UNIT-1

Reporting and analyzing data is crucial because it allows organizations to gain valuable insights from their collected information, enabling informed decision-making, identifying trends, tracking progress, and ultimately improving operations, customer experiences, and overall performance by basing actions on concrete evidence rather than assumptions.

Key reasons to report and analyze data:

Informed decision-making:

By understanding patterns and trends within data, businesses can make strategic choices based on facts rather than intuition.

Performance tracking:

Regularly monitoring key metrics through reports helps identify areas of success and areas needing improvement, allowing for course correction.

Problem identification:

Analyzing data can reveal hidden issues or potential risks that might not be apparent otherwise.

Customer insights:

Analyzing customer data can provide valuable information about their needs, preferences, and behaviors, leading to better product and service development.

Operational efficiency:

Data analysis can identify areas where processes can be optimized to improve productivity and reduce costs.

Market understanding:

Analyzing market trends and competitor data helps businesses stay ahead of the competition.

Communication and transparency:

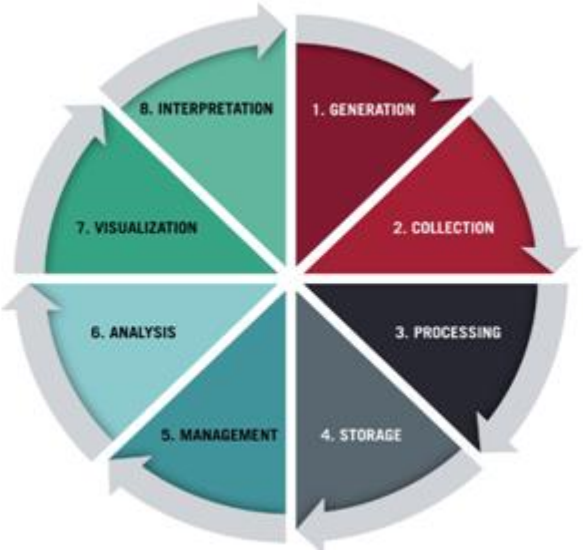
Presenting data in clear reports allows for effective communication of key findings to stakeholders.

The lifecycle of data, transforming raw data into valuable information, typically involves stages like data collection, data cleaning/processing, data storage, data analysis, interpretation, and visualization where raw data is cleaned, structured, analyzed to extract meaningful insights and patterns, ultimately producing valuable information that can be used for decision-making.

Key stages in the data lifecycle:

- **Data Generation:** Data is created from transactions, sensors, user inputs, etc.
- **Data Collection:**

Gathering raw data from various sources like surveys, sensors, databases, or web interactions.



- **Data Cleaning/Preprocessing:**

Removing errors, inconsistencies, and irrelevant data to prepare it for analysis.

- **Data Transformation:**

Structuring and organizing data into a usable format, including data aggregation, normalization, and feature engineering.

- **Data Storage:**

Storing processed data in appropriate databases or data warehouses for easy access and retrieval.

- **Data Analysis:**

Applying statistical methods, machine learning algorithms, or other analytical techniques to identify patterns, trends, and relationships within the data.

- **Interpretation:**

Explaining the results of the analysis, drawing meaningful conclusions, and providing context to the findings.

- **Data Visualization:**

Presenting the analyzed data in a clear and understandable format using graphs, charts, or dashboards to communicate insights effectively.

Important considerations in the data lifecycle:

- **Data Quality:** Ensuring the accuracy, completeness, and consistency of data throughout the lifecycle.

- **Data Governance:** Establishing rules and policies to manage data access, security, and privacy.

- **Data Archiving:** Storing historical data for future reference or legal compliance.

- **Data Destruction:** Securely deleting data when it is no longer needed.

Summarize

- **Data Generation:** Data is created from transactions, sensors, user inputs, etc.
- **Data Collection:** Data is gathered from multiple sources.
- **Data Storage:** Data is stored in databases, data warehouses, or cloud storage.
- **Data Processing:** Cleaning, transforming, and organizing data for analysis.
- **Data Analysis:** Applying statistical and AI techniques to extract insights.
- **Data Visualization:** Presenting insights in a readable format.
- **Data Utilization:** Using analyzed data for decision-making.
- **Data Archiving or Deletion:** Storing historical data or removing outdated data.

What is Business Intelligence (BI)?

1. Definition

Business Intelligence (BI) refers to **technologies, strategies, and practices** used to collect, analyze, and present business data for decision-making.

2. Components of BI

- **Data Warehousing** – Centralized storage of structured data.
- **Data Mining** – Discovering patterns and relationships in data.
- **Reporting and Dashboards** – Visual representation of key performance indicators (KPIs).
- **Predictive Analytics** – Forecasting future trends using AI/ML.

3. Benefits of BI

- Improved **decision-making**.
- Enhanced **operational efficiency**.
- Better **customer insights**.
- Competitive **advantage**.

BI and Data Warehousing (DW) in Today's Perspective

- **Real-time BI:** Companies now demand real-time analytics instead of traditional batch processing.
- **Cloud-Based DW:** Data is now stored and analyzed on cloud platforms like AWS, Azure, and Google Cloud.
- **Big Data Integration:** BI systems now process structured and unstructured data.
- **AI and Machine Learning:** Automating analytics and predictive insights.

Self-Service BI: Non-technical users can analyze data without IT support.

Components or Building Blocks of Data Warehouse

Data warehouse Architecture and its components

Architecture is the proper arrangement of the elements. We build a data warehouse with software and hardware components. To suit the requirements of our organizations, we arrange these building we may want to boost up another part with extra tools and services. All of these depends on our circumstances

1. Source Data Component

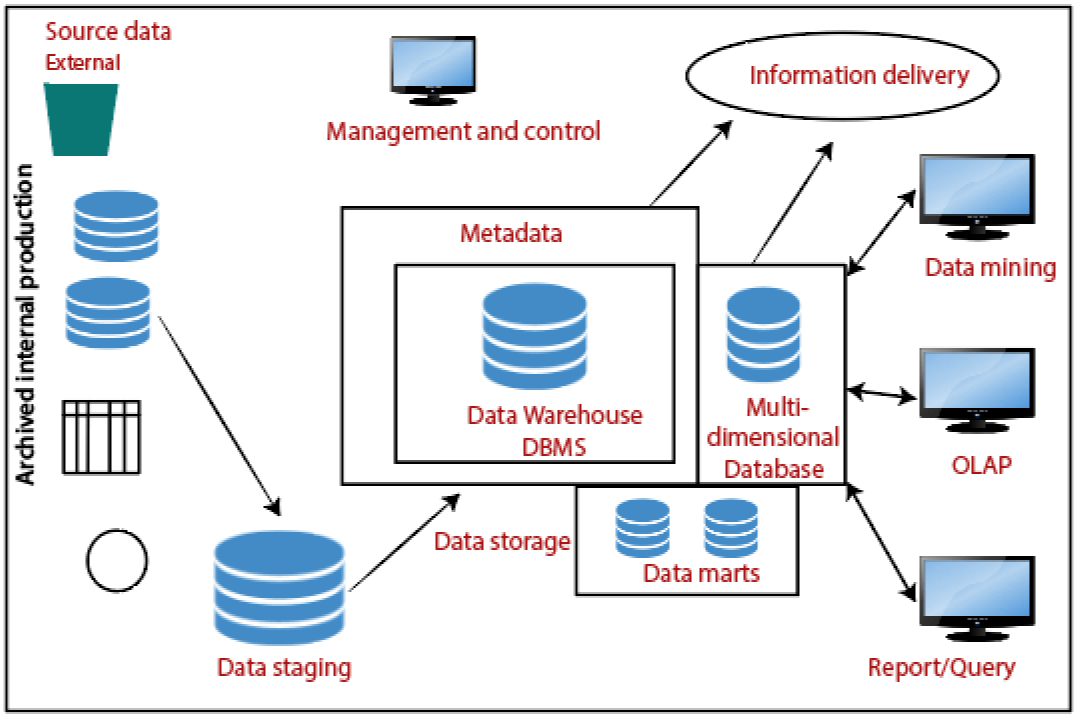
2. Data Staging Component

3. Data Storage Components

4. Data marts

5. Information delivery system

6. Data warehouse management and Control



Components or Building Blocks of Data Warehouse

The figure shows the essential elements of a typical warehouse. We see the Source Data component shows on the left. The Data staging element serves as the next building block. In the middle, we see the Data Storage component that handles the data warehouses data. This element not only stores and manages the data; it also keeps track of data using the metadata repository. The Information Delivery component shows on the right consists of all the different ways of making the information from the data warehouses available to the users.

1.Source Data Component

Source data coming into the data warehouses may be grouped into four broad categories:

Production Data: This type of data comes from the different operating systems of the enterprise. Based on the data requirements in the data warehouse, we choose segments of the data from the various operational modes.

They perform conversions, summarization, key changes, structural changes and condensation. The data transformation is required so that the information can by used by decision support tools. The transformation produces programs, control statements, JCL code, COBOL code, UNIX scripts, and SQL DDL code etc., to move the data into data warehouse from multiple operational systems.

The functionalities of these tools are listed below:

- To remove unwanted data from operational db

- Converting to common data names and attributes
- Calculating summaries and derived data
- Establishing defaults for missing data
- Accommodating source data definition change.

Issues to be considered while data sourcing, cleanup, extract and transformation:

Data heterogeneity: It refers to DBMS different nature such as it may be in different data modules, it may have different access languages, it may have data navigation methods, operations, concurrency, integrity and recovery processes etc

Internal Data: In each organization, the client keeps their "**private**" spreadsheets, reports, customer profiles, and sometimes even department databases. This is the internal data, part of which could be useful in a data warehouse.

Archived Data: Operational systems are mainly intended to run the current business. In every operational system, we periodically take the old data and store it in achieved files.

External Data: Most executives depend on information from external sources for a large percentage of the information they use. They use statistics associating to their industry produced by the external department.

2.Data Staging Component

After we have been extracted data from various operational systems and external sources, we have to prepare the files for storing in the data warehouse. The extracted data coming from several different sources need to be changed, converted, and made ready in a format that is relevant to be saved for querying and analysis.

We will now discuss **the three primary functions** that take place in the staging area.

1) Data Extraction: This method has to deal with numerous data sources. We have to employ the appropriate techniques for each data source.

2) Data Transformation

Data transformation also contains purging source data that is not useful and separating outsource records into new combinations. Sorting and merging of data take place on a large scale in the data staging area. When the data transformation function ends, we have a collection of integrated data that is cleaned, standardized, and summarized.

3) Data Loading: Two distinct categories of tasks form data loading functions. When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the information into the data warehouse storage. The initial load moves high volumes of data using up a substantial amount of time.

3.Data Storage Components

A) Data warehousing

Data storage for the data warehousing is a split repository. The data repositories for the operational systems generally include only the current data. Also, these data repositories include the data structured in highly normalized for fast and efficient processing.

The data source for data warehouse is coming from operational applications. The data entered into the data warehouse transformed into an integrated structure and format. The transformation process involves conversion, summarization, filtering and condensation. The data warehouse must be capable of holding and managing large volumes of data as well as different structure of data structures over the time.

Data warehousing is the process of storing data, and data mining is the process of analyzing that data. Both are key components of business intelligence (BI).

Data warehousing

- A centralized repository that stores data from various sources
- A relational database that can store large amounts of data
- A collection of databases that can be integrated to provide new insights
- A time-variant model that stores historical data and continuously adds new data

B) Meta data

It is data about data. It is used for maintaining, managing and using the data warehouse. It

is classified into two:

a) **Technical Meta data:** It contains information about data warehouse data used by warehouse

designer, administrator to carry out development and management tasks. It includes,

- Info about data stores
- Transformation descriptions. That is mapping methods from operational db to warehouse db
- Warehouse Object and data structure definitions for target data
- The rules used to perform clean up, and data enhancement

- Data mapping operations
- Access authorization, backup history, archive history, info delivery history, data acquisition history, data access etc.

b) **Business Meta data:** It contains info that gives info stored in data warehouse to users. It includes, Subject areas, and info object type including queries, reports, images, video, audio clips etc.

- Internet home pages
- Info related to info delivery system
- Data warehouse operational info such as ownerships, audit trails etc.,
- Meta data helps the users to understand content and find the data. Meta data are stored in a separate data stores which is known as informational directory or Meta data repository which helps to integrate, maintain and view the contents of the data warehouse.

4. Data marts

A data mart is a database that stores and organizes data for a specific business unit or department. It's a subset of a company's larger data storage system.

Data mart is used in the following situation:

- Extremely urgent user requirement
- The absence of a budget for a full scale data warehouse strategy
- The decentralization of business needs
- The attraction of easy to use tools and mind sized project

5. Information delivery system

An "information delivery system" in a data warehouse refers to the set of tools and processes that allow users to access, analyze, and visualize data stored within the warehouse, effectively presenting relevant information in a format that supports informed decision-making, typically through reports, dashboards, and interactive queries, delivered via user interfaces like web applications or dedicated BI tools.

Key points about information delivery systems in data warehouses:

Its purpose is to provide info to business users for decision making. There are five categories:

- Data query and reporting tools
- Application development tools
- Executive info system tools (EIS)
- OLAP tools

- Data mining tools

Query and reporting tools are used to generate query and report. There are two types of reporting tools. They are:

1. Production reporting tool used to generate regular operational reports

2. Desktop report writer are inexpensive desktop tools designed for end users.

Managed Query tools: used to generate SQL query. It uses Meta layer software in between users and databases which offers a point-and-click creation of SQL statement. This tool is a preferred choice of users to perform segment identification, demographic analysis, territory management and preparation of customer mailing lists etc.

Application development tools: This is a graphical data access environment which integrates

OLAP tools with data warehouse and can be used to access all db systems

OLAP Tools: are used to analyze the data in multi dimensional and complex views. To enable multidimensional properties it uses MDDB and MRDB where MDDB refers multi dimensional data base and MRDB refers multi relational data bases.

Data mining tools: are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes

6. Data warehouse management and Control

The management of data warehouse includes ecurity and priority management

- Monitoring updates from multiple sources
- Data quality checks
- Managing and updating meta data
- Auditing and reporting data warehouse usage and status
- Purging data
- Replicating, sub setting and distributing data

- Backup and recovery
- Data warehouse storage management which includes capacity planning, hierarchical storage management and purging of aged data etc.,