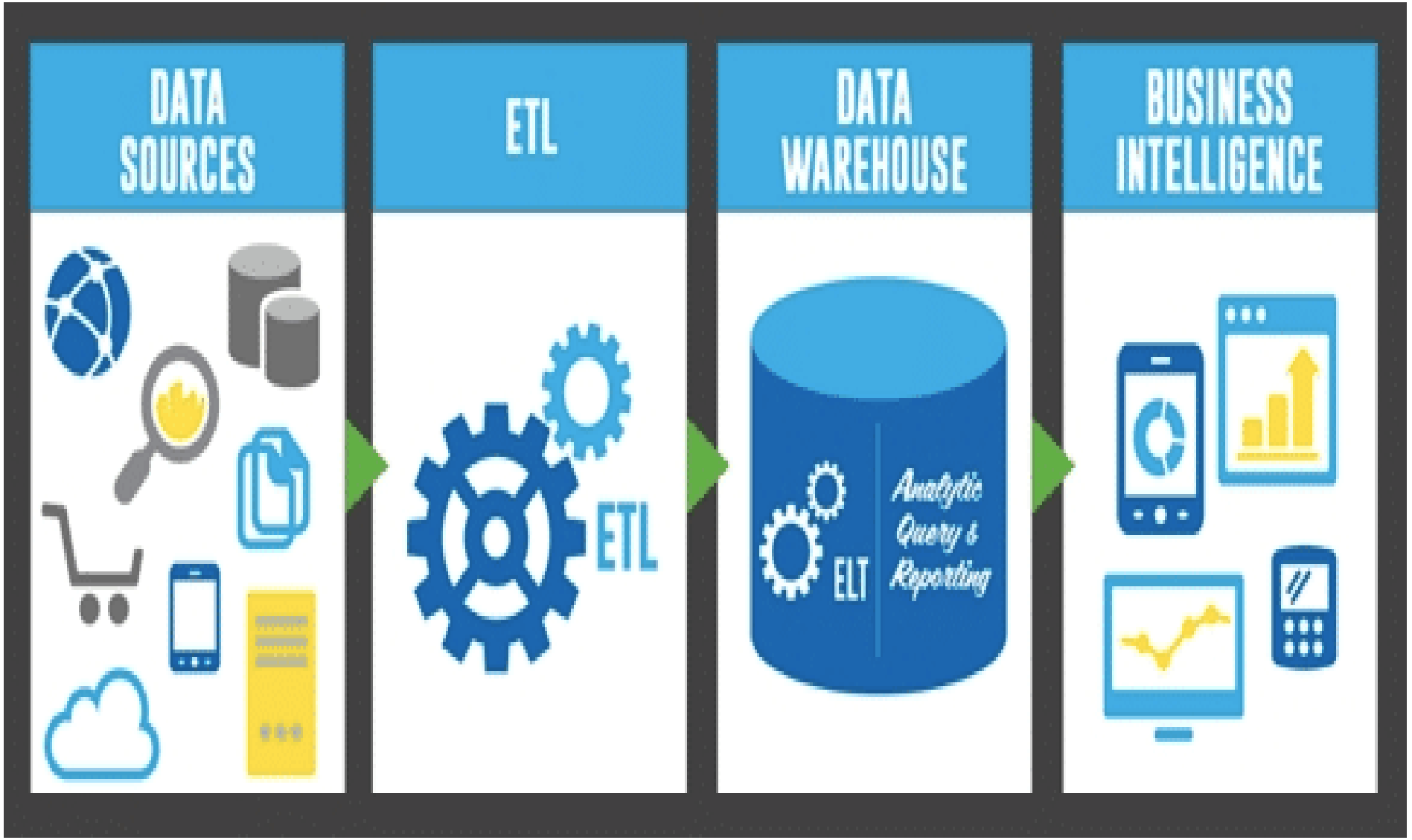
NEED FOR DATA WAREHOUSE



NEED FOR DATA WAREHOUSE



NEED FOR DATA WAREHOUSE



GOALS OF DATA WAREHOUSE

- **Information Accessibility :** Data in a data warehouse must be easy to comprehend, both by the business users and developers alike. It should be properly labeled to facilitate easy access. The business users should be allowed to slice and dice the data in every possible way
- **Information Credibility :** The data in the data warehouse should be credible, complete, and of desired quality

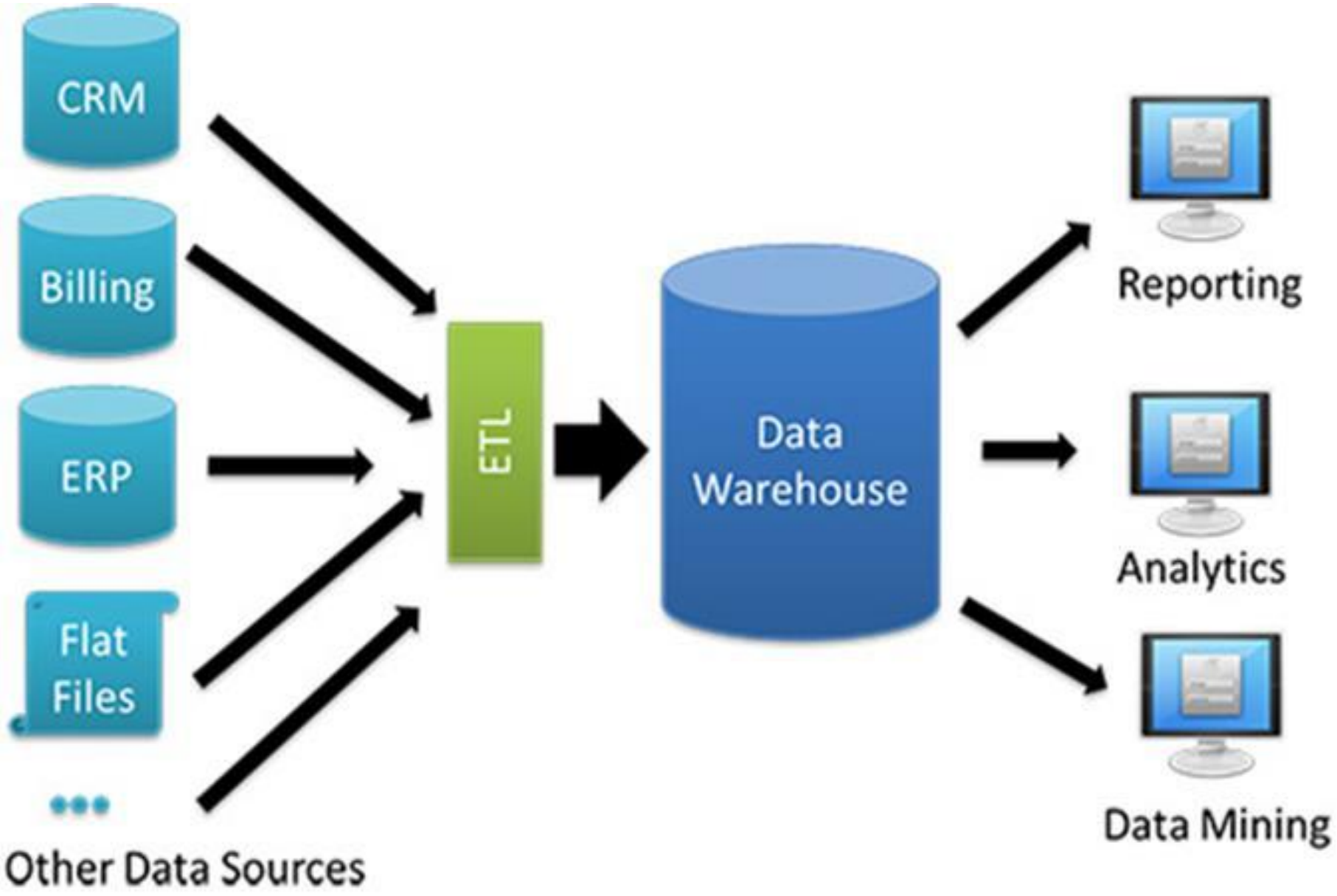
GOALS OF DATA WAREHOUSE

- **Flexible to change : Business situations change, users' requirements change, technology changes, and tools to access data may also change. The data warehouse must be adaptable to change**
- **Support for fact-based decision making : The data warehouse should have enough data to support more precise decision making. What is also required is that the business users should be able to access the data easily**

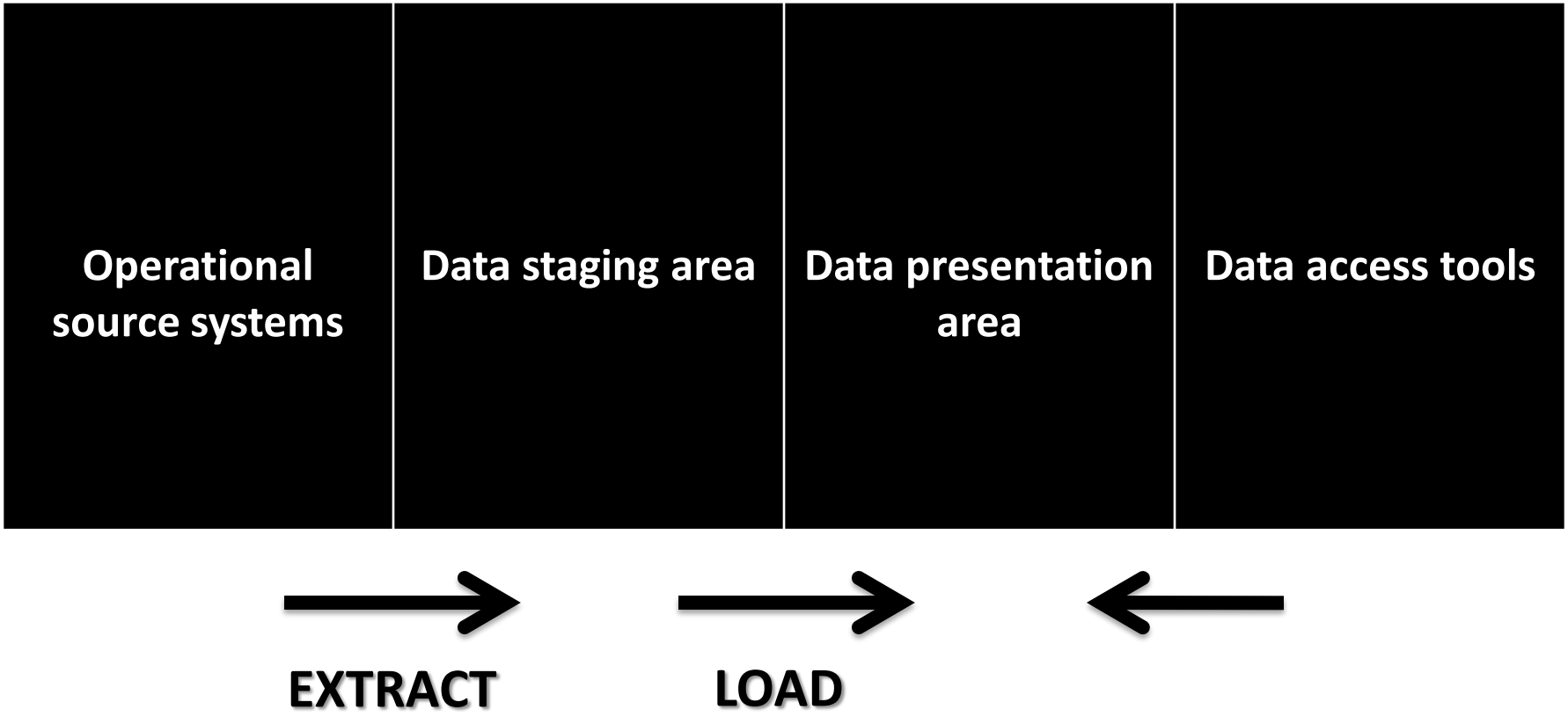
GOALS OF DATA WAREHOUSE

- **Support for the data security : The data warehouse maintains the company's confidential information. This information falling into wrong hands will do more damage than not a data warehouse at all**
- **Information consistency : Information consistency is about a single/consistent version of truth. Users from across the organization make use of the data warehouse to view a single and consistent version of truth**

WHAT CONSTITUTES A DW



WHAT CONSTITUTES A DW



WHAT CONSTITUTES A DW

- **Operational source systems:** These systems maintain transactional or operational data. They are outside DW. There could be any number of such systems feeding data to the DW. They may maintain little historical data
- **Data staging area:** The staging area comprises storage space for the data that has been extracted from various operational sources. It also consists of a set of processes related to data quality

WHAT CONSTITUTES A DW

- **Data presentation area:** Data presentation area is the interface or the front face of the DW with which the business community interacts via the data access tools
- **Data access tools:** Data access tools can be ad hoc query tools used to query the data presentation area. A data access tool can also be a reporting tool

KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

- Designing a Data Warehouse is an essential part of business development
- For designing, there are two most common architectures named Kimball and Inmon but question is which one is better, which one serves user at low redundancy

KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

- Kimball approach of designing a data warehouse was introduced by Ralph Kimball
- This approach starts with recognizing business process and questions that data warehouse has to answer
- These sets of information are being analyzed and then documented well
- The ETL software brings all data from multiple data sources called data marts and then is loaded into a common area called staging, Then this is transformed into OLAP cube

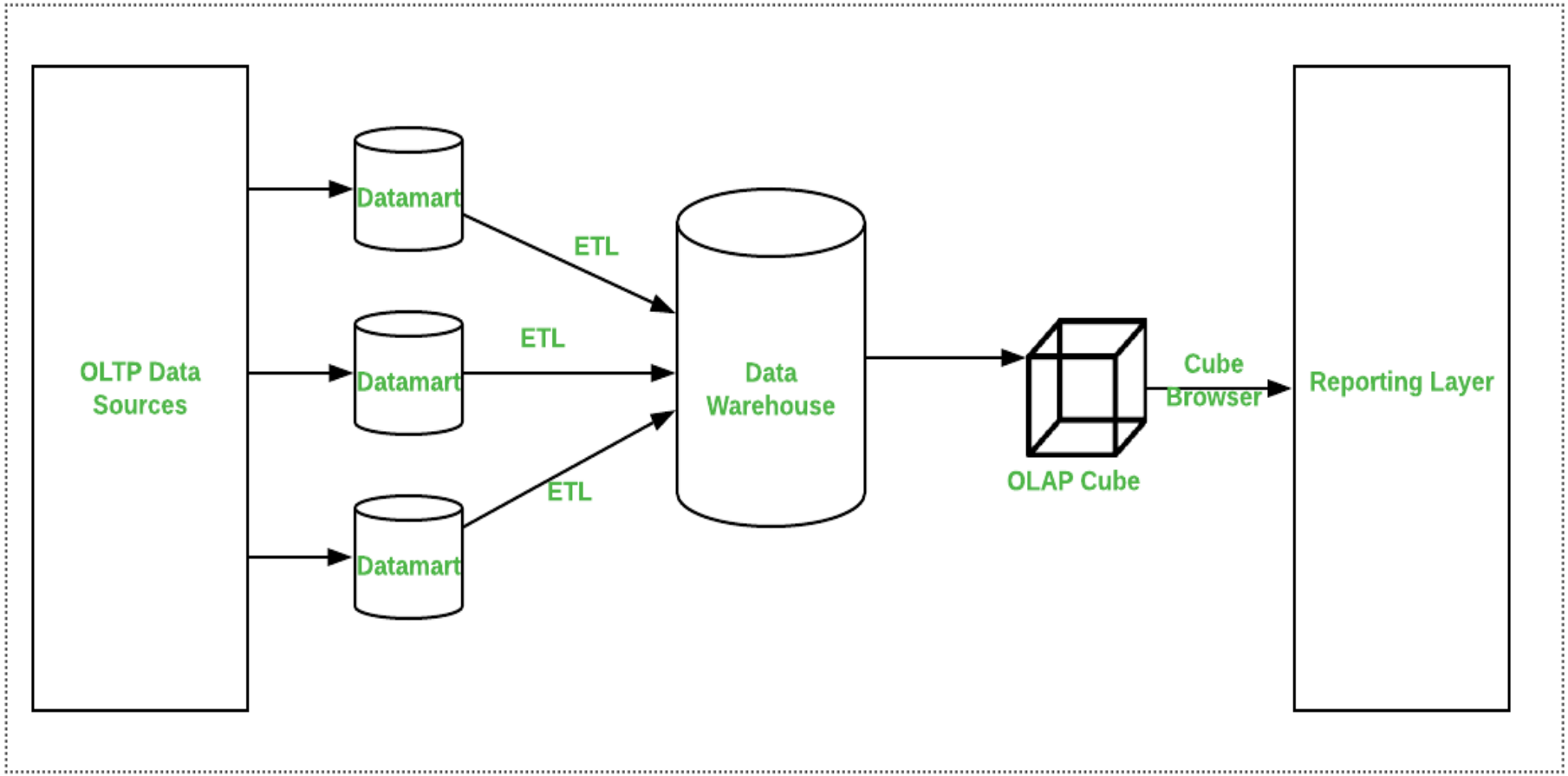
KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

Applications of Kimball approach

- Setup and Built is quick
- Generating report against multiple star schema is very successful
- Database operation are very effective
- Occupies less space in database and management is easy

KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

Kimball approach



Kimball Model

KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

- Inmon approach of designing a data warehouse was introduced by Bill Inmon
- This approach starts with corporate data model. This model recognizes key areas and also takes care of customer, product, and vendor
- This model serves for creation of a detailed logical model which is used for major operations
- Details, model is then used to develop a physical model

KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

- **Inmon approach**
- **This physical model is normalized and makes data redundancy less**
- **This is a complex model that is difficult to be used for business purposes for which data marts are created and each department is able to use it for their purposes**

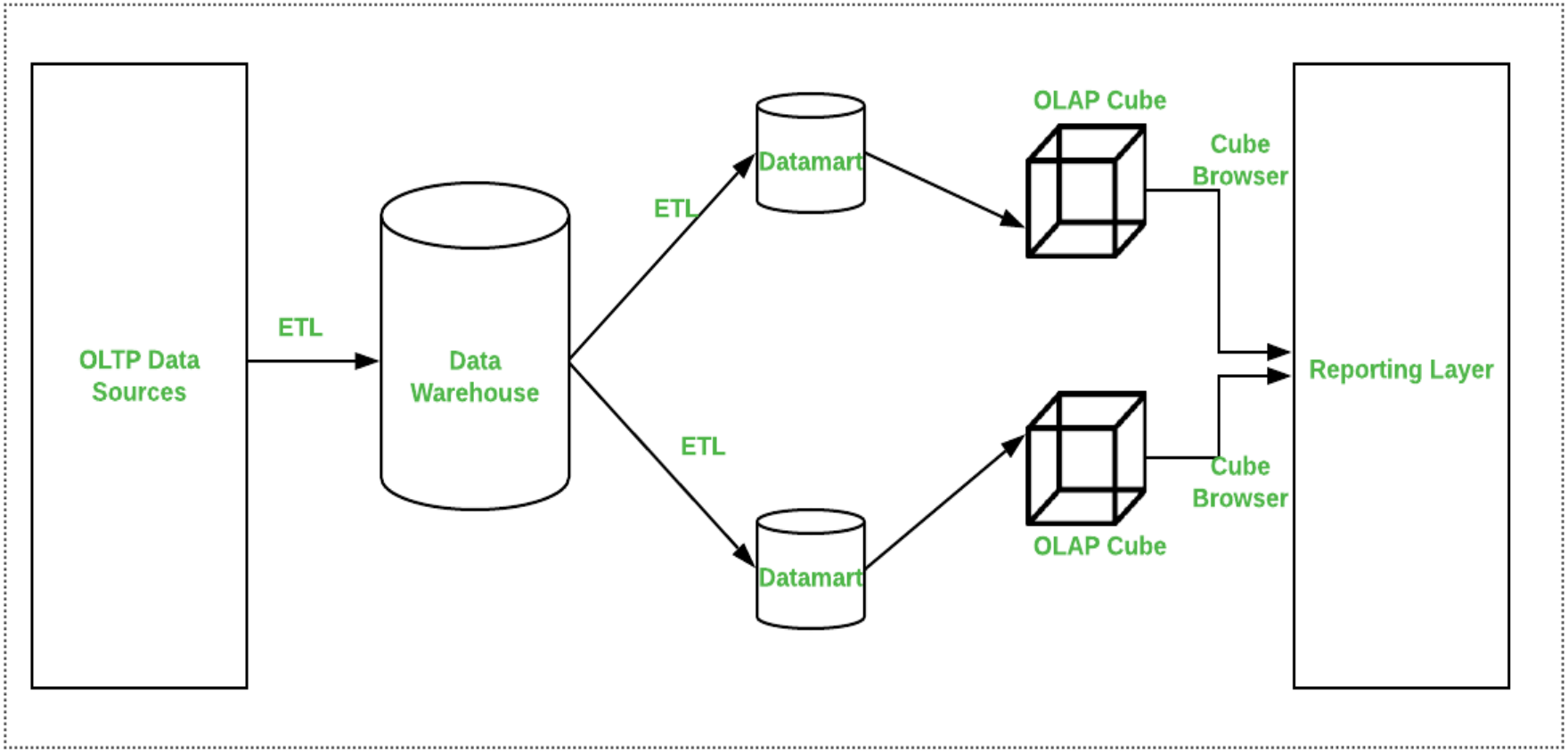
KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

Applications of Inmon approach

- The data warehouse is very flexible to changes
- Business process can be understood very easily
- Reports can be handled across enterprise
- ETL process is very less prone to errors

KIMBALL VS. INMON DATA WAREHOUSE ARCHITECTURES

Inmon approach



Inmon Model

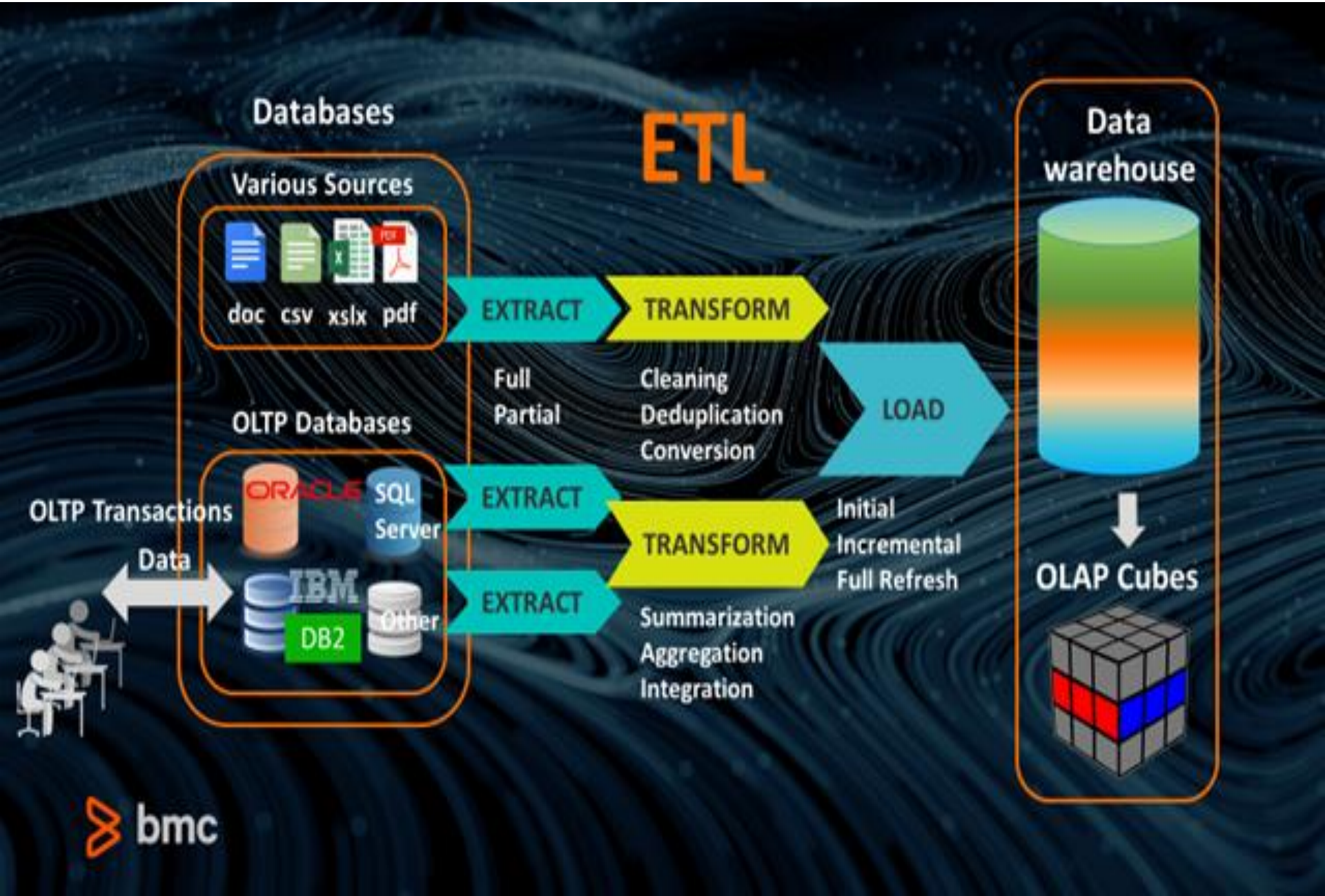
ETL : EXTRACT, TRANSFORM & LOAD

- ETL is a three stage process in database usage, especially in data warehousing. It allows integration and analysis of data stored in different sources
- After collecting the data from multiple sources (extraction), the data is reformatted (from host format to warehouse format) and cleansed (to detect and rectify errors) to meet the information needs (transformation) and then sorted, summarized, consolidated, and loaded into desired end target (loading)

ETL : EXTRACT, TRANSFORM & LOAD

- **ETL TOOLS**

ETL : EXTRACT, TRANSFORM & LOAD



ETL : EXTRACT, TRANSFORM & LOAD

- ETL allows creation of efficient and consistent databases

- So we can say, ETL is
- Extracted data from different data sources
- Transforming the extracted data into a relevant format to fit information needs
- Loading data into the final target database, usually a data warehouse

ETL : EXTRACT, TRANSFORM & LOAD

DATA EXTRACTION

- Extraction is the operation of extracting data from the source system for further use in a data warehouse environment. This the first step in the ETL process
- Depending upon the type of source data, the complexity of extraction may vary. The storage of intermediate version of data is very necessary. This data is required to be backed up and archived

ETL : EXTRACT, TRANSFORM & LOAD

DATA TRANSFORMATION

- A series of rules/functions is applied to the data extracted from the source to obtain derived data that is loaded into the end target
- Depending upon the data source, manipulation of data may be required. If the data source is good, its data may require very less transformation and validation. But data from some sources might require one or more transformation types to meet the operational needs and make data fit in the end target

ETL : EXTRACT, TRANSFORM & LOAD

DATA TRANSFORMATION

Some transformation types are:

- Selecting only certain columns to load
- Summarizing multiple rows of data
- Splitting a column into multiple columns
- Joining together data derived form multiple sources
- Deriving a new calculated value

ETL : EXTRACT, TRANSFORM & LOAD

DATA LOADING

- The load phase loads the data into the end target, usually the DW. Depending on the requirements of the organization, this process varies widely
- The timing and scope to replace or append into the DW are strategic design choices dependent on the time available and the business needs
- More complex systems can maintain a history of all changes to the data loaded in the DW

DATA INTEGRATION

DATA INTEGRATION

- It is the integration(combining) of data present in different sources for providing a unified view of the data
- It is the ability to consolidate data from several different sources while maintaining the integrity and reliability of the data
- Data integration becomes increasingly important in cases of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets

DATA INTEGRATION

- Probably the most well known implementation of data integration is building an enterprise's data warehouse
- The benefit of a data warehouse enables a business to perform analyses based on the data in the data warehouse. This would not be possible to do on the data available only in the source system. The reason is that the source systems may not contain corresponding data

DATA INTEGRATION

Significant in a variety of situations; both

- Commercial (e.g., two similar companies trying to merge their database)
- Scientific (e.g., combining research results from different bioinformatics research repositories)

DATA INTEGRATION

SCHEMA INTEGRATION

- Reconciles schema elements
- Multiple data sources may provide data on the same entity type. The main goal is to allow applications to transparently view and query this data as one uniform data source, and this is done using various mapping rules to handle structural differences
- Schema integration is developing a unified representation of semantically similar information, structured and stored differently in the individual databases

DATA INTEGRATION

SCHEMA INTEGRATION

PROBLEM

CustID	TransactionID	ProductID	UnitQuantity
C101	T1001	P1010	10

CustomerNumber	TransactionID	ProductID	UnitQuantity
C201	T1007	P1111	22

SOLUTION

CustID	TransID	ProductID	UnitQuantity
C101	T1001	P1010	10
C201	T1007	P1111	22

DATA INTEGRATION

INSTANCE INTEGRATION

- Matches tuples and attribute values
- Data integration from multiple heterogeneous data sources has become a high-priority task in many large enterprises. Hence to obtain the accurate semantic information on the data content, the information is being retrieved directly from the data. It identifies and integrates all the instance of the data items that represents the real-world entity, distinct from the schema integration

DATA INTEGRATION

INSTANCE INTEGRATION

PROBLEM

Employee Leave

EmployeeNo	EmployeeName	SSN	---	---
10014	Fred Aleck	E12233	---	---

Employee Attendance

EmployeeNo	EmployeeName	SSN	---	---
10014	Aleck Fred	E12233	---	---

Employee Payroll

EmployeeNo	EmployeeName	SSN	---	---
10014	F Aleck	E12233	---	---

DATA INTEGRATION

INSTANCE INTEGRATION

SOLUTION

Employee Leave

EmployeeNo	EmployeeName	SSN	---	---
10014	Fred Aleck	E12233	---	---

Employee Attendance

EmployeeNo	EmployeeName	SSN	---	---
10014	Fred Aleck	E12233	---	---

Employee Payroll

EmployeeNo	EmployeeName	SSN	---	---
10014	Fred Aleck	E12233	---	---

DATA INTEGRATION

NEED AND ADVANTAGES

- It is of benefit to decision-makers, who have access to important information from past studies in order to gain meaningful insights
- It helps to reduce costs, overlaps and redundancies; reduces exposure to risks
- It helps to monitor key variables like trending patterns and consumer behavior across geographies

DATA INTEGRATION

COMMON APPROACHES

- **Federated databases**

DATA INTEGRATION

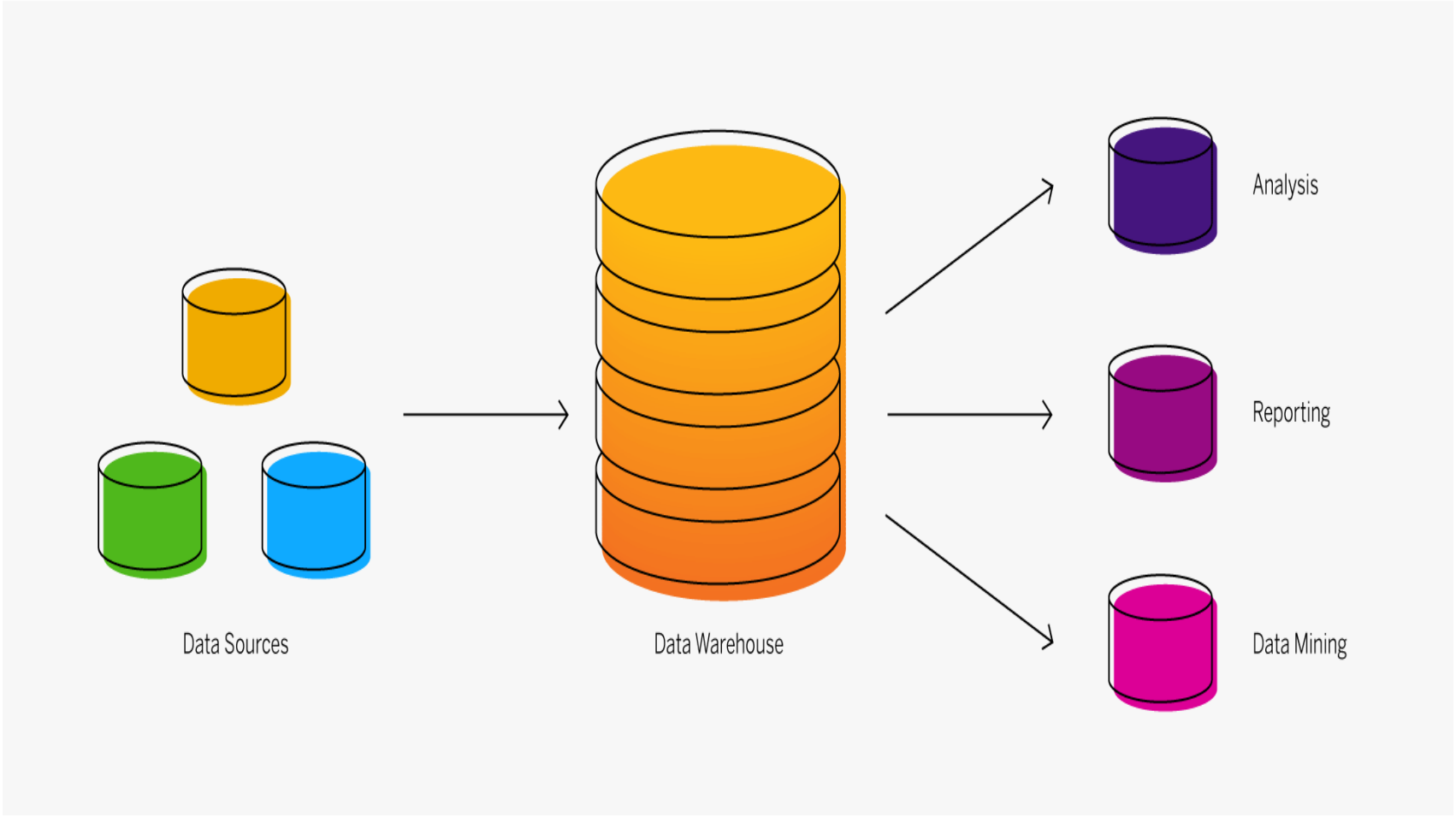
COMMON APPROACHES

- **Federated databases: Type of meta-database management system which transparently integrates multiple autonomous databases into a single federated database**
- **A collection of databases that are treated as one entity and viewed through a single user interface**
- **The federated database consists of a collection of heterogeneous databases**

DATA INTEGRATION

COMMON APPROACHES

- **Data warehouse:**



DATA INTEGRATION

COMMON APPROACHES

- Data warehouse: The various primary concepts used in data warehousing would be:
- ETL (Extract Transform Load)
- Component-based (Data Mart)
- Dimensional Models and Schemas
- Metadata driven

DATA QUALITY

DATA QUALITY

- Most of the data which is received by organizations is often duplicated, inconsistent, ambiguous and incomplete. This data needs to be collected and cleaned up, because bad data leads to bad decisions and bad decisions lead to bad business
- But it has been seen that most attention is paid to data integrity, which has a narrow scope than data quality
- We will look at both data integrity and data quality

DATA QUALITY

- **Data integrity reflects the degree to which the attributes of data associated with a certain entity accurately describe the occurrence of that entity**

  Examples of data integrity are,
- **Primary key: is Unique or not null and can be referred**
- **Foreign key: has values present in the primary key column that it refers to**
- **Not null: makes sure it must have a value**
- **Check constraint: adds a business rule**

DATA QUALITY

Data quality is described by several dimensions such as accuracy/correctness, completeness, consistency and timeliness

- Correctness/Accuracy: Accuracy of data is the degree to which the captured data correctly reflects/describes the real world entity/object or an event. Examples of data accuracy are,
- The temperature recorded in the thermometer is the real temperature
- The age of the patient as maintained in the hospital's database is the real age

DATA QUALITY

- **Consistency: This is about the single version of truth. Consistency means data throughout the enterprise should be in sync with each other.**

  Examples of data inconsistency are,

- **A customer has cancelled and surrendered his credit card. Yet the card billing status reads as "due"**
- **An employee left the organization. Yet his email address(with organizations domain) is still active**

DATA QUALITY

- **Completeness: The completeness of data is the extent to which the expected attributes of data are provided**

  Examples of data completeness are,
- **Data of all students of a university are available**
- **Data of all patients of a hospital are available**
- **All the data of all the clients of an IT organization is available**

DATA QUALITY

- Timeliness: The timeliness of data is extremely crucial. Right data to the right person at the right time is important for business. Delayed supply of data could prove harmful to the business. No matter how important the data is, if it is not supplied at the right time, it becomes inconsequential and useless. Examples are,
- The airlines are required to provide the most recent information to their passengers
- The population census results are published two years after the survey is completed

DATA QUALITY

How do we maintain data quality?

- From the technical standpoint, data quality results from the process of going through the data and scrubbing it, standardizing it, and de duplicating records, as well as doing some of the data enrichment
- Clean up your data by standardizing it using rules
- Use algorithms to detect duplicates which are obivious

DATA QUALITY

PROBLEM: DATA INCONSISTENCY

Invoice 1

BillNo	CustomerName	SSN	---	---
101	Mr. Aleck Steven	E12233	---	---

Invoice 2

BillNo	CustomerName	SSN	---	---
205	Mr. S. Aleck	E12233	---	---

Invoice 3

BillNo	CustomerName	SSN	---	---
314	Mr. Steven Aleck	E12233	---	---

DATA QUALITY

SOLUTION: CONSISTENT DATA

Invoice 1

BillNo	CustomerName	SSN	---	---
101	Mr. Aleck Steven	E12233	---	---

Invoice 2

BillNo	CustomerName	SSN	---	---
205	Mr. Aleck Steven	E12233	---	---

Invoice 3

BillNo	CustomerName	SSN	---	---
314	Mr. Aleck Steven	E12233	---	---

DATA PROFILING

DATA PROFILING

- Data profiling is a process of examining data from an existing source and summarizing information about that data
- You profile data to determine the accuracy, completeness, and validity of your data
- The value of your data depends on how well you profile it
- Today, only about 3% of data meets quality standards. That means poorly managed data is costing companies millions of dollars in wasted time, money, and untapped potential

DATA PROFILING

- Data profiling is the process of reviewing source data, understanding structure, content and interrelationships
- Data profiling can uncover data quality issues in data sources, and what needs to be corrected in ETL
- Data profiling can highlight data which suffers from serious or numerous quality issues, and the source of the issues (e.g. user inputs, errors in interfaces, data corruption)

DATA PROFILING

- Data profiling is the process of statistically examining and analyzing the content in a data source, and hence collecting information about that data. It consists of techniques used to analyze the data we have for accuracy and completeness
- Data profiling helps us make a thorough assessment of data quality
- It assists the discovery of anomalies in data
- Data profiling is also used to assess and validate metadata

DATA PROFILING – WHY?

- Is the data complete? Are there blank or null values?
- Is the data unique? How many distinct values are there? Is the data duplicated?
- What range of values exist, and are they expected? What are the maximum, minimum, and average values for given data? Are these the ranges you expect?
- Answering these questions helps you ensure that you are maintaining quality data, which companies are increasingly realizing

DATA PROFILING – HOW?

- **Data quality:** Analyze the quality of data at the data source
- **NULL values:** Look out for number of NULL values in an attribute
- **Primary key selection:** It is very essential to choose perfect primary key column
- **Empty string values:** The empty string values may create problems later, so it is important to look at such values
- **String length:** Proper value need to be assigned as string length of a particular column

DATA PROFILING – HOW?

- **Data format:** Sometimes, the format in which certain data is written in some columns may not be user friendly
- For example, if there is a string column that stores the marital status of a person in the form of "M" for "Married" and "U" for "Unmarried"

END