Analytical Study of Data Warehouse

Research · January 2019

DOI: 10.13140/RG.2.2.22600.65285

1 author:

Raj Sinha
Lovely Professional University
40 PUBLICATIONS **37** CITATIONS

SEE PROFILE

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Analytical Study of Data Warehouse

Raj Sinha*

Abstract
A data warehouse is a system that stores data from a company's operational databases as well as external sources. Data warehouse platforms are different from operational databases because they store historical information, making it easier for business leaders to analyze data over a specific period of time. Data warehouse platforms also sort data based on different subject matter, such as customers, products or business activities. First, we determined the business objectives for the system. Then we collected and analyzed information about the enterprise. We identified the core business processes that the company needed to track, and constructed a conceptual model of the data. Then we located the data sources and planned data transformations. Finally, we set the tracking duration.
***Keywords*:** *Company, Information, Storage, Business, Data*.

Introduction

Data Warehouse is centralized data repositories storage for analytical and reporting purposes. According to Ralph Kimball, "Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model". The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization.

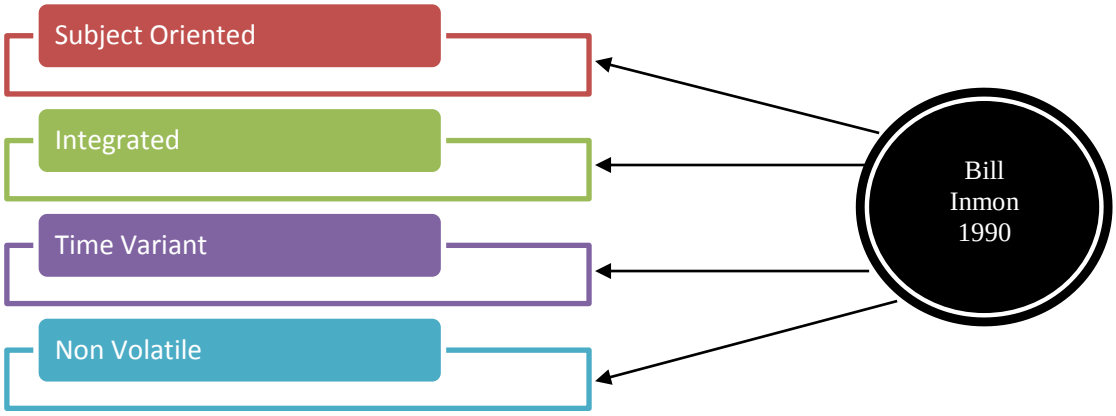


Fig: 1.0 Bill Inmon definition of data warehouse

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inmon:

- ✓ Subject oriented
- ✓ Integrated
- ✓ Non Volatile
- ✓ Time Variant

* Guest Faculty (MCA, MBA) LNMIEDSC, Patna

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Subject Oriented: It provides information around a subject rather than the organization's ongoing operations. In simple words, data warehouse concentrates on a particular subject area. For example: let's us take the example of sales company.

- Operational System: Details (Sales Receipt, Contract, Delivery)
- Subject area: Sales

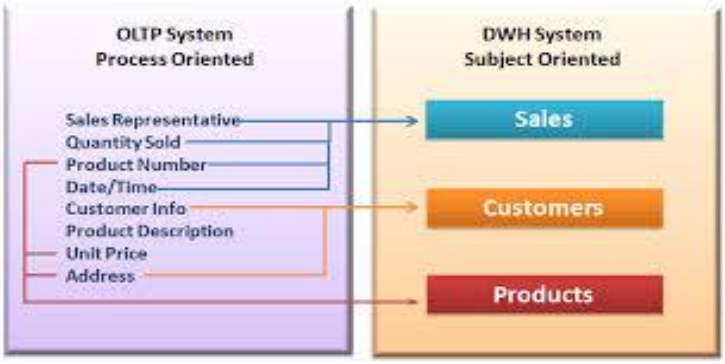


Fig: 1.1 Subject Oriented

Integrated: It consists of data that has been combined from numerous sources.

- Obtain data: mainframes (Product Info), flat files (Contract Info), SQL Server database (Sales Details), etc
- Combine: Data warehouse (Mainframes (Product Info) + flat files (Contract Info) +SQL Server database (Sales Details))

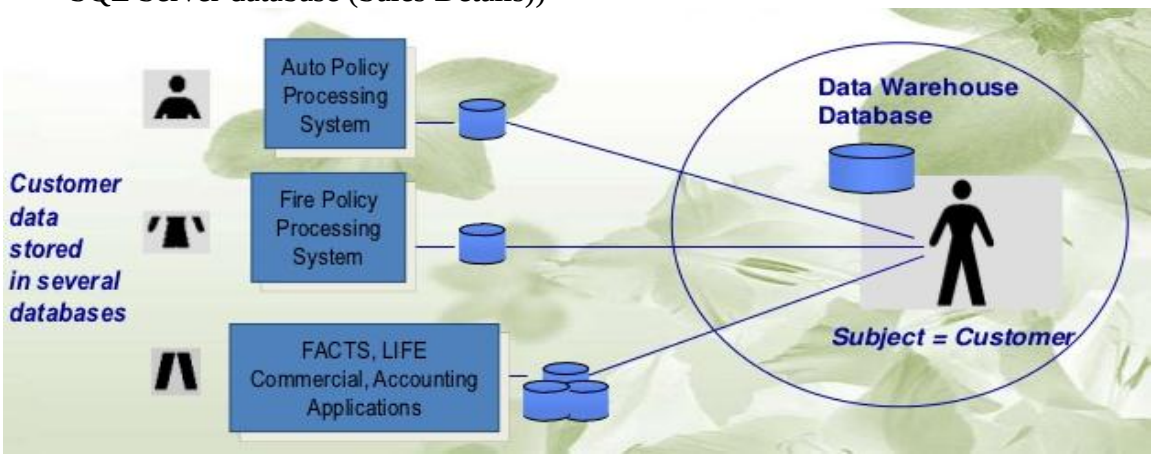


Fig: 1.2 Integrated

Time Variant: All data in the data warehouse is identified with a particular time period. Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

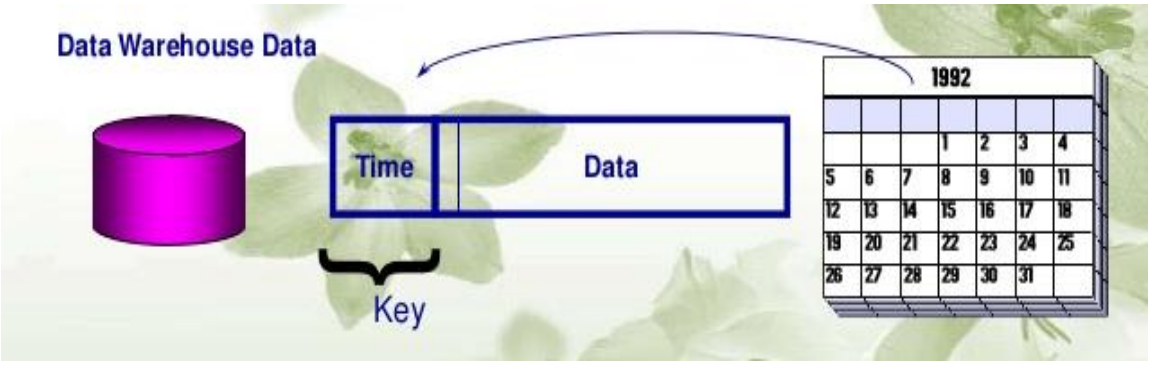


Fig: 1.3 Time Variant

Non-volatile: It means that, once entered into the data warehouse, data should not change. A different version of data is stored in the data warehouse indicating if any insert, update or delete have occurred in the database. Hence, the original data is never altered. In short, Data is stable in a data warehouse.

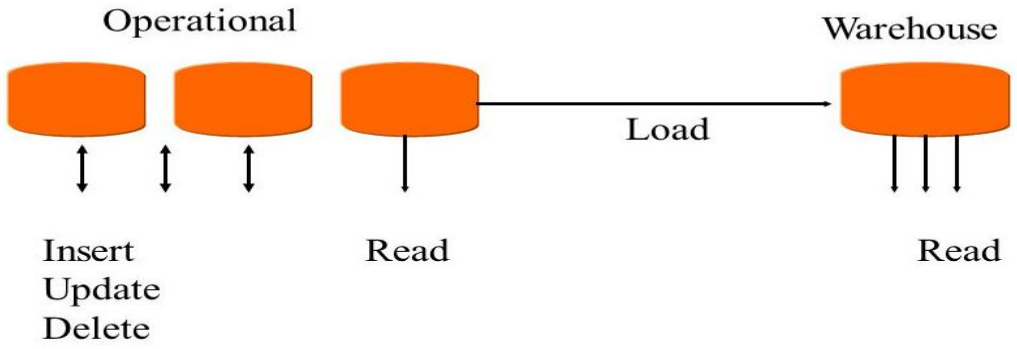


Fig: 1.4 Non Volatile

Difference between Bill Inmon and Ralph Kimball

Ralph Kimball provided a more concise definition of a data warehouse: "A data warehouse is a copy of transaction data specifically structured for query and analysis."This is a functional view of a data warehouse.

Kimball did not address how the data warehouse is built like Inmon did; rather he focused on the functionality of a data warehouse.

REQUIREMENTS	INMON	KIMBALL
Data Stability	Source systems changes frequently	Stable source systems
Staff requirement	Large	Small
Delivery	Slow and Long	Quick turnaround
Cost	Low upfront cost	High expenditure

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

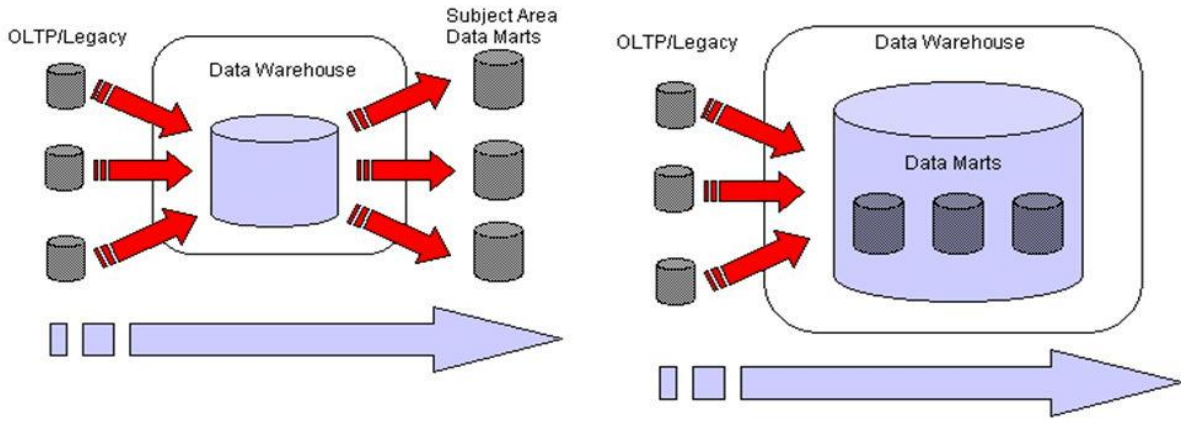


Fig: 1.5 Inmon vs Kimball

Data warehouse system is also known by the following name:

- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse

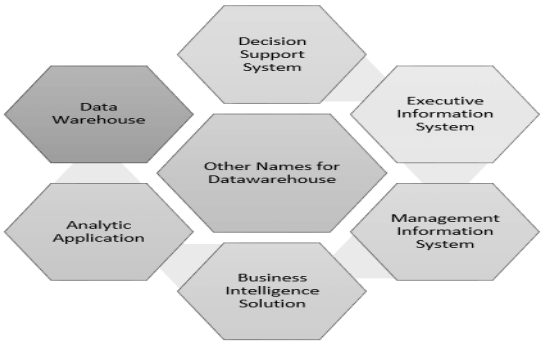
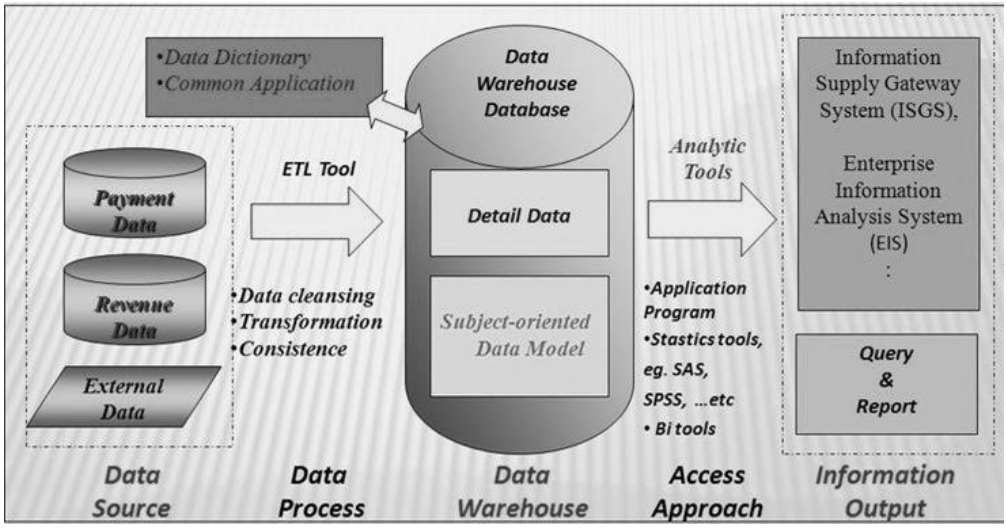


Fig: 1.6 Other Names of data warehouse

In the Airline system, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc. It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations. Healthcare sector also used Data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A



Objectives

- Reconcile different views of the same data
- Provide a consolidated picture of enterprise data
- Create a virtual "one-stop-shopping" data environment
- Improve quality of data
- Minimize inconsistent reports
- Capture and provide access to metadata
- Provide capability for data sharing
- Integrate data from multiple sources
- Merge historical data with current data

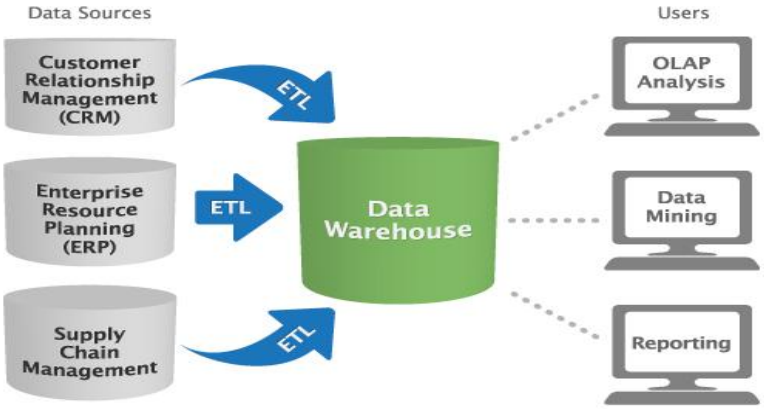


Fig: 1.7 Objectives of data warehouse

Characteristics: The key characteristics of a data warehouse are as follows:

- Some data is denormalized for simplification and to improve performance
- Large amounts of historical data are used
- Queries often retrieve large amounts of data
- Both planned and ad hoc queries are common
- The data load is controlled

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

NOTE: In general, fast query performance with high data throughput is the key to a successful data warehouse.

Need of data warehouse

- Ensure consistency.
- Make better business decisions.
- Improve organization bottom line with the decrease costs, maximize efficiency and increase sales.
- 40-60% reduction in time to analyze data
- 100% confidence in your data
- Higher quality insights
- Better data security

Fig: 1.8 Need of data warehouse

Types of data warehouse: Three main types of Data Warehouses are:

- Enterprise Data Warehouse
- Operational Data Store
- Data Mart

Data warehouse architecture: There are mainly 3 types of architecture:

- Single-tier architecture
- Two-tier architecture
- Three-tier architecture

Single-tier architecture: The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

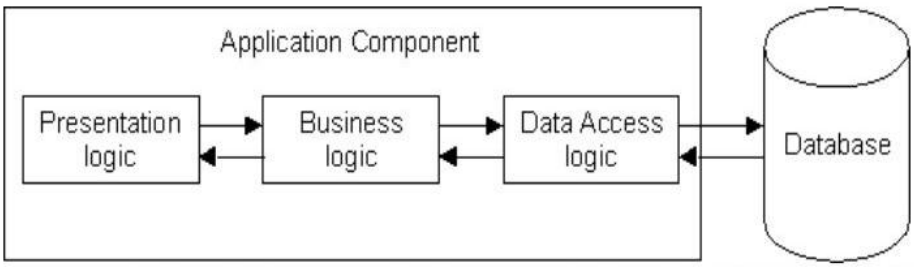


Fig: 1.9 Single tier architecture

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Two-tier architecture: Two-layer architecture separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

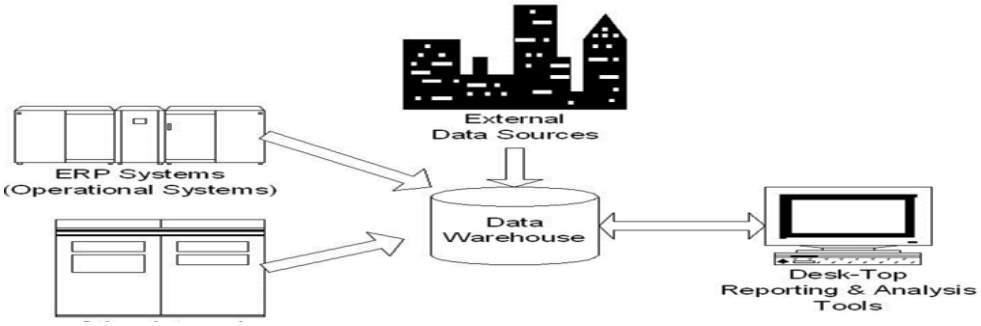


Fig: 1.10 Two-Tier architecture

Three-tier architecture: This is the most widely used architecture. It consists of the Top, Middle and Bottom Tier.

- Bottom Tier: The bottom tier of the architecture is the data warehouse database server. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.
- Middle Tier: The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database.
- Top-Tier: The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools.

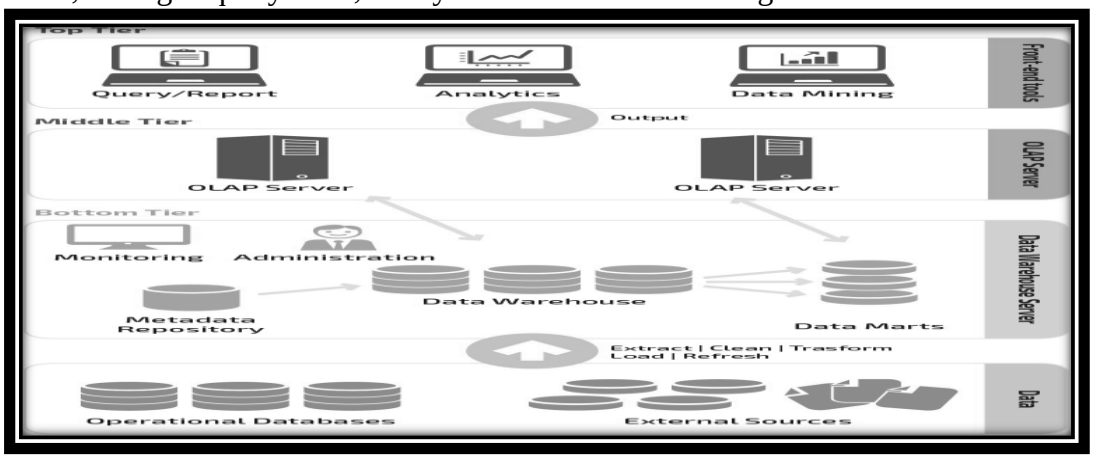


Fig: 1.11 Three-tier data warehouse

Components of a data warehouse: There are mainly five components of Data Warehouse:

- Data Warehouse Database
- Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)
- Metadata
- Query Tools
- Data warehouse Bus Architecture

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Data Warehouse Database: The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins; aggregates are resource intensive and slow down performance.

Hence, alternative approaches to Database are used as listed below-

- In a data warehouse, relational databases are deployed in parallel to allow for scalability. Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.
- New index structures are used to bypass relational table scan and improve speed.
- Use of multidimensional database (MDDBs) to overcome any limitations which are placed because of the relational data model. Example: Essbase from Oracle.

Sourcing, Acquisition, Clean-up and Transformation Tools (ETL): They are also called Extract, Transform and Load (ETL) Tools. An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, concatenate, etc. and then load the data to Data Warehouse system. The data is loaded in the DW system in the form of dimension and fact tables. These ETL Tools have to deal with challenges of Database & Data heterogeneity.

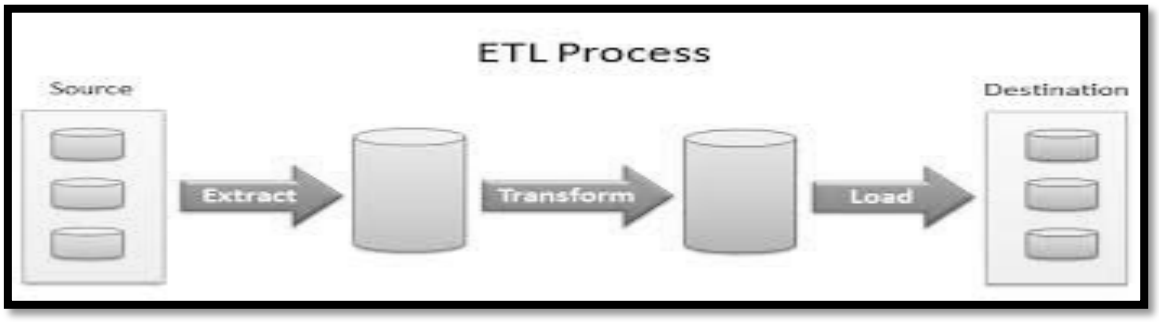


Fig: 1.12 Extract, Transform and Load Process

Metadata: Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse. For example, the index of a book serves as a metadata for the contents in the book. They can be classified into following categories:

- Technical Metadata: Basically used by Data warehouse designers and administrators.
- Business Metadata: It contains detail to understand information stored in the data warehouse.

Query Tools: Query tools allow users to interact with the data warehouse system in order to provide information to businesses to make strategic decisions. This is the primary objects of data warehousing.

These tools fall into four different categories:

- Query and reporting tools
- Application Development tools
- Data mining tools
- OLAP tools

Query and reporting tools: They are further divided into:

Reporting tools: They are further divided into:

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

- Production reporting tools: allows organizations to generate regular operational reports. It also supports high volume batch jobs like printing and calculating. Some popular reporting tools are Brio, Business Objects, Oracle, PowerSoft, SAS Institute. Desktop report writer: designed for end-users for their analysis
- Managed query tools: This kind of access tools helps end users to resolve snags in database and SQL and database structure by inserting meta-layer between users and database.

Application Development tools: Sometimes built-in graphical and analytical tools do not satisfy the analytical needs of an organization. In such cases, custom reports are developed using Application development tools.

Data mining tools: Data mining is a process of discovering meaningful new correlation, pattens, and trends by mining large amount data. Data mining tools are used to make this process automatic.

OLAP tools: These tools are based on concepts of a multidimensional database. It allows users to analyze the data using elaborate and complex multidimensional views.

Data warehouse Bus Architecture: Data warehouse Bus determines the flow of data in your warehouse. The data flow in a data warehouse can be categorized as Inflow, up flow, down flow, Outflow and Meta flow.

General stages of data warehouse

Step 1: Determine Business Objectives
Step 2: Collect and Analyze Information
Step 3: Identify Core Business Processes
Step 4: Construct a Conceptual Data Model
Step 5: Locate Data Sources and Plan Data Transformations
Step 6: Set Tracking Duration
Step 7: Implement the Plan

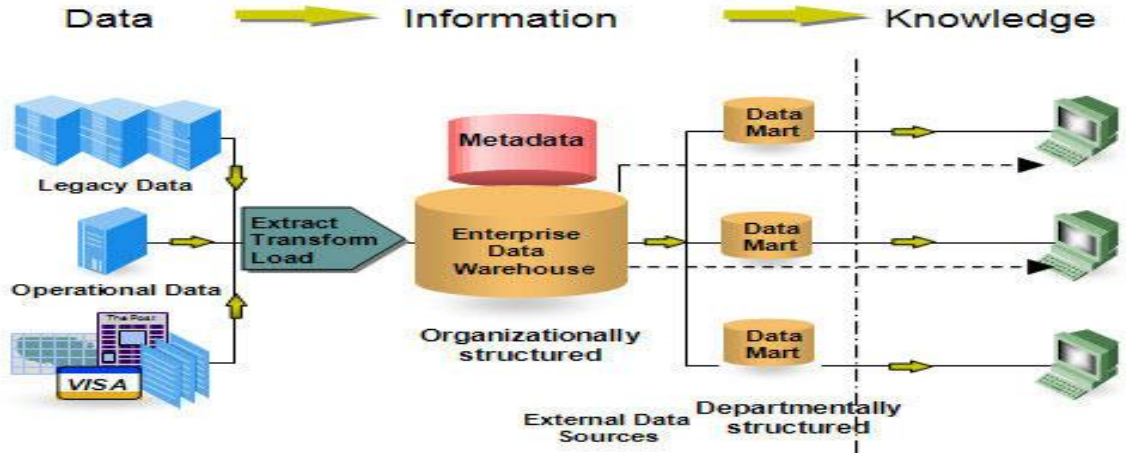


Fig: 1.13 General Stages of Data warehouse

Application of data warehouse

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

International Journal of Management, IT & Engineering
Vol. 8 Issue 1(1), January 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Implementation of data warehouse: Basic data warehouse implementation phases are:

- Current situation analysis
- Selecting data interesting for analysis, out of existing database
- Filtering and reducing data
- Extracting data into staging database
- Selecting fact table, dimensional tables and appropriate schemes
- Selecting measurements, percentages of aggregations and warehouse methods
- Creating and using the cube

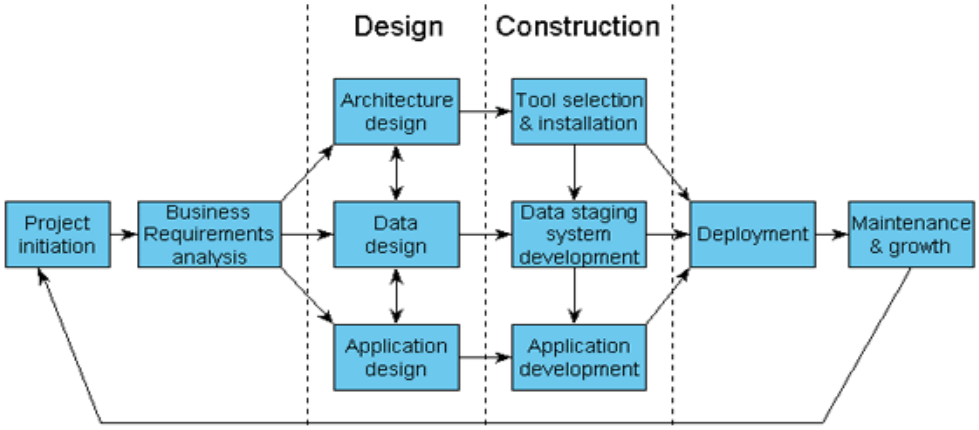


Fig: 1.15 Implementation of data warehouse

Trends of data warehouse: Oracle, a well-known player in the market, in 2014 identified the top 10 trends in data warehousing which are as follows:

- The "datafication" of the enterprise spawns more-capable data warehouses.
- Physical and logical consolidation reduces costs.
- Hadoop optimizes data warehousing environments by accelerating data transformation.
- Customer experience (CX) strategies gain real-time insight to improve marketing campaigns.
- Engineered systems become the de facto standard for large-scale information management activities.
- On-demand sandbox analytics environments meet rising demand for rapid prototyping and information discovery.
- Data compression enables high-value analytics.
- In-database analytics simplifies data-driven analysis.
- In-memory technologies supercharge data warehouse performance.
- Data warehouses become more critical to business operations.

Data warehouse tools: Some most prominent one are:

- MarkLogic
- Oracle
- Amazon RedShift

Advantages of data warehouse

- Integrating data from multiple sources;
- Performing new types of analyses; and
- Reducing cost to access historical data.
- Ensures Data Quality and Consistency

- Saves Time and Money
- Tracks Historically Intelligent Data
- Generates high Return On Investment
- Better enterprise intelligence**.**
- More cost-effective decision-making
- Competitive advantage

Disadvantages of data warehouse

- Costly to maintain
- Underestimation of resources of data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long-duration projects
- Complexity of integration

Conclusion

A data warehouse is "a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources". Data warehouse allows business users to quickly access critical data from some sources all in one place. Being a subject-oriented, integrated, time-variant and volatile, data warehousing caters several advantages to enterprises and users when implemented for business purposes. The successful application of DWH delivers great results and improves the overall functioning of every organization.

References

1. Barry, D., Data Warehouse from Architecture to Implementation, Addison-Wesley, 1997.Hall, 2000.
2. http://www.informit.com/store/database
3. https://www.glowtouch.com/the-benefits-of-data-warehousing-and-etl/
4. Joe Caserta and Ralph Kimball.," The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data"
5. Kachur, R. J. The Data Warehouse Management Handbook. Upper Saddle River, N.J.: Prentice
6. Suknović, M. (2016). Data warehousing and data mining- A case study. Yugoslav Journal of operations Research, 15(1).
7. Mattéo Golfarelli and Stefano Rizzi," Data Warehouse Design: Modern Principles and Methodologies".Organizational Sciences, Belgrade, 2003.
8. Ponniah, P. (2004). Data warehousing fundamentals: a comprehensive guide for IT professionals. John Wiley & Sons.
9. Anahory, S., & Murray, D. (1997). Data warehousing in the real world: a practical guide for building decision support systems. Harlow, UK: Addison-Wesley.