JGI JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

Department of CS & IT
Module 1

Data Warehousing and Data Mining

23MCAG203

by
Dr. S.K. Manju bargavi
Professor
Jain University

Text Books	
TB-1	Data Warehousing Fundamentals, Paulraj Ponnaiah, Wiley Student Edition. 2001 (Module 1 to 3)
TB-2	Data Mining Introductory and advanced topics, Margaret H Dunham, Pearson Education, 1st Edition, 2002 Module (4 and 5)
Reference Books	
RB-1	Data Warehousing in the Real World, Sam Anahory and Dennis Murray. Pearson Education Asia, 1st Edition, 1997
RB-2	The Data Warehouse Life Cycle Tool kit, Ralph Kimball, Wiley, 2nd Edition, 2008.
RB-3	The Data Mining Techniques, Arun K Pujari, University Press, 3rd Edition, 2003
RB 4	Data Mining – Concepts and Techniques, Jiawei Han and Micheline Kamber, Harcourt India, 2nd Edition, 2006
RB 5	Introduction to Data Mining, Pang-Ning Tan, Micheal Steinbach and Vipin Kumar, Pearson Education, 1st Edition, 2006.

Data, Information, Knowledge

- Data
  - Items that are the most elementary descriptions of things, events, activities, and transactions
  - May be internal or external
- Information
  - Organized data that has meaning and value
- Knowledge
  - Processed data or information that conveys understanding or learning applicable to a problem or activity

Data

- Raw data collected manually or by instruments
- Quality is critical
  - Quality determines usefulness
    - Contextual data quality
    - Intrinsic data quality
    - Accessibility data quality
    - Representation data quality
  - Often neglected or casually handled
  - Problems exposed when data is summarized

Data

- Cleanse data
  - When populating warehouse
  - Data quality action plan
  - Best practices for data quality
  - Measure results
- Data integrity issues
  - Uniformity :- Sample conforms to a uniform distribution
  - Version :- Versioning of Data
  - Completeness check :- The degree to which all data in a data set is available
  - Conformity check :- the data values of the same attributes must be represented in a uniform format and data types.

List the major categories of data sources for an MSS/BI

Internal sources; usually the reporting systems of the functional areas.

External sources:(commercial databases, government and industry reports, etc.) and personal data.

Benefits of commercial databases

Provide external data at a timely manner and at a reasonable cost. Because of economies of scale, such services are comprehensive and inexpensive

Database Vs. Datawarehouse

- Databases are typically the term used to describe operational data stores and are transactional in their structure. As a result databases are usually highly normalized, whereas data warehouses are highly de-normalized.

- Technically a data warehouse is a database, however, a data warehouse is an integrated, time-variant, nonvolatile, subject-oriented repository of detail and summary data used for decision support and business analytics within an organization.

Character	Data Warehouse	Transactional Database
Workloads	Analytics, reporting, big data	Transaction processing
Data source	Data collected and normalized from many sources	Data captured as-is from a single source, such as a transactional system
Data capture	Bulk write operations typically on a predetermined batch schedule	Optimized for continuous write operations as new data is available to maximize transaction throughput
Data normalization	De-normalized schemas, such as the Star schema or Snowflake schema	Highly normalized, static schemas
Data storage	Optimized for simplicity of access & high-speed query performance using columnar storage	Optimized for high throughout write operations to a single row-oriented physical block
Data access	Optimized to minimize I/O and maximize data throughput	High volumes of small read operations

Data warehouse

A data warehouse is a physically separate database from a company's operational environments. Its purpose is to provide decision support from its *data repository* that makes operational data accessible in a form that is readily acceptable for decision support and other user's applications.

Data warehousing is the process of taking internal data, cleansing it, and storing it in a data warehouse where it can be accessed by various decision makers in the decision-making process. External information is also brought into the data warehouse.

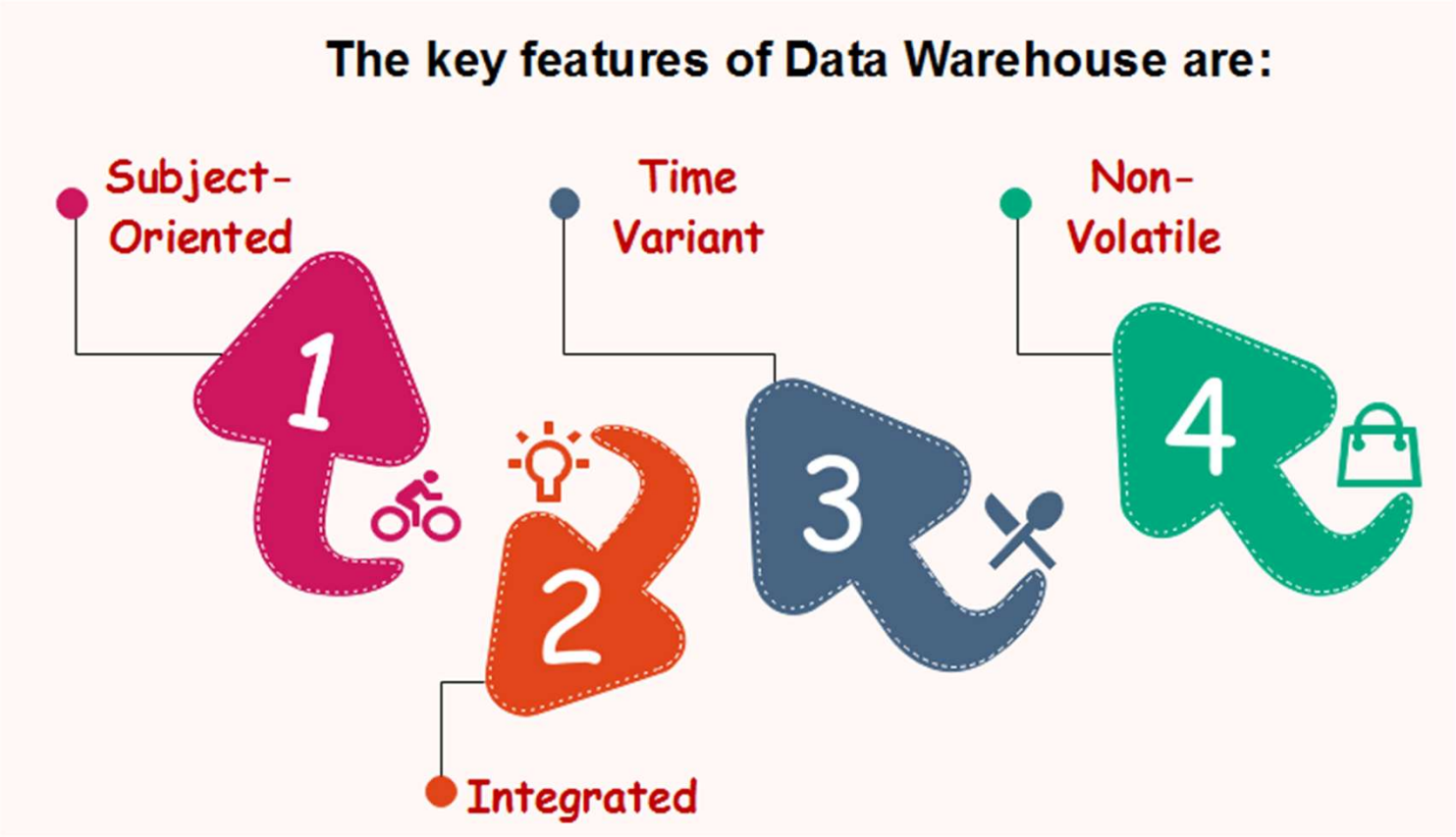
Data Warehouse

- A data warehouse can be defined as a collection of organizational data and information extracted from operational sources and external data sources.
- The data is periodically pulled from various internal applications like sales, marketing, and finance; customer-interface applications; as well as external partner systems. This data is then made available for decision-makers to access and analyze.
- So what is data warehouse? For a start, it is a comprehensive repository of current and historical information that is designed to enhance an organization's performance.

Data warehouse

- A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of managements decision making process.
- It is the process whereby organizations extract value from their informational assets through use of special stores called data warehouses

Characteristics of Data Warehouse

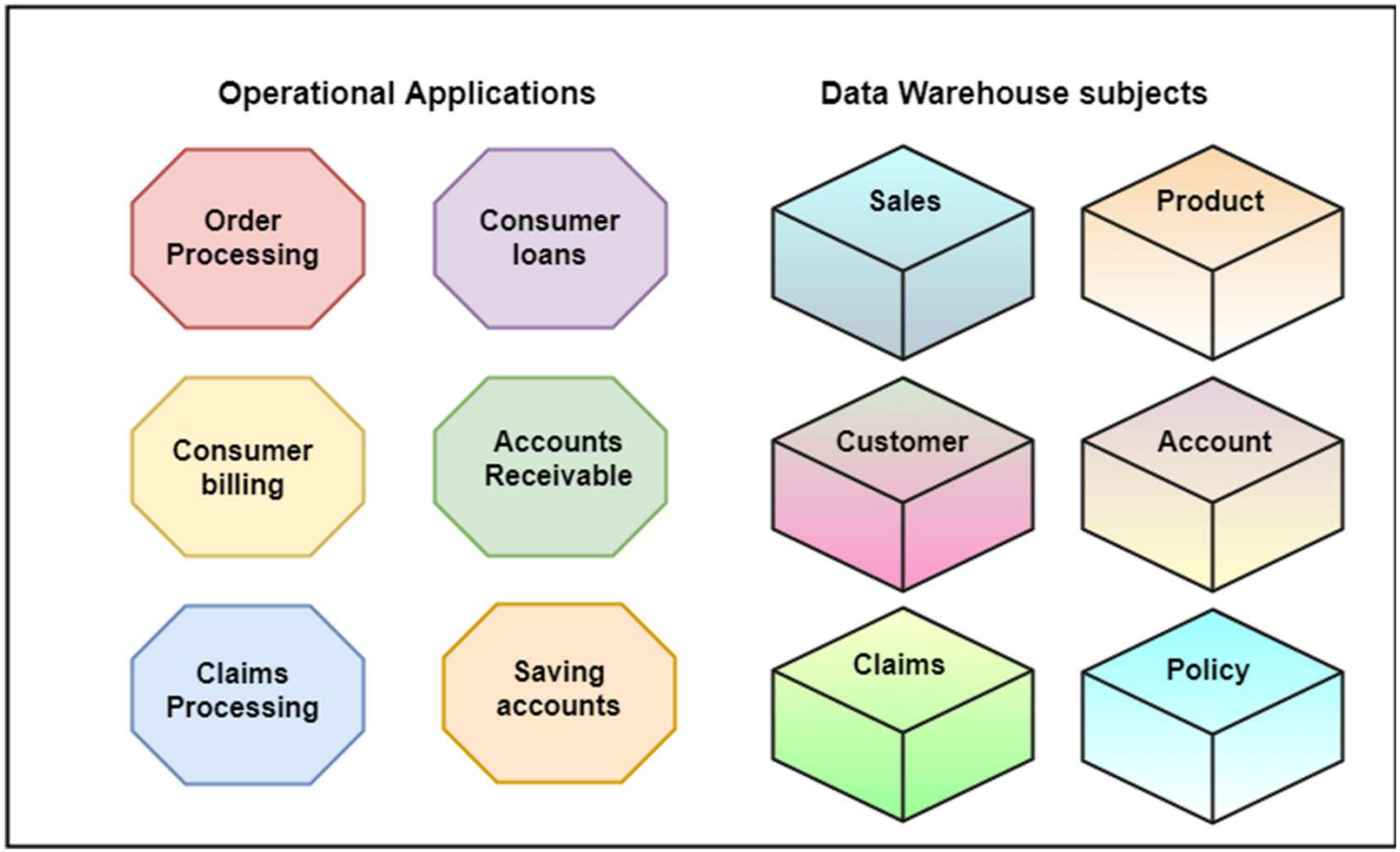


Subject-Oriented

- A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations.

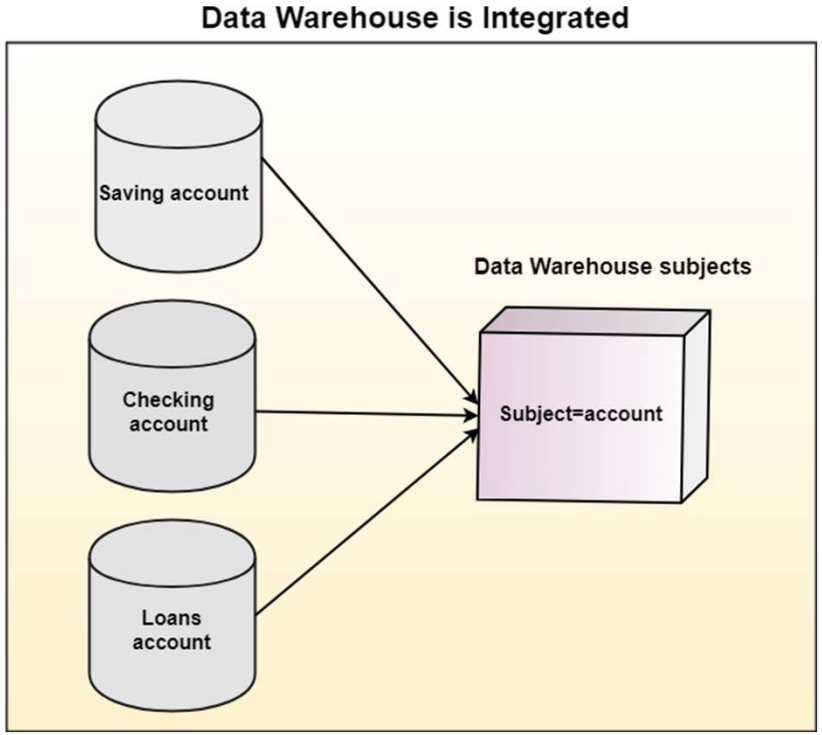
- This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.

Data Warehouse is Subject-Oriented



Integrated

- A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.

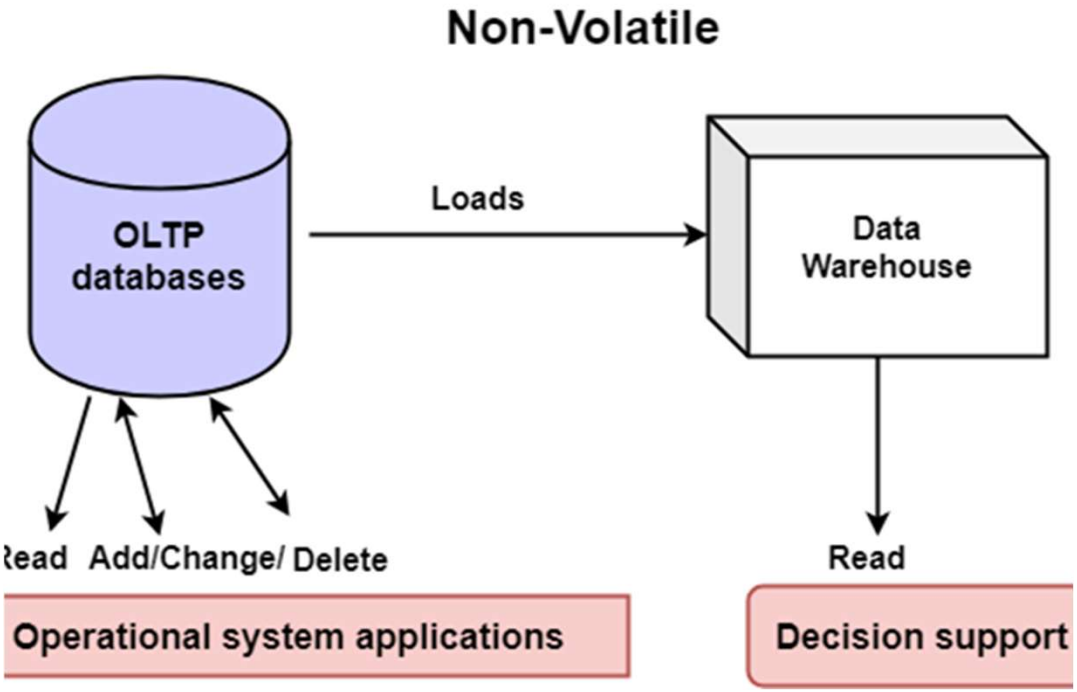


Time-Variant

- Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.

Non-Volatile

- The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS.
- It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval.
- Non-Volatile defines that once entered the warehouse, and data should not change.



Key Characteristics of Data Warehouse

- Data Granularity
  - In a data warehouse, **data granularity** is the level of detail in a model or decision making process. It tells you how detailed your data is: Lower levels of detail equal finer, more detailed, data granularity. Finer, more granulated data will allow you to perform more precise data analysis.
  - For example, time-series data for sales volume can be measured in years, months, weeks, or days — with days as the lowest level of granularity. Performing data analysis on the more granular daily data will result in better insights on sales than yearly data. .

Benefits of Data Warehouse

- Understand business trends and make better forecasting decisions.
- Data Warehouses are designed to perform well enormous amounts of data.
- The structure of data warehouses is more accessible for end-users to navigate, understand, and query.
- Queries that would be complex in many normalized databases could be easier to build and maintain in data warehouses.
- Data warehousing is an efficient method to manage demand for lots of information from lots of users.
- Data warehousing provide the capabilities to analyze a large amount of historical data.

Types

- Enterprise Data Warehouse (EDW)
  - This type of warehouse serves as a key or central database that facilitates decision-support services throughout the enterprise. The advantage to this type of warehouse is that it provides access to cross-organizational information, offers a unified approach to data representation, and allows running complex queries.

Types

- Operational Data Store (ODS)
  - This type of data warehouse refreshes in real-time. It is often preferred for routine activities like storing employee records. It is required when data warehouse systems do not support reporting needs of the business.

Types

- Data Mart
  - A data mart is a subset of a data warehouse built to maintain a particular department, region, or business unit. Every department of a business has a central repository or data mart to store data. The data from the data mart is stored in the ODS periodically. The ODS then sends the data to the EDW, where it is stored and used.

What is OLTP?

- OLTP or online transactional processing is a software program or operating system that supports transaction-oriented applications in a three-tier architecture. It facilitates and supports the execution of a large number of real-time transactions in a database.

What is OLTP?

- The defining characteristic of OLTP transactions is atomicity and concurrency.
- Concurrency prevents multiple users from changing the same data simultaneously. Atomicity (or indivisibility) ensures that all transactional steps are completed for the transaction to be successful. If one step fails or is incomplete, the entire transaction fails.
- Atomic statefulness is a computing condition in which database changes are permanent, requiring transactions to be completed successfully. OLTP systems enable inserting, deleting, changing, and querying data in a database.

What is OLTP?

- OLTP systems activities consist of gathering input data, processing the data, and updating it using the data collected. OLTP is usually supported by a database management system (DBMS) and operates in a client-server system. It also relies on advanced transaction management systems to facilitate multiple concurrent updates.

OLTP Transaction Examples

- OLTP systems facilitate many types of financial and non-financial transactions such as:
  - Automated teller machines (ATMs)
  - Online banking applications
  - Online bookings for airline ticketing, hotel reservations, etc.
  - Online and in-store credit card payment processing
  - Order entry
  - E-commerce and in-store purchases
  - Password changes and sending text messages
  - OLTP systems are found in a broad spectrum of industries with a concentration in client-facing environments.

OLTP Characteristics

1. Short response time

- OLTP systems maintain very short response times to be effective for users. For example, responses from an ATM operation need to be quick to make the process effective, worthwhile, and convenient.

2. Process small transactions

- OLTP systems support numerous small transactions with a small amount of data executed simultaneously over the network. It can be a mixture of queries and Data Manipulation Language (DML) overload. The queries normally include insertions, deletions, updates, and related actions. Response time measures the effectiveness of OLTP transactions, and millisecond responses are becoming common.

OLTP Characteristics

3. Data maintenance operations

- Data maintenance operations are data-intensive computational reporting and data update programs that run alongside OLTP systems without interfering with user queries.

4. High-level transaction volume and multi-user access

- OLTP systems are synonymous with a large number of users accessing the same data at the same time. Online purchases of a popular or trending gadget such as an iPhone may involve an enormous number of users all vying for the same product. The system is built to handle such situations expertly.

OLTP Characteristics

5. Very high concurrency

- An OLTP environment experiences very high concurrency due to the large user population, small transactions, and very short response times. However, data integrity is maintained by a concurrency algorithm, which prevents two or more users from altering the same data at the same time. It prevents double bookings or allocations in online ticketing and sales, respectively.
- A mobile money transfer application is a good example where concurrency is very high as thousands of users can be making transfers simultaneously on the platform at every time of the day.

OLTP Characteristics

6. Round-the-clock availability

- OLTP systems often need to be available round the clock, 24/7, without interruption. A small period of unavailability or offline operations can significantly impact a large number of people and an equally huge transaction quantity.
- Downtimes can also pose potential losses to organizations, e.g., an online banking system downtime has adverse consequences to the bank's bottom line. Therefore, an OLTP system requires frequent, regular, and incremental backup.

OLTP Characteristics

7. Data usage patterns

- OLTP systems experience periods of both high data usage and low data usage. Finance-related OLTP systems typically see high data usage during month ends when financial obligations are settled.

8. Indexed data sets

- Index data sets are used to facilitate rapid query, search, and retrieval.

OLTP Characteristics

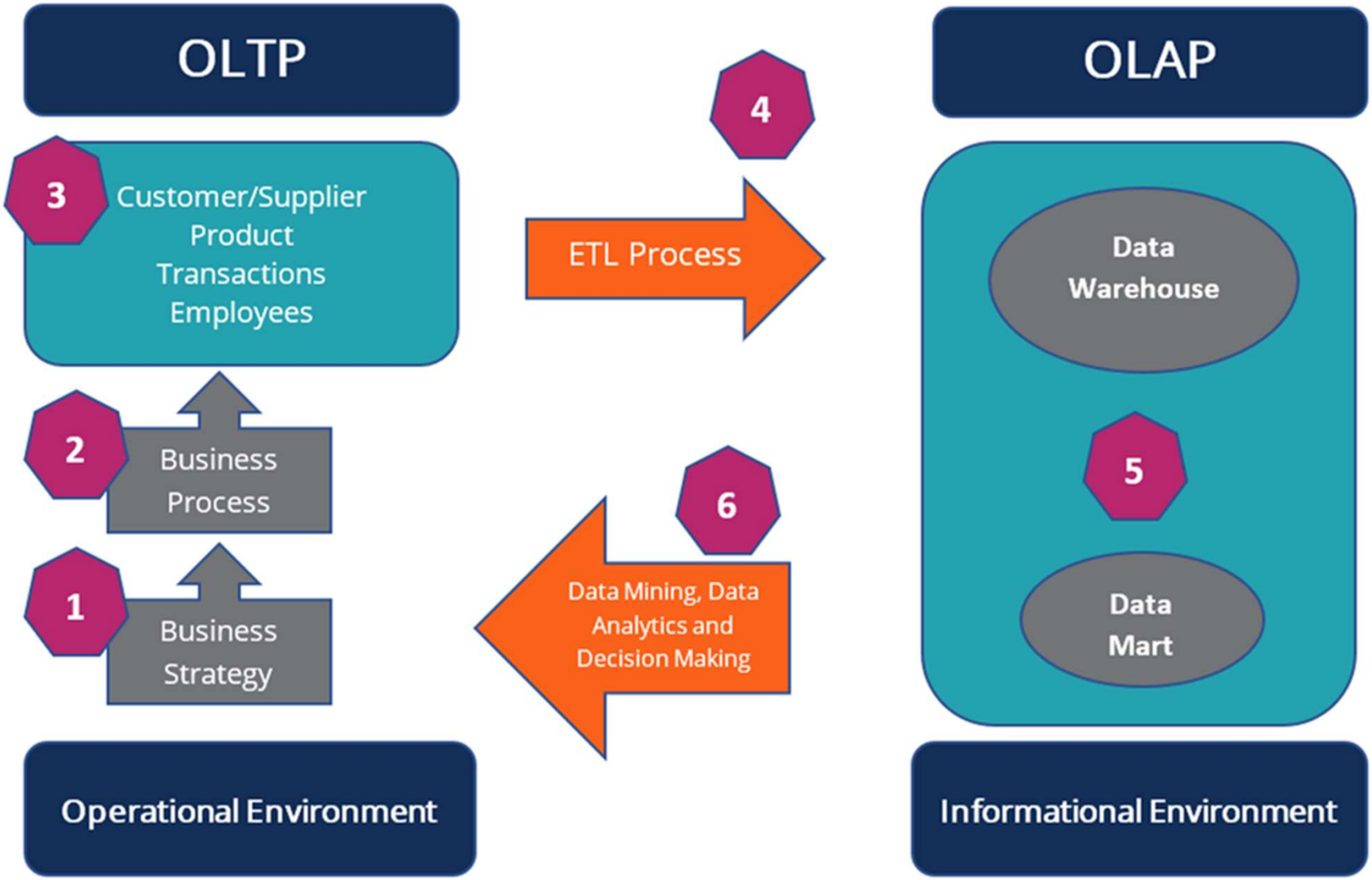
9. Normalized schema

- OLTP systems utilize a fully normalized schema for database consistency.

10. Storage

- OLTP stores data records for the past few days or about a week. It supports sophisticated data models and tables.

OLTP Architecture & System Design



OLTP Architecture & System Design

1. Business Strategy

- The business strategy influences the OLTP systems design. The strategy is formulated at the senior management and the level of the board of directors.

2. Business Process

- They are processes by the OLTP system that will accomplish the goals set by the business strategy. The processes comprise a set of activities, tasks, and actions.

OLTP Architecture & System Design

3. Product, Customer/Supplier, Transactions, Employees

- The OLTP database contains information on products, transactions, employees, and customers, and suppliers.

4. Extract, Transform, Load (ETL) Process

- The ETL process extracts data from the OLTP database and transforms it into the staging area, which includes data cleansing and optimizing the data for analysis. The transformed data is then loaded into the online analytical processing (OLAP) database, which is synonymous with the data warehouse environment.

OLTP Architecture & System Design

5. Data Warehouse and Data Mart

- Data warehouses are central repositories of integrated data from one or more incongruent sources. A data mart is an access layer of the data warehouse that is used to access specific/summarized information of a unit or department.

6. Data Mining, Analytics, and Decision Making

- The data stored in the data warehouse and data mart is used for analysis, data mining, and decision making.

Advantages of Data Warehouse

- Now that we are aware of data warehouse meaning and how they work, it is time to know the benefits of data warehouses and how exactly they can help your business grow and scale. Whether you own a digital marketing agency or have a traditional brick-and-mortar setup, data warehousing can yield several benefits for your business.

Advantages of Data Warehouse

Photo taken from intellipaat.com/blog/tutorial/data-warehouse-tutorial/data-warehouse-quality-management/

Advantages of Data Warehouse

1. Saves Time
   - In the modern fast-paced world of cut-throat competition, your capacity as a business to swiftly make refined decisions is essential to outpace your opponents.
   - A DWH provides you access to all your required data in a matter of minutes, so you and your employees don't have to dread an approaching deadline. All you need to do is deploy your data model to acquire data within seconds. Most warehousing solutions allow you to do that without using a complex query or machine learning.
   - With data warehousing, your business won't have to rely on the 24/7 availability of a technical expert to troubleshoot problems associated with retrieving information. This way, you can save plenty of time.

Advantages of Data Warehouse

2. Improves Data Quality

- The refined quality of data helps guarantee that your company's policies are based on precise information about your corporate exertions.
- By understanding the data warehousing meaning, you can transform data from multiple sources into a shared arrangement. Consequently, you can ensure the reliability and quality of your corporate data. This way, you can identify and remove replicated data, poorly recorded data, and any other errors.
- Implementing a data quality management program and improving data integrity can be both costly and laborious for your company. You can easily use a data warehouse to eliminate a number of these annoyances while saving money and boosting your organization's overall efficiency.

Advantages of Data Warehouse

3. Improves Business Intelligence

- You can use a data warehouse to gather, assimilate, and derive data from any source and set up a process to leverage business analytics. As a result, your BI will improve by leaps and bounds, owing to the capability of effortlessly integrating data from distinct sources.
- Let's face it: cross-checking numerous databanks can be tough, and at times, inconvenient. But, with a data warehouse in place, everyone on your team can have an integrated understanding of all the relevant information in a timely manner.

Advantages of Data Warehouse

4. Leads to Data Consistency

- Another important benefit of using central data stores is the evenness of big data. Your business can benefit from a data storage or data mart in a similar arrangement. As data warehousing stores large amounts of data from diverse sources, such as a transactional system, in a consistent fashion, each source will generate outcomes that are synchronized with other sources.
- This guarantees improved quality and consistency of data. Consequently, you and your team can feel assured that your data is correct, which will result in more cognizant corporate decisions.

Advantages of Data Warehouse

5. Stores Historical Data

- As a data warehouse allows you to store large volumes of historical data from databases, you can easily investigate different time phases and inclinations that can be ground-breaking for your company. Thus, with the right and real-time data in your hands, you can make superior corporate decisions concerning your business strategies.
- Moreover, predicting the results of your business processes is a significant aspect of being a resourceful business person. It can be challenging to forecast the future without a tangible understanding of your historical achievements and letdowns.

Advantages of Data Warehouse

6. Increases Data Security

- But, with data warehousing, you can save yourself from the hassle of additional data security. (5 Million US Dollars)
- As a business that deals with customer information regularly, your first and foremost priority is to protect your existing and prospective consumers' information. Hence, to evade all future nuisances, you take all the necessary actions to escape data breaches. Using a warehousing solution, you can keep all your data sources consolidated and protected. This will significantly decrease the threat of a data breach

Advantages of Data Warehouse

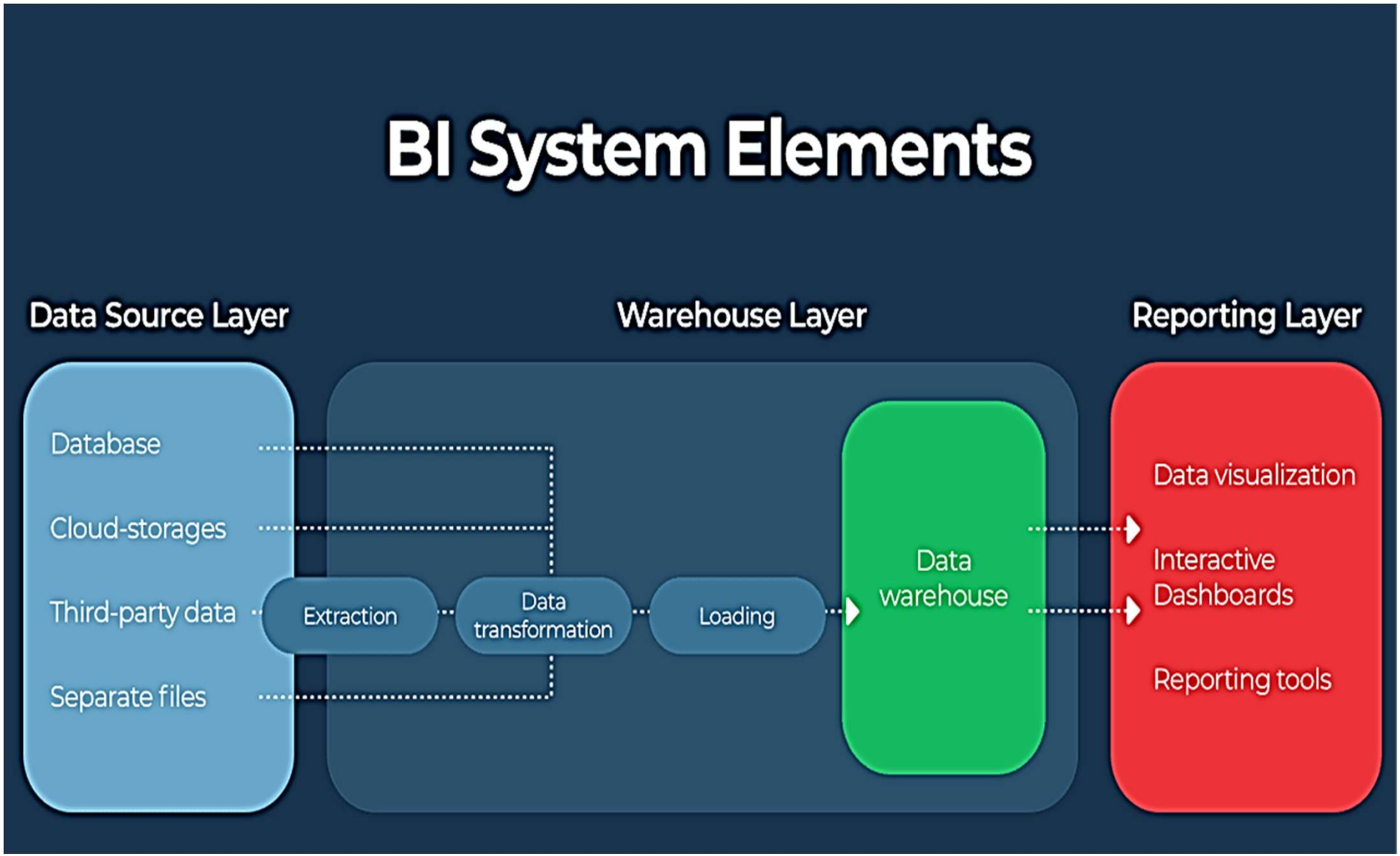


Photo taken from www.cleveroad.com/blog/bi-developer-roles-and-responsibilities

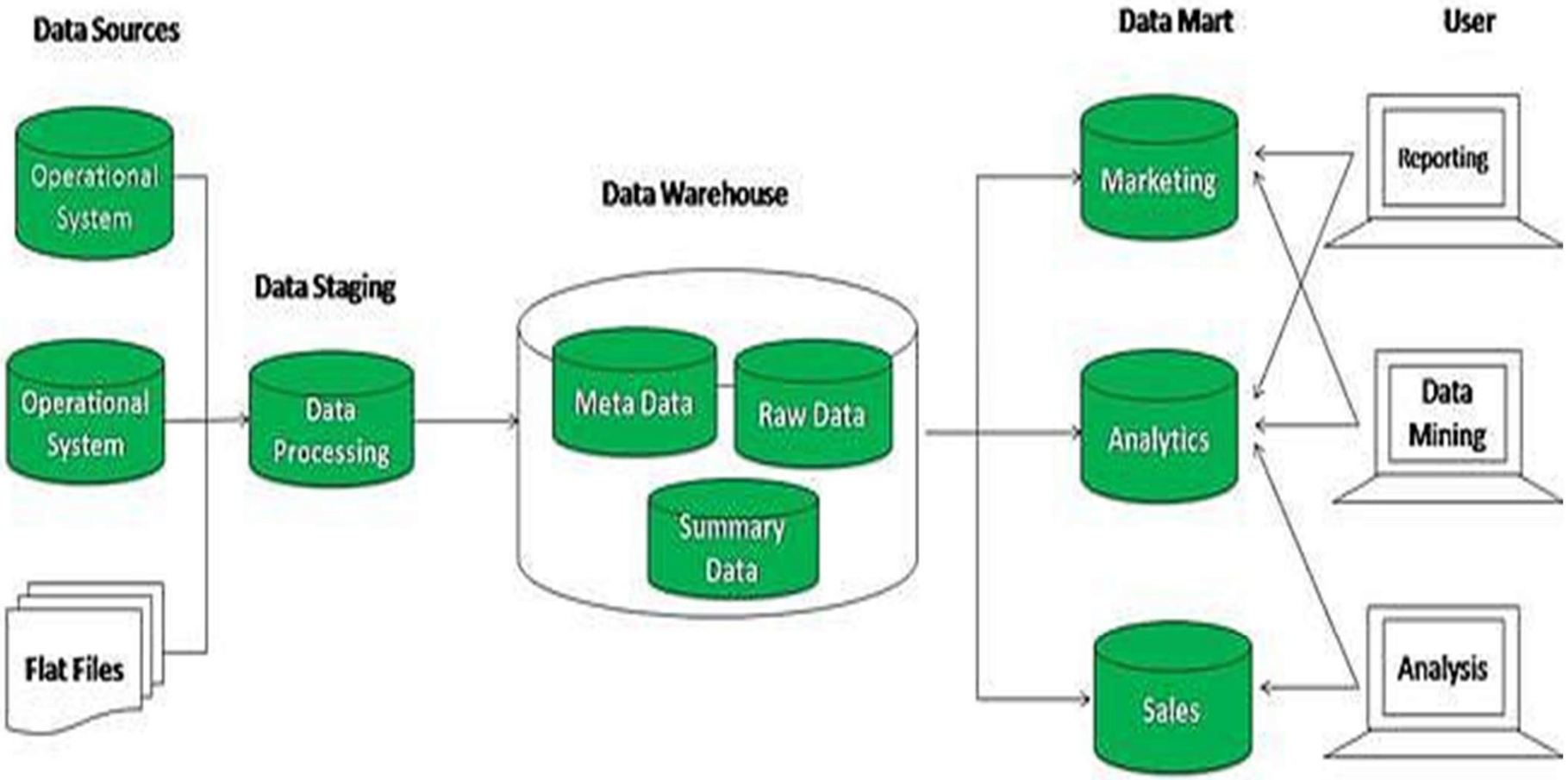
Advantages of Data Warehouse

- For example, suppose you own a fashion brand. You plan to launch a promotional campaign for your new clothing line. Setting up a central repository enables you to access and analyze historical data from your previous campaigns in order to identify which approach worked the best, and how you might emulate it in upcoming promotions.
- You can’t expect to store and analyze such comprehensive past data in any conventional databank. Thus, using EDW gives you an advantage in your business procedures.

Advantages of Data Warehouse

- A data warehouse allows improved security by offering cutting-edge safety characteristics erected into its setup. Consumer information is a valuable resource for any company. But once safety becomes a problem, this information becomes your main burden.
- These are just a few advantages that data warehousing has to offer for your business. It provides you with improved business intelligence, robust decision support, superior business practices, and effective analytics processing.

Tools for Data Warehouse



Data Warehouse Applications

- As discussed before, a data warehouse helps business executives to organize, analyze, and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess "closed-loop" feedback system for the enterprise management. Data warehouses are widely used in the following fields –
- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

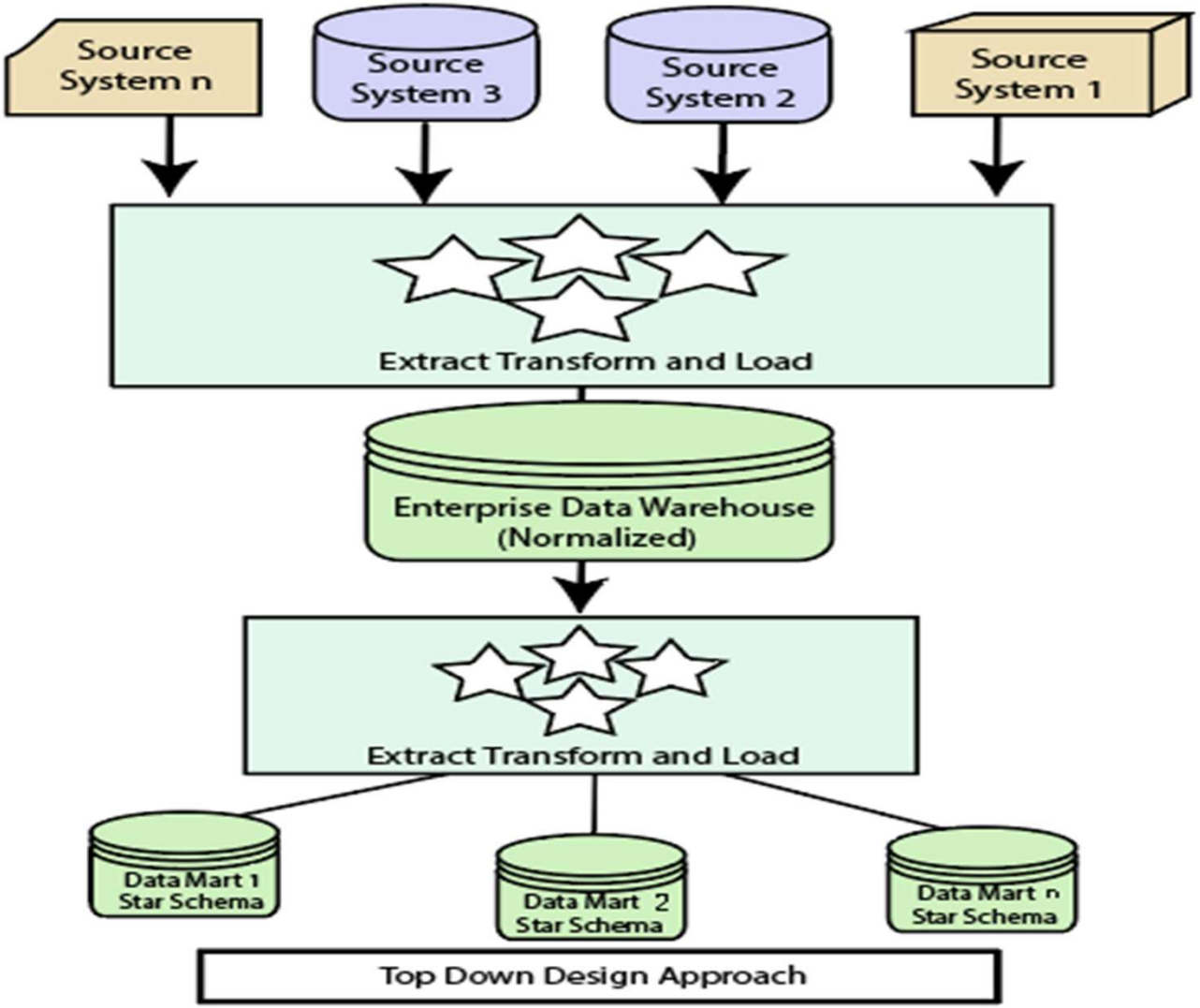
Data Warehousing

- A data warehouse is a single data repository where a record from multiple data sources is integrated for online business analytical processing (OLAP). This implies a data warehouse needs to meet the requirements from all the business stages within the entire organization.
- Thus, data warehouse design is a hugely complex, lengthy, and hence error-prone process. Furthermore, business analytical functions change over time, which results in changes in the requirements for the systems. Therefore, data warehouse and OLAP systems are dynamic, and the design process is continuous.

Data Warehousing

- Data warehouse design takes a method different from view materialization in the industries. It sees data warehouses as database systems with particular needs such as answering management related queries.
- The target of the design becomes how the record from multiple data sources should be extracted, transformed, and loaded (ETL) to be organized in a database as the data warehouse.
- There are two approaches
  - "top-down" approach
  - "bottom-up" approach

Top-down Design Approach



Top Down Design Approach

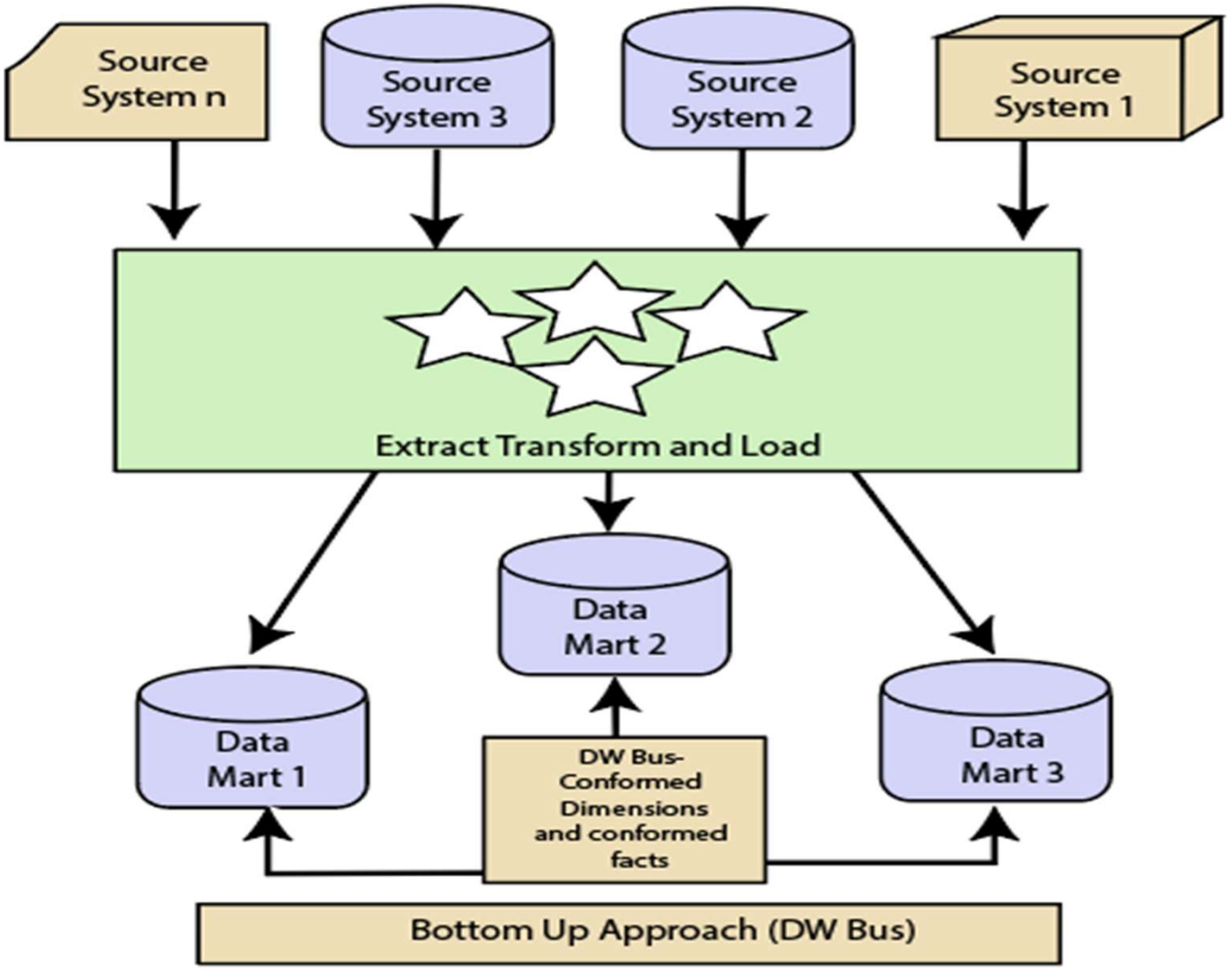
Top-down Design Approach

- In the "Top-Down" design approach, a data warehouse is described as a subject-oriented, time-variant, non-volatile and integrated data repository for the entire enterprise data from different sources are validated, reformatted and saved in a normalized (up to 3NF) database as the data warehouse.
- The data warehouse stores "atomic" information, the data at the lowest level of granularity, from where dimensional data marts can be built by selecting the data required for specific business subjects or particular departments. This approach is a data-driven approach as the information is gathered and integrated first and then business requirements by subjects for building data marts are formulated.
- The advantage of this method is which it supports a single integrated data source. Thus data marts built from it will have consistency when they overlap.

Top-down Design Approach

- **Advantages of top-down design**
  - Data Marts are loaded from the data warehouses.
  - Developing new data mart from the data warehouse is very easy.
- **Disadvantages of top-down design**
  - This technique is inflexible to changing departmental needs.
  - The cost of implementing the project is high.

Bottom-Up Design Approach



Bottom Up Design Approach

Bottom-Up Design Approach

- In the "Bottom-Up" approach, a data warehouse is described as "a copy of transaction data specifically architecture for query and analysis," term the star schema. In this approach, a data mart is created first to necessary reporting and analytical capabilities for particular business processes (or subjects). Thus it is needed to be a business-driven approach in contrast to data-driven approach.
- Data marts include the lowest grain data and, if needed, aggregated data too. Instead of a normalized database for the data warehouse, a denormalized dimensional database is adapted to meet the data delivery requirements of data warehouses.

Bottom-Up Design Approach

- Using this method, to use the set of data marts as the enterprise data warehouse, data marts should be built with conformed dimensions in mind, defining that ordinary objects are represented the same in different data marts. The conformed dimensions connected the data marts to form a data warehouse, which is generally called a virtual data warehouse.

- The advantage of the "bottom-up" design approach is that it has quick ROI, as developing a data mart, a data warehouse for a single subject, takes far less time and effort than developing an enterprise-wide data warehouse. Also, the risk of failure is even less. This method is inherently incremental. This method allows the project team to learn and grow.

Bottom-Up Design Approach

- **Advantages of bottom-up design**
  - Documents can be generated quickly.
  - The data warehouse can be extended to accommodate new business units.
  - It is just developing new data marts and then integrating with other data marts.

- **Disadvantages of bottom-up design**
  - the locations of the data warehouse and the data marts are reversed in the bottom-up approach design.

Differentiate between Top-Down Design Approach and Bottom-Up Design Approach

Top-Down Design Approach	Bottom-Up Design Approach
Breaks the vast problem into smaller subproblems.	Solves the essential low-level problem and integrates them into a higher one.
Inherently architected- not a union of several data marts.	Inherently incremental; can schedule essential data marts first.
Single, central storage of information about the content.	Departmental information stored.
Centralized rules and control.	Departmental rules and control.
It includes redundant information.	Redundancy can be removed.
It may see quick results if	Less risk of failure, favorable

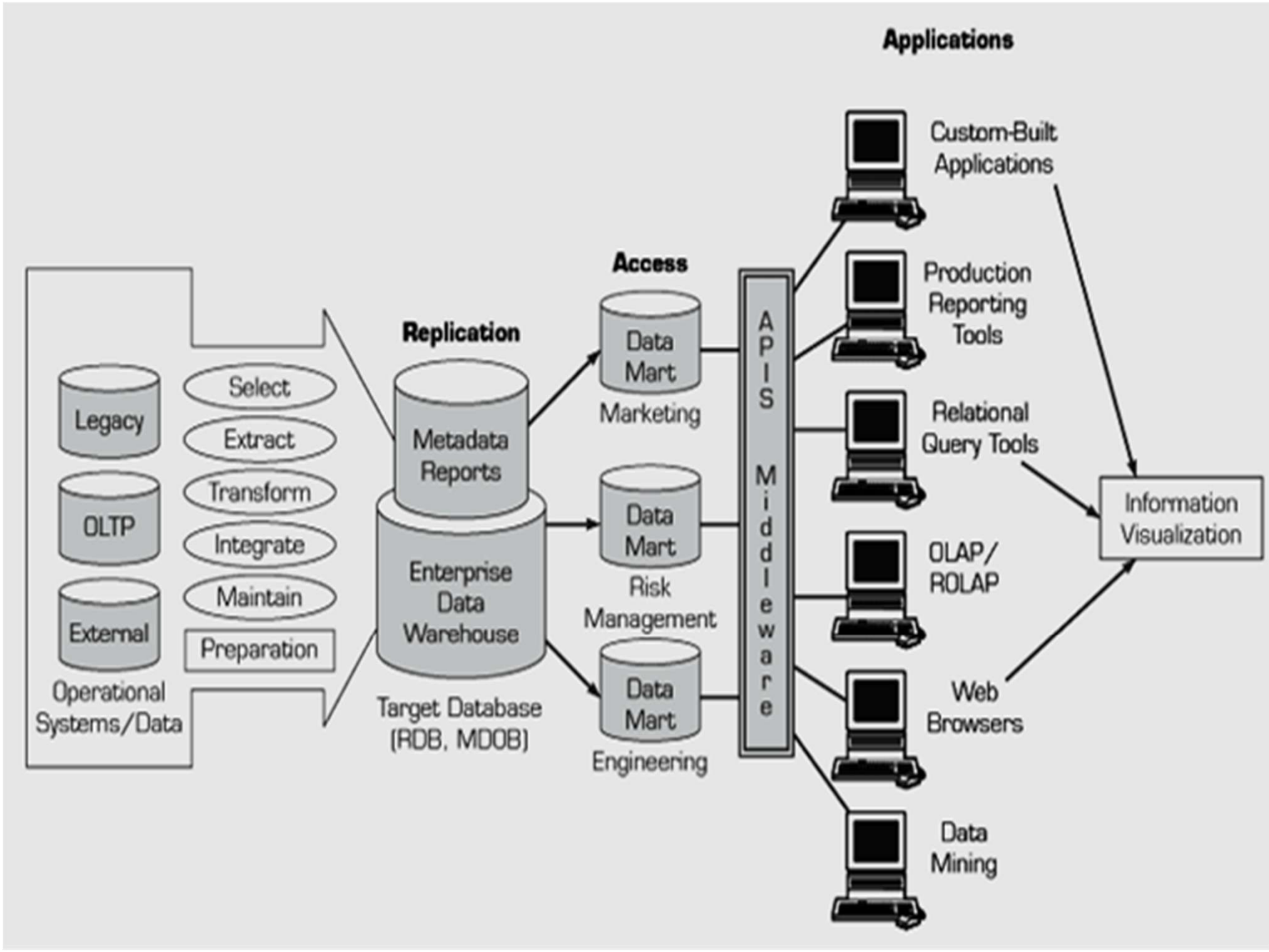


Figure1.1: Data Warehouse Framework and Views

Tools for Datawarehouse

1. Hevo Data
2. Amazon Web Services Data Warehouse Tools
3. Google Data Warehouse Tools
4. Microsoft Azure Data Warehouse Tools
5. Oracle Autonomous Data Warehouse
6. Snowflake
7. IBM Data Warehouse Tools
8. Teradata Vantage
9. SAS Cloud
10. SAP Data Warehouse Cloud

https://www.geeksforgeeks.org/top-15-popular-data-warehouse-tools/

Planning

- improper planning and inadequate project management tend to result in failures.

Factors causing failures

- Improper planning
- Inadequate project management
- Company not ready for a data warehouse
- Insufficient staff training
- Improper team management
- No support from top management

- **Decisions Decide the type of data warehouse to be built**
  - where to keep the data warehouse
  - where the data is going to come from
  - whether you have all the needed data
  - who will be using the data warehouse
  - how they will use it
  - at what times will they use it

Key issues

- Value and expectation
- Risk assessment
- Top-down or bottom-up
- Build or Buy
- Single vendor or best of breed

Key Issues

Value and Expectations	Risk Assessment	Top-Down or Bottom-Up	Build or Buy	Single Vendor or Best-of-Breed
Asses the value to be derived from the proposed data warehouse,	More than calculating the loss from the project costs	Plan and define overall requirements	Find the proper balance between in-house and vendor software.	High level of integration or products best suited for objectives
Make a list of realistic benefits and expectations	Take into account the opportunities that will be missed if there is NO data warehouse	Look at the pros and cons of these methods		
		Weight these options and document them		

Driving Force

- Business Requirements, Not Technology
- Understand the requirements
- Focus on
  - user's needs
  - Data needed
  - How to provide information
- Use a preliminary survey to gather general requirements before planning

Preliminary Survey

- Mission and functions of each user group
- Computer systems used by the group
- Key performance indicators
- Factors affecting success of the user group
- Who the customers are and how they are classified
- Types of data tracked for the customers, individually and as groups
- Products manufactured or sold
- Categorization of products and services
- Locations where business is conducted
- Levels at which profits are measured—per customer, per product, per district
- Levels of cost details and revenue
- Current queries and reports for strategic information

Key Issues in Planning a Data Warehouse

Here are some of the **difficulties** of **Implementing Data Warehouses**:

- Implementing a data warehouse is generally a massive effort that must be planned and executed according to established methods.
- Construction, administration, and quality control are the significant operational issues which arises with data warehousing.
- Some of the important and challenging consideration while implementing data warehouse are: the design, construction and implementation of the warehouse.
- The building of an enterprise-wide warehouse in a large organization is a major undertaking.
- Manual Data Processing can risk the correctness of the data being entered.

Key Issues in Planning a Data Warehouse

- An intensive enterprise is the administration of a data warehouse, which is proportional to the complexity and size of the warehouse.
- The complex nature of the administration should be understood by an organization that attempts to administer a data warehouse.
- There must be a flexibility to accept and integrate analytics to streamline the business intelligence process.
- To handle the evolutions, acquisition component and the warehouse's schema should be updated.
- A significant issue in data warehousing is the quality control of data. The major concerns are: quality and consistency of data.

Justification

1. Calculate the current technology costs to produce the applications and reports supporting strategic decision making. Compare this with the estimated costs for the data warehouse and find the ratio between the current costs and proposed costs. See if this ratio is acceptable to senior management.
2. Calculate the business value of the proposed data warehouse with the estimated dollar values for profits, dividends, earnings growth, revenue growth, and market share growth. Review this business value expressed in dollars against the data warehouse costs and come up with the justification.
3. Do the full-fledged exercise. Identify all the components that will be affected by the proposed data warehouse and those that will affect the data warehouse. Start with the cost items, one by one, including hardware purchase or lease, vendor software, in-house software, installation and conversion, ongoing support, and maintenance costs. Then put a dollar value on each of the tangible and intangible benefits, including cost reduction, revenue enhancement, and effectiveness in the business community.

Challenges for Data Warehousing Project Management

DATA ACQUISITION

- Large number of sources
- Many disparate sources
- Different computing
- platforms
- Outside sources
- Huge initial load
- Ongoing data feeds
- Data replication considerations
- Difficult data integration
- Complex data transformations
- Data cleansing

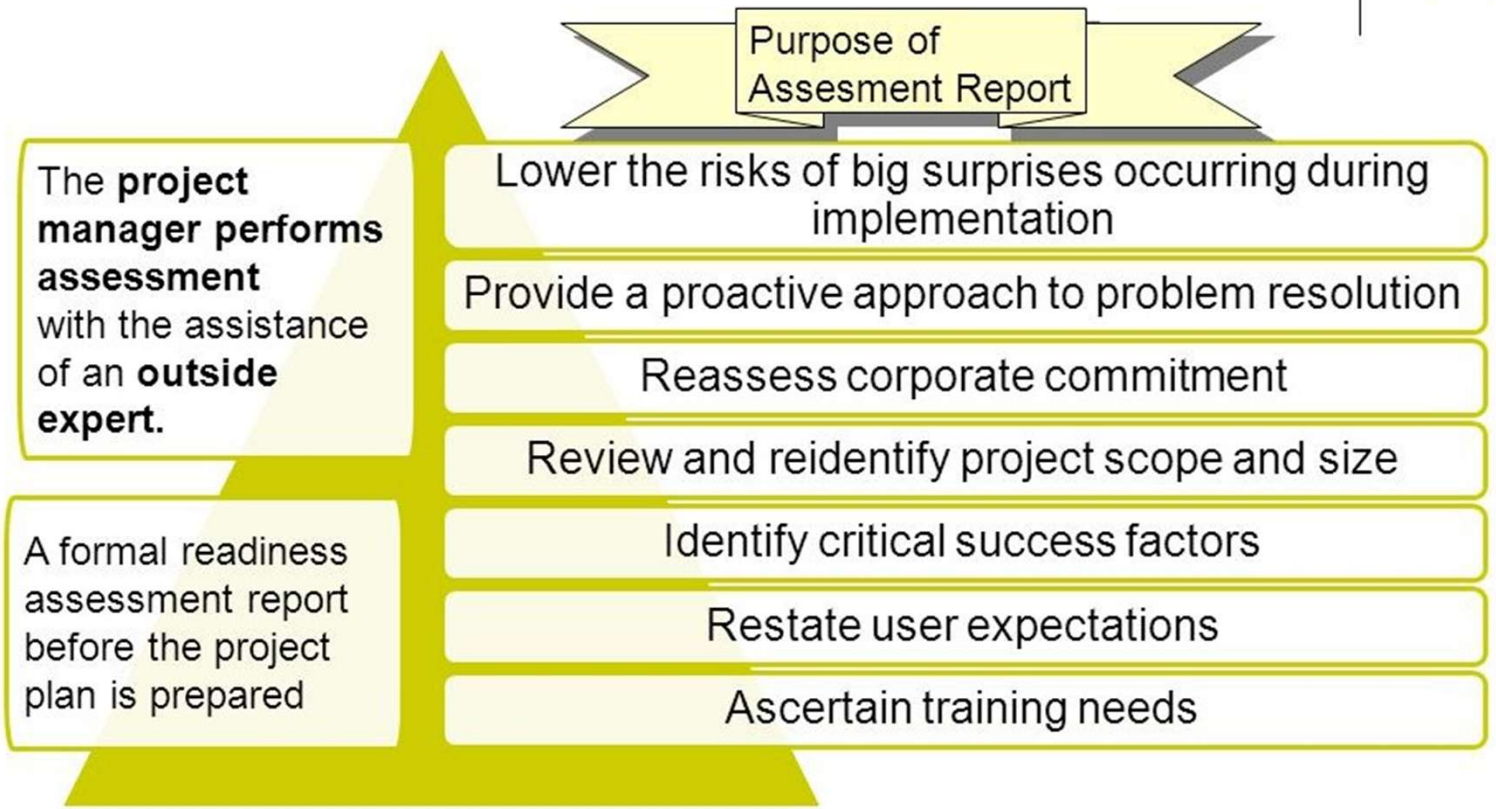
DATA STORAGE

- Storage of large data
- volumes
- Rapid growth
- Need for parallel
- processing
- Data storage in staging area
- Multiple index types
- Several index files
- Storage of newer data types
- Archival of old data
- Compatibility with tools
- RDBMS & MDDBMS

INFO. DELIVERY

- Several user types
- Queries stretched to limits
- Multiple query types
- Web-enabled
- Multidimensional analysis
- OLAP functionality
- Metadata management
- Interfaces to DSS apps.
- Feed into Data Mining
- Multi-vendor tools

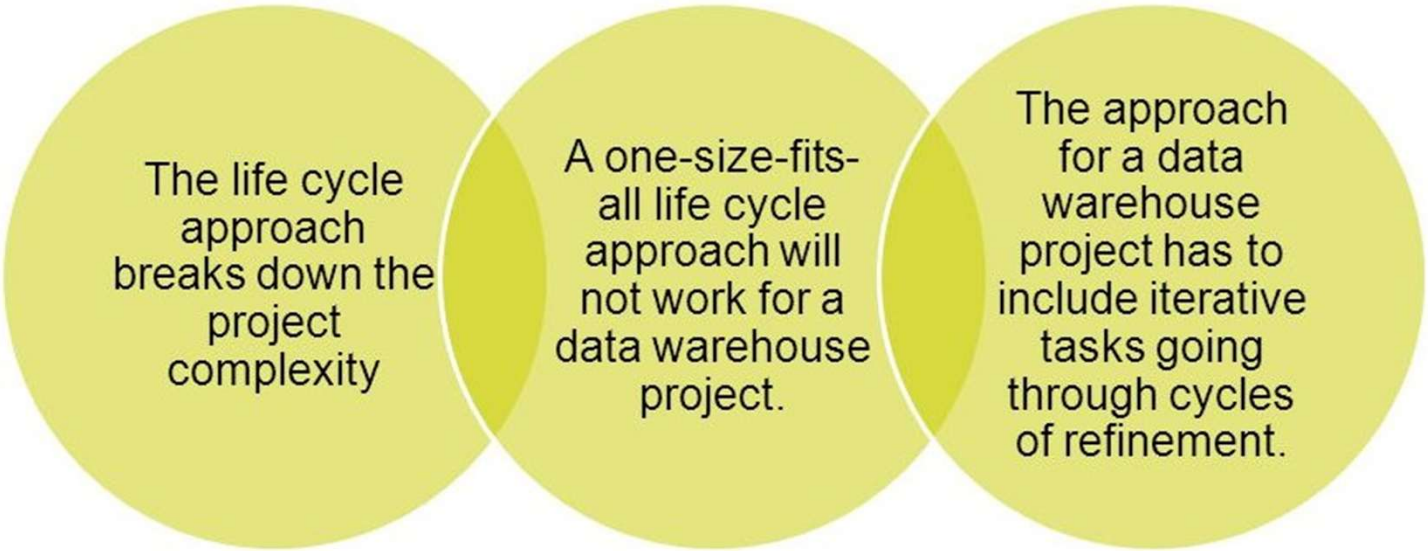
Readiness Assesment Report



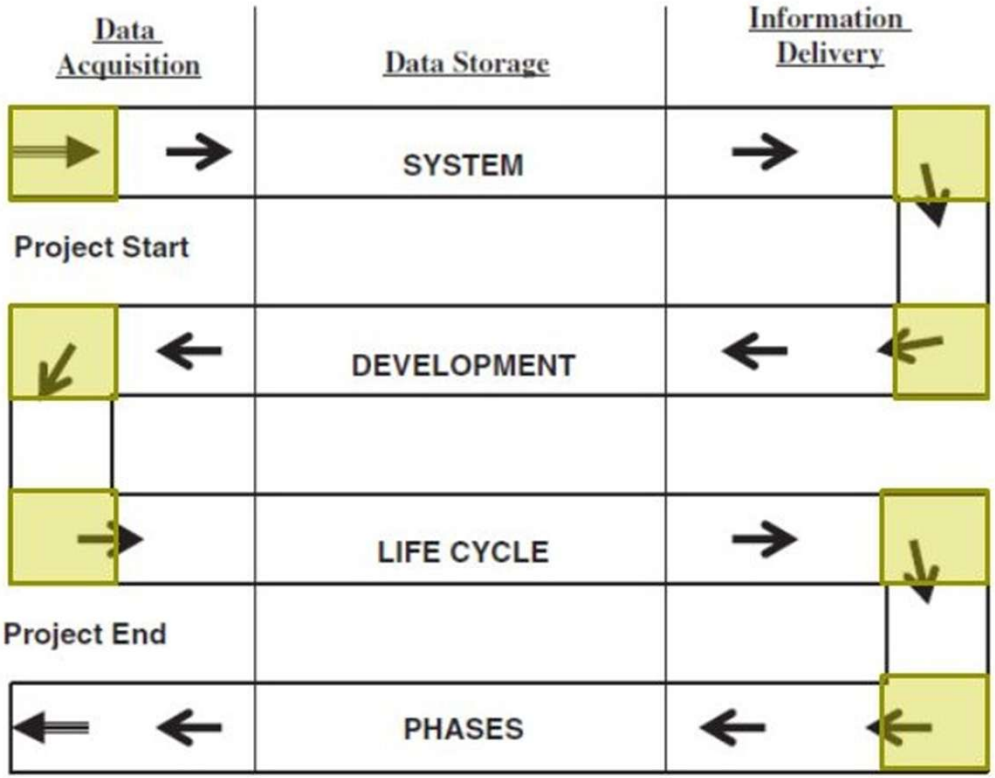
Advantages of the life cycle approach

1. Accomplishes all the major objectives in the system development process.
2. Enforces orderliness and enables a systematic approach to building computer systems.
3. Breaks down the project complexity and removes any ambiguity with regard to the responsibilities of project team members.
4. Implies a predictable set of tasks and deliverables.

Life Cycle Approach



System Development Life Cycle for data warehousing



How Data Warehousing projects are different

- Operational systems are modeled to automate and/or record events in an existing business workflow.
- Data warehouse projects are increasingly being driven by (e..g, demanded by) business users with strong business justifications.
- Data warehouse front-end applications tend to support ad-hoc queries, reporting and dashboards, as opposed to static data entry forms and canned reports
- DW systems are built up with data from existing transaction systems so *data profiling* and *integration* are critical steps
- *Security* and *Compliance* are two areas that require special attention since data from across the organization will be combined together
- Operational systems capture *revenues* fairly accurately, however sources for *costs* and how those costs are allocated against revenues often raise complex issues.
- Technology resources are non-trivial: Data storage, Archiving, Security, OLAP engine, web integration, etc.

Major steps in DW Projects

- *Business and Technical Justification* – In this phase, the project's sponsors detail the business justification, opportunities and benefits as well as the technical justifications for the DW project. Staffing and other necessary resources are identified.
  - **Business Justification:**
    – Review business initiatives and processes
    – Enlist BI sponsors and stakeholders (e.g., potential BI users)
    – Document business benefits and outcomes in terms of adding business value
    – Project scope and Budgeting
  - **Technical Justification:**
    – Product evaluations for proof-of-concept and technology roadmaps
    – Assessing necessary technical skills/expertise
    – Assessing data quality

Major steps in DW Projects

- *Gathering Business Requirements (KPI's)* – In this phase business users are interviewed to determine what measurements / metrics they require. These are called *Key Performance Indicators* (KPI's) and are generally calculated by summing and combining OLTP transactions data. This stage clearly requires deep involvement of the business managers and others in higher level decision making positions.
- *System Design / Modeling* – In this phase, the overall system is designed using conceptual modeling at three levels:
  - *System Architecture Design* – Overall technology architecture (hardware and DBMS software integration) are designed. This step can be done in parallel with data and application design.
  - *Data Modeling* – Data models (*Dimensions* and *facts*) are created and mappings / pipelines from existing operational systems are designed.
  - *BI Application Design* – Applications are designed at the conceptual level. For example, reports, user interfaces, etc. can be mocked up and reviewed by users.
- This stage is carried out by systems analysts in conjunction with the business stakeholders.

Major steps in DW Projects

- *System Development* – In this phase, the designs are implemented in hardware and software. DBMS vendors are selected, data warehouse schemas are created, ETL code is written/configured, and BI applications are coded. This stage is carried out almost exclusively by technologists (programmers, DBAs, etc.) although business users may be called upon for testing.
- *Phased deployment* – In this phase, users are brought on-line (*on boarded*) to the data warehouse.
- *Maintenance and evolution* – Data warehouses undergo continuous evolution as new KPI's and data sources are defined and integrated.

Kimball Lifecycle