Fundamentals of Data Warehouse

Data Warehouse :- centralized repo, storing, consolidating, managing large volumes of data from various sources, designed to support Business Intelligence (BI), efficient analysis & decision making.

Evolution of Data Warehouse

- RDBMS (Early 1980s)
  - ↳ Improved access to valuable info
  - ↳ Transactional databases not always optimized for reporting or analytical needs.

- Genesis of Data Warehousing (Late 1980s)
  - ↳ 'Business Data Warehouse' by IBM
- Bill Inmon's contributions
  - ↳ Approach → centralized repo modeled to 3NF.
    → Top-down
  - ↳ Defn → A warehouse is subject-oriented, integrated, time-variant and non-volatile data collection for management decision-making.

- Ralph Kimball's contributions
  - ↳ Approach → star schema modeling
    → easy to understand for end users
  - ↳ Defn → A warehouse is the conglomerate of all data marts within the enterprise, with information stored in the dimensional model

Need for data warehouse

- Enhancing the turnaround time for analysis & reporting (data from single source)
- Improved BI (decision based on reliable facts)

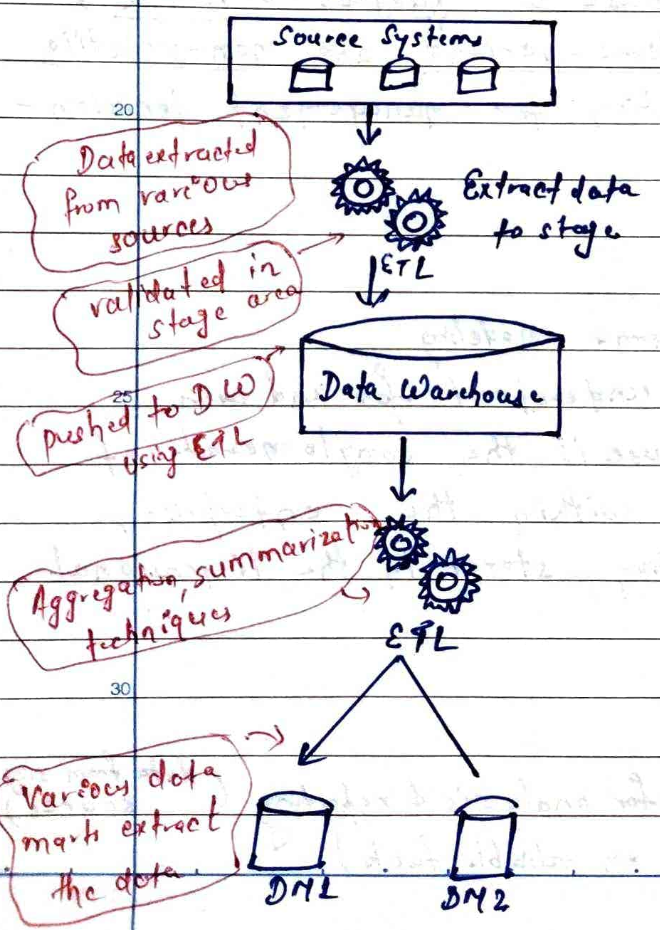
- Benefit of historical data (time-period analysis, trend analysis)
- Standardization of data (data from heterogenous sources → single format)
- Immense ROI (Return on Investment) (Additional revenues / reduced expenses)

Benefits of Data Warehouse

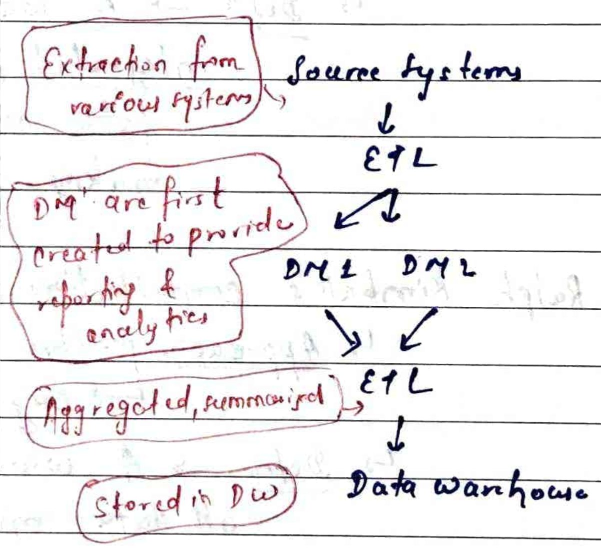
- Faster and accurate Data Analytics
- Increase Revenue and Returns
- Better Efficiency
- Access to Historical Insights
- Improved data security
- Scalability
- Works on premises and on cloud

Data Warehousing design approaches

1) Top down approach (Bill Inmon)

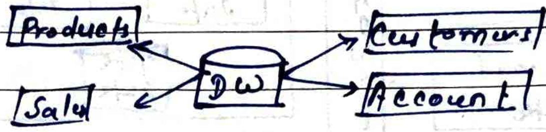


2) Bottom-up approach (Ralph Kimball)

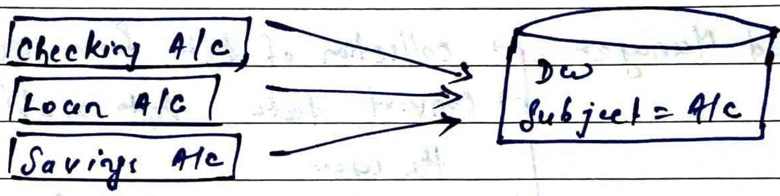


Characterstics of a data warehouse

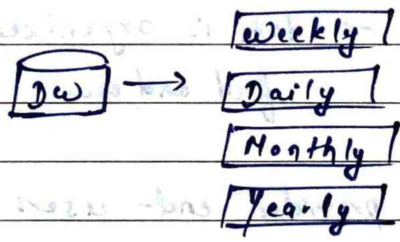
- Subject Oriented → provides info about a specific theme
  → focuses on data demonstration and analysis to make different decisions.



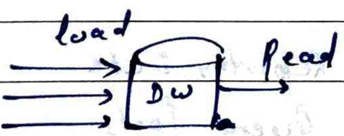
- Integrated → a common system to measure all similar data from multiple systems
  → consistent, readable & coded



- Time-variant → data held in various intervals such as weekly, monthly, and yearly.
  → history
  → data can't be changed, modified or updated once stored.



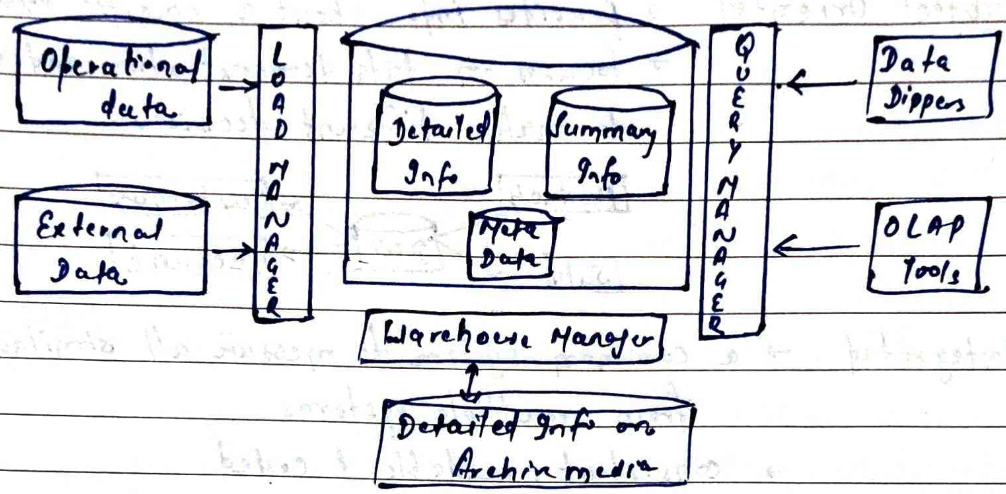
- Non-Volatile → data in DW → permanent



whereas in operational DB, we can Select / Insert / Delete / Update

How DW works?

Components of a DW



- Load Manager
  - → collection of data from various sources
  - → convert data into usable form for the users
  - → import / export data from operational systems.
  - → includes program for pooling out the data, validation, accuracy, extraction, cleaning etc.

- Warehouse Manager
  - → large, physical db that holds vast amt. of info
  - → data is organized such that easy to find and use

- Query Manager → provides end-users with access to the stored warehouse info.
  - ↳ through various tools

- End user access tools
  - → Reporting Data ↳ tools for EIS
  - → Query tools ↳ tools for OLAP
  - → Data dippes

OLTP & OLAP

OLTP → Online Transaction Processing
OLAP → Online Analytical Processing

OLTP → captures and maintains transaction data in a database.
→ emphasis on fast processing, because OLTP DBs are read, written and updated frequently.

OLAP → applies complex queries to large amount of historical data, aggregated from OLTP databases.
→ emphasis on response time to these complex queries.

	OLTP	OLAP
Characteristics	Handles a large number of small transactions	Handles large volumes of data with complex queries
Query types	Simple standardized queries	Complex queries
Operations	Based on INSERT, UPDATE, DELETE commands	Based on SELECT commands to aggregate data for reporting
Response time	Milliseconds	Seconds, minutes, or hours depending on the amount of data to process
Design	Industry-specific, such as retail, manufacturing, or banking	Subject-specific, such as sales, inventory, or marketing
Source	Transactions	Aggregated data from transactions
Purpose	Control and run essential business operations in real time	Plan, solve problems, support decisions, discover hidden insights
Data updates	Short, fast updates initiated by user	Data periodically refreshed with scheduled, long-running batch jobs
Space requirements	Generally small if historical data is archived	Generally large due to aggregating large datasets
Backup and recovery	Regular backups required to ensure business continuity and meet legal and governance requirements	Lost data can be reloaded from OLTP database as needed in lieu of regular backups
Productivity	Increases productivity of end users	Increases productivity of business managers, data analysts, and executives
Data view	Lists day-to-day business transactions	Multi-dimensional view of enterprise data
User examples	Customer-facing personnel, clerks, online shoppers	Knowledge workers such as data analysts, business analysts, and executives
Database design	Normalized databases for efficiency	Denormalized databases for analysis

Data Granularity

- Granularity refers to the level of detail in data stored in a data warehouse.
- Multiple granular level exist in data warehouses to meet various analytical requirements.
- Operational data → lowest level
- Fine granularity requires substantially permanent data storage.
- Allows users to navigate from summarized info to finer details
- Balancing the level of detail with performance requirements is essential.

Metadata and Warehousing

In DW, data is stored using a common schema controlled by a common dictionary.

Metadata should contain following info :-

- Data Structure (Programmer's view, Analysts' view)
- Data sources
- Data transformation details
- Model of data
- Connection b/w data model & data warehouse
- Data extraction history.

Data Warehousing Application

- Investment & Insurance → analyze customer, market trends
- Healthcare :- forecast treatment's outcomes, research
- Retail :- Distributing, marketing, pricing policies
- Social Media Websites :- fb, twitter, impressions, locations, members.
- Banking :- spending patterns, special offers, deals
- Govt :- store and analyze taxes
- Airlines :- flight freq., road profitability analyses
- Public Sectors :- helps govt. & agencies manage their data & records.

Types of data warehouses

(i) Enterprise
- central repo db
- central place when all business info from diff. sources are made available.

(ii) Operational
- data refreshed in near real-time
- used for routine commercial activity.

(iii) Data Mart
- subset of DW
- supports specific region, business unit etc
- contains subset of data in DW → enhancing user experiences by reducing volume of data.

Popular data warehouse platform

- Google Big Query
  - → cost-effective, built-in machine learning capabilities.
  - → integrated (can be) with Cloud ML & TensorFlow.
  - → scalable & serverless

- AWS Redshift
  - → cloud based
  - → can process petabyte of data fast.
  - → suitable for high speed data analytics.

- Snowflake
  - → make business more data-driven
  - → set up an enterprise-grade cloud DW
  - → dependent on Azure, Amazon web Services, Google Cloud services

- Microsoft Azure Synapse
  - → robust platform for data management, analytics, integration & more
  - → AI, Blockchain etc.